INTOSAI WGBD
INTOSAI Working Group on Big Data

CASE STUDIES ON THE USE OF BIG DATA IN AUDITING

INTOSAI WGBD
INTOSAI Working Group on Big Data

INTOSAI Working Group on Big Data

July 2025

FOREWORD

In today's world where the wave of digitalization is sweeping the globe, big data technology is reshaping the paradigms and boundaries of public sector auditing. Big data auditing is emerging as a new driver for Supreme Audit Institutions (SAIs) worldwide to promote efficiency and effectiveness.

The INTOSAI Working Group on Big Data (WGBD) has compiled this collection of case studies, bringing together 28 innovative practices in the field of big data audit. These cases showcase cross-regional and multilevel big data audit applications, aiming to provide audit professionals around the world with practical solutions for digital transformation. The cases vividly illustrate how big data technology empowers public sector auditing, driving a paradigm shift — from sampling to full-population analysis, from post supervision to real-time monitoring, and from compliance audit to in-depth performance evaluation.

We would like to extend our gratitude to 17 SAIs — the SAIs of Austria, Brazil, Chile, China, Denmark, Egypt, Finland, Indonesia, Pakistan, Portugal, Republic of Korea, Russia, Slovakia, Türkiye, Ukraine, the United States, and Vietnam — for sharing their invaluable experiences and innovative applications in big data audit. We also express our appreciation to the Secretariat of WGBD for their dedicated efforts and contributions in compiling this publication.

I hope you find the case studies useful.

Hou Kai

Auditor General of the National Audit Office of China

Chair of INTOSAI Working Group on Big Data

ABOUT INTOSAI WORKING GROUP ON BIG DATA

INTOSAI Working Group on Big Data (WGBD) is a specialized working group approved by INTOSAI under Strategic Goal Three: Knowledge Sharing and Services. Its objectives are to identify the challenges and opportunities faced by SAIs in the era of big data, to summarize the knowledge and experience in the field of big data audit, and to strengthen bilateral and multilateral technical cooperation on big data.

Chair

China

Vice Chair

United States

Members

- Argentina
- Azerbaijan
- Austria
- Bangladesh
- Belgium
- Bhutan
- Brazil
- Canada
- Chile
- China
- Denmark
- Ecuador
- Egypt
- Estonia
- Fiji
- Finland
- France
- Georgia
- India
- Indonesia
- Iran
- Kenya
- Kuwait
- Malaysia
- Mexico
- Netherlands
- New Zealand
- Norway
- Pakistan
- Peru
- Philippines
- Portugal
- Republic of Korea
- Russia
- Samoa
- Senegal
- Slovakia
- Thailand
- Türkiye
- Ukraine
- United Kingdom
- United States
- Vietnam
- Zambia

Observers

- AFROSAI-E
- Bulgaria
- European Court of Auditors
- INTOSAI Development Initiative
- Ireland
- Malta

CONTENTS

01 Equal Employment Opportunity Commission (EEOC) Oversight of the Length of the Charge Intake Process

I. Introduction

Equal Employment Opportunity Commission (EEOC) is the primary federal agency that enforces federal laws prohibiting employment discrimination and investigates allegations of employment discrimination. It received over 60,000 discrimination charges in fiscal year 2021. EEOC also provides guidance to employers and employees to help ensure compliance with federal anti-discrimination laws. According to EEOC, its enforcement and outreach activities play a critical role in advancing equal opportunity in the workplace.

GAO was asked to review EEOC's charge investigation and outreach processes. This report examines (1) the steps EEOC has taken since fiscal year 2011 to address its pending charge inventory, and what is known about the length and quality of its investigation process; and (2) how EEOC evaluates its outreach efforts, and the challenges officials said they face when providing outreach.

GAO analyzed EEOC charge data for fiscal years 2011 through 2021 and reviewed EEOC documentation. GAO conducted nine discussion groups with officials in EEOC field offices selected based on geography and workload, and interviewed EEOC headquarters and field officials.

II. Main Content

From fiscal years 2011 to 2021, the EEOC reduced its pending charge inventory—the number of open charges as of the end of the fiscal year. Specifically, the pending charge inventory decreased from 82,199 to 43,520 charges during this time. EEOC officials identified steps that contributed to the reduction. For example, in fiscal year 2018, EEOC

implemented an online portal, which allows individuals to file an inquiry and helps them determine if they meet the criteria to proceed with filing a charge.

EEOC monitors the quality of its investigations; however, GAO found that EEOC does not monitor the length of the intake process across field offices. For the purposes of this report, the intake process begins when an individual files an inquiry and ends when an EEOC official interviews the individual about the alleged incident. GAO's analysis of EEOC data from the online portal shows that the average length of the intake process varied among EEOC's 53 field offices. EEOC officials said they could monitor data on the length of the intake process, but do not routinely do so because they have focused on ensuring that individuals who file inquiries do not miss the statutory deadline for filing a charge. By not monitoring the length of the intake process across field offices, EEOC is missing information that could allow it to better identify and provide support to offices that take longer to complete the intake process.

The Length of the Intake Process Varies Greatly among EEOC Field Offices. EEOC tracks certain measures related to its outreach efforts such as the number of events it hosts and attendees, and is taking steps to develop measures to assess the impact of its efforts. However, Outreach and Education Coordinators (OEC) GAO spoke with identified challenges with providing outreach such as a need for additional training and support. EEOC headquarters officials identified multiple steps they take to provide training and support to OECs. For example, in fiscal year 2021, EEOC held an OEC conference that provided training on multiple topics including how to market outreach events.

Determine the Audit Objectives

The audit has two objectives. First objective is to see the steps that EEOC has taken to address its pending charge inventory since fiscal year 2011 along with the quality and length of its investigation process. Second objective is to evaluate its outreach efforts to employers and employees, and the challenges when providing outreach.

Define Audit Methods

To evaluate the appropriateness of the EEOC charge investigation process, the audit used two methods: one was to analyze the charge data collected, and the other was to interview with field officials for measuring outreach efforts.

Audit Steps, Methods, and Results

To address our objectives, we analyzed EEOC data on charges from FY 2011 through FY 2021. We also reviewed EEOC planning documents for the charge investigation and outreach processes, and guidance specific to the charge investigation process. Additionally, we assessed EEOC's actions against its internal guidance and agency wide goals, as outlined in its planning documents, as well as against federal internal control standards.

1.Collect and review charge data

We reviewed charge data from FY 2011 through FY 2021 to capture data from both before and after EEOC put a renewed emphasis on its Priority Charge Handling Procedures around 2014. EEOC provided us with direct access to EEOC's Enterprise Data Warehouse, known as the data enclave, which houses all of EEOC's charge data. The National Opinion Research Center (NORC) at the University of Chicago manages the data enclave. From NORC's database management system within the Electronic Data Warehouse, we accessed the data tables that had details about charges along with additional data tables that had charge related information on the charging party (the individual, agency, or organization who filed the charge), respondent (the entity against which a charge is filed), EEOC staff who processed the charge, and more. The data included inquiries or charges that opened or closed from FY 2011 through FY 2021. We filtered out any charges processed by Fair Employment Practices Agencies, rather than EEOC. We confirmed the accuracy of the charge data we downloaded by conferring with EEOC/NORC to verify the total counts for new charges, closed charges, pending charges, and initial inquiries.

In addition to conferring with NORC and EEOC, we assessed the reliability of these data by

- reviewing EEOC's documentation related to its stored data;
- interviewing officials from EEOC's Office of Enterprise Data and Analytics; and

- performing logic checks to identify obvious errors in the data, among other steps.

Based on the steps taken, we determined that the data included in our report were sufficiently reliable for our purposes.

2.Conduct descriptive analysis

Using the confirmed data, we conducted a descriptive analysis employing selective variables that reflect various charge-related characteristics. Many states and localities have their own laws prohibiting employment discrimination and their own agencies responsible for enforcing state and local laws, often referred to as "Fair Employment Practices Agencies." Among various metrics of analysis, we focused on the time for charge intake processing to see if EEOC improved the time taken to process a discrimination charge by which it reduced its pending charge inventory.

Analysis Results

EEOC has taken steps to reduce its pending charge inventory—in the past decade it has reduced its pending inventory by almost half—but demand for EEOC's services is high as evidenced by the more than 145,000 inquiries it received in fiscal year 2021. This demand underscores the importance of EEOC efficiently managing its inventory to carry out its mission. The length of the intake process warrants EEOC's attention since the amount of time it takes certain offices to complete the intake process may result in some individuals not taking the next step to file a discrimination charge. Monitoring office-level data on how long the intake process takes would allow headquarters officials to identify offices where the intake process takes longer and target intake related support, as needed, to these offices.

III. Analysis and Inspiration

The Chair of the EEOC should monitor field office-level data on the length of the intake process. For example, EEOC could monitor these data by including them in the monthly internal management reports it produces. By paying greater attention to the intake process as a result of monitoring these data, EEOC may be able to identify ways to make the process more efficient.

IV. Questions

1. What steps has the Equal Employment Opportunity Commission (EEOC) taken to address its pending charge inventory since fiscal year (FY) 2011, and what is known about the quality and length of its investigation process?

2. How EEOC evaluates its outreach efforts to employers and employees, and the challenges officials said they face when providing outreach?

(Written by Applied Research and Methods, U.S. Government Accountability Office)

02 Case Study of Text Analysisfor Performance Audits

I. Introduction

The Austrian Court of Audit is performing an audit of the Agency for European Integration and Economic Development (AEI). The focus of the audit is on the payment flows of the agency, which is in the spotlight in Austria because of its handling of EU funding. The core business of the AEI are so-called twinning projects, in which know-how from the Austrian administration is exported to other countries. The EU subsidizes those projects, the AEI is suspected of having misused the funds. Since its foundation in 2003 the agency carried out 250 international projects with a funding volume amounting to more than 220 million euros. Those projects are under investigation by multiple authorities. The defendants deny the allegations. By now, the agency has filed for bankruptcy.

However, the team of auditors is faced with a challenge that is common for a lot of audits: unstructured data in all forms and file formats in thousands of files. To mine this treasure trove of information, the Department for Digitalization, Information Technology and Data Analysis developed a search engine for auditors and used natural language processing to provide guidance.

II. Main Content

Objectives, Steps & Methods

The objective of the audit was to assess those five areas:

- legal basis, e.g. with regard to supervisory powers
- structure of AEI and its subsidiary

- operational business according to the distribution of roles between the association and the LLC
- economic situation of the association and the LLC as well as the identification of risk potentials
- role of selected ministries with focus on their project participation
- perception of control and steering possibilities in the bodies of the AEI by selected ministries
- In order to help to achieve those goals, the Department for Digitalization, Information Technology and Data Analysis setup a dedicated search engine for auditors to find relevant information efficiently.

For this the data analysts of the ACA used Apache Solr, which is an open-source search platform that allows you to create powerful search engines for applications It is based on the Apache Lucene library, which provides the core search functionality. Solr allows you to index and search various types of content, including text, documents, images. It also provides advanced features such as faceting, spell checking and highlighting. One of the advantages of Solr is its scalability and ability to handle large volumes of data. It can be deployed on a single server. It is therefore not necessary to upload confidential data to external servers. The crawl all of the data from thousands of documents, we relied on Apache Tika. It can automatically extract text and metadata from documents during indexing, making it possible to search for keywords and phrases contained within the documents.

After setting up Apache Solr, we built an interface for auditors in R using Shiny, which allows you to search in the documents. Furthermore, we implemented different types of searches to fulfill the needs of the auditors: queries that support wildcards and Boolean operators as well as an exact matching of a text block. In addition, the fact that the data quality provided was not consistently good was taken into account by incorporating a fuzzy search and a word root search. To integrate this tool into our existing application portfolio, we linked each result of the query to the original document in a different application where our auditors routinely review documents.

As a next step, further needs of the auditors were identified and these were translated into NLP tasks. To summarize the analysis, we can describe the tasks in following areas:

- **Named Entity Recognition:** This is a process of identifying and classifying named entities in unstructured data. Named entities refer to specific types of words or phrases that represent entities with proper names, such as people, organizations, locations or dates. We provided the auditors with a summary of the most important persons in the dataset and identified which organizations and locations occur.
- **Analysis of communication:** One part of the seized data concerned the e-mail correspondence of the key persons of the association. To get a better understanding how the AEI worked internally, we visualized communication flows. So auditors were able to see the most important persons of the network.
- **Topic Modelling:** The goal of topic modelling in general is to find groups of words that frequently appear together in different documents, and to assign these groups to specific topics. These topics can then be used to gain insights into the content of the documents and to organize them into different categories or themes. There are several methods for topic modelling, we used the Latent Dirichlet Allocation (LDA) algorithm. As a next step, we identified the most important document types, like protocols of assemblies and made it possible for auditors to only search in this part of the dataset.

III. Analysis Results

The audit of the AEI is ongoing, so it is not currently possible to share the auditors' results. But we have seen growing interest in the tool and how natural language processing can aid an audit from other auditors of the Austrian Court of Audit.

IV. Questions

The main problems that occurred during the natural language processing we related to data quality. It would be interesting to see approaches using novel techniques to improve the input. A major unresolved issue are handwritten notes. Making them searchable is a task that has not been solved yet. Furthermore, we're curious to use how sentiment analysis and machine translations using transformers could guide and help our audits.

03 Case Study of the Analysis of health data of military-aged men regarding the Body Mass Index

I. Introduction

In Austria the costs of treating chronic diseases place a growing burden on the healthcare system and increase every year, similar to most other countries around the world. An analysis for the year 2017 showed that approximately 60% of the costs of the regional health insurance funds in Austria are currently being spent on medical assistance, remedies, medical aids, transportation and inpatient costs for chronic diseases.

The causes of chronic diseases are diverse and often a combination of genetic, environmental and lifestyle-related factors. The main causes include an unhealthy lifestyle (such as an unbalanced diet, obesity, lack of physical activity, smoking, alcohol consumption), genetic predisposition, environmental factors, psychosocial factors, infections and aging. If the state invests more money in preventive measures to avoid chronic diseases, a significant amount of money can be saved in the public healthcare system.

As the Austrian Court of Audit already noted in 2014, the majority of public funds in the healthcare sector in Austria is allocated to curative care, while less than 2% were spent on health promotion and disease prevention.

II. Main Content

Determine the Audit Objectives

Data regarding the main causes for chronic diseases are needed, to implement a control system so that the government gets needed information for their decisions. In a first step we wanted to get data regarding the development of obesity over the years. The idea was that the needed data is acquired from the federal ministry of defence because each young man

has to make a medical check at the army, for the evaluation of their fitness. Each medical check documents beside other things the weight and height of each young man.

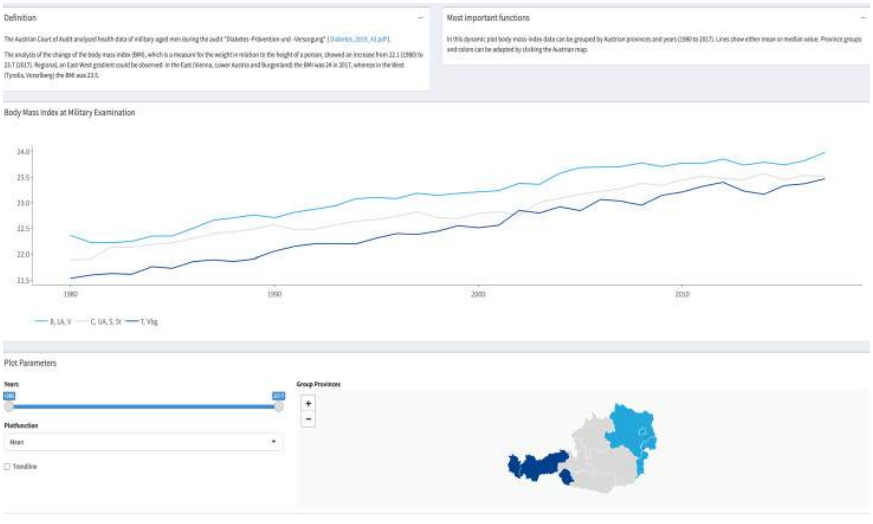
Define Audit Methods, Audit Steps and Methods

The body mass index can be calculated easily with the height and weight of each young man, therefor you get an impression of the increase of obesity over the years for the young male population in Austria.

Based on the ideas above, the auditors calculated the body mass index for all young man over a period of 38 years from 1980 to 2017. The data for the calculation came from the federal ministry of defence.

Analysis Results

The result of the data analysis were used for a visualization and a dashboard for decision makers like the minister of health, minister of education,….With the help of this dashboard, it is immediately apparent that the body mass index and, consequently, overweight rates among young men in Austria have increased significantly in all federal states from 1980 to 2017. Interestingly, the increase has been lower in the western part of Austria (dark blue in the graphic) compared to the eastern part (turquoise in the graphic). This is attributed to the fact that the western part of Austria is mountainous, while the eastern part is flat.



III. Analysis and Inspiration

With this audit the Austrian Court of Audit tried to show the public, that data analysis is very helpful for many areas in public administration.

For example if you want to reduce the spendings in the public health system you have to look in a first step what are the triggers for the costs. When you realize that a big part of the spendings in the public health system are used for chronic diseases, then you have to evaluate what are the causes for chronic diseases. When you realize that one cause for chronic diseases is obesity and that from 1980 to 2017 the body mass index for young man increased, then the minister for health and the minister for education and some other decision makers know that they have to take measures to stop this development. And also the public will understand these measures of the decision makers, if they see the visualizations and dashboards of the used data. Data analysis is an important and extremely useful tool if it's wisely used and properly defined.

IV. Questions

- *How can the public institutions be motivated to use and share their valuable data. For the decision makers it's quite important that they make decisions on data and facts.*

- *In the era of "fake news", more processed data from public institutions which are audited, reviewed and visualized are helpful and needed for the society and more transparency.*

(Written by Department of Data Analysis, Informationsecurity and new technologies, Audit Office of Austria)

04 Case Study of Payroll Oversight in Brazil

I. Introduction

The payment of salaries, pensions, and retirement benefits to public servants is a relevant concern in public expenditure in Brazil. The total expenditure on personnel in the federal public administration has averaged over $76.4 billion annually in the past 5 years, reaching 12.7% of GDP or 33.5% of public expenditure in the period from 2015 to 2021. However, the execution of these expenses is fragmented, with more than 12 million servants or pensioners belonging to over a hundred different careers in about 800 organizations throughout the country, without integrated data or systems. Monthly payments include components such as bonuses, allowances, function incorporations, and job accumulations that increase the complexity of the system, making it prone to errors, fraud, and inefficiencies, making the task of overseeing them a major challenge.

To cope with this task with a reduced team, an initiative has been implemented by TCU (Brazil's Court of Accounts) in recent years to proactively identify and correct potential irregularities continuously in each monthly payment cycle, using digital tools such as data cross-checking, automated analysis, dashboards, and computerized systems. Furthermore, a collaborative approach was adopted so that the organizations being audited could take timely preventive or corrective measures based on the detected evidence, reducing the risk and amount of irregular expenses.

II. Main Content

Determine the Audit Objectives

Payments of salaries, pensions, and retirement benefits are made to a significant number of beneficiaries monthly, for a diversity of careers each with their own regulations. Therefore, it is necessary to identify as soon as possible whether these payments are being made in compliance with applicable legislation so that undue expenses do not persist over time. When irregularities are identified, the responsible organizations must be notified to take the necessary measures and report the solution given to TCU.

Define Audit Methods

Payments to public servants and their pensioners are recorded in various information systems under the responsibility of public organizations within the jurisdiction of the TCU. Thus, it was possible to build a consolidated database that allowed an integrated view of the payments made collectively by the audited organizations. Based on this information, audit trails can be constructed to explore typical situations that may indicate evidence of irregularities in payments made by each organization. Through an online registration system, the evidence is forwarded to the organizations for clarification and necessary action, if needed. The actions taken are reported back to the TCU, which keeps track of the progress of all identified potential irregularities. In case of non-clarification of the irregularity by any organization after some time, the TCU may take the necessary measures.

Audit Steps, Methods, and Results

1. Data acquisition.

The data used by Continuous Auditing is obtained from a series of sources to be combined in a single repository at the TCU, called LabContas. The audited organizations monthly send data on payments for the immediately preceding month based on cooperation agreements. Information on job positions in organizations not covered by cooperation agreements is obtained from the annual databases of official government agencies. Information on local

and state payroll is also obtained through agreements with local control agencies.

2. Application of programs - audit trails - on the acquired data indicate the occurrence of irregularities.

Once the databases are ready, automated programs automatically cross-check the data to identify suspicious payments. The programs are developed by the audit team and refined as clarifications are provided by the audited organizations. Irregularities can arise for different reasons. There may be inconsistencies in data that should be consistent. There may be payments that do not comply with laws and regulations. There may be payments that are legal but are paid in duplicate. There may be legal payments, but the amount is incorrect or exceeds the allowed limit.

3. Sending evidence to the originating organizations for clarification.

Once the audit trails have been conducted on the data, indications of irregularities are forwarded to the organizations that executed the payments in order to seek clarifications on the identified situations. The notified organizations are required to investigate the situation described and provide a detailed account of the measures taken in response to the potential irregularity. Through an electronic system, the organization reports whether (i) the irregularity is valid and has been rectified; (ii) the irregularity is valid and measures to rectify it are currently underway; (iii) the irregularity is valid but cannot be rectified due to another ongoing process; (iv) the irregularity is invalid (false positive).

4. Analysis of clarifications.

The clarifications provided by the audited organizations are analyzed by the audit team on a regular basis. When the identified situations are resolved by the organizations or are determined to be false positives, the evidence is archived. Certain evidence is automatically archived without the need for human intervention when the audit trails identify that the issue has ceased to occur in subsequent months. Some evidence is returned to the originating organizations for additional clarifications.

5. Preparation of the final report.

The above steps occur monthly. At the conclusion of each audit cycle, which occurs at the end of each year, a report is generated presenting the key irregularities identified, the amount saved from the suspension of irregular payments, and the number of evidence that remain pending clarification for the following year.

Analysis Results

The data cross-check was able to identify much evidence of irregular payments and improper accumulations. Additionally, it was possible to identify the most frequent irregularities and organizations with the highest incidence of irregularities, and therefore, in need of improving their internal controls. It was also observed that irregularities, once identified, tend to consistently reduce over time due to the implementation of controls by the responsible organizations. The programs that execute the audit trails can generate many false positives. The quality of the audit trails depends on the quality of the data available for auditing.

III. Analysis and Inspiration

The highlight of the case study is the use of integrated data from different sources combined with the use of business intelligence to identify large-scale irregularities with a limited audit team. The cost of conducting the audit is low compared to the savings resulting from the initiative. The collaborative approach with the audited organizations for seeking clarification on irregularities has allowed for broadening the scope of results. A data-driven audit has proven to be an effective mechanism for timely identification and resolution of irregularities with a significant impact on public personnel expenses.

IV. Questions

1. How can the coverage and consistency of acquired data be increased to meet the objectives of the audit?

2. How can we measure the appropriate level of delivering new audit trails to detect more evidence of irregularities while ensuring that organizations are not overwhelmed and are able to provide clarification on the most significant evidence?

3. The data cross-referencing and collaborative work carried out by the initiative can identify and address irregularities on an individualized and large-scale basis, but how can structural changes be promoted to address the root causes of common problems?

(Written by Audit Department for Personnel, Brazil's Court of Accounts)

05 Report on the Use of Artificial Intelligence in the Federal Public Administration

I. Introduction

There is a risk that Artificial Intelligence Technology will not be leveraged by the government to enhance digital services. It is also understood that AI technology should be strategically addressed at the national level in view of Brazil's competitiveness vis-à-vis other nations. On the other hand, the excitement generated by new technologies, such as Artificial Intelligence, can also be a source of waste of public money when the technology is not understood by managers and uncertainties are not considered.

The guidelines issued by the Digital Government Secretariat indicate the importance of the government promoting the digital economy, strongly based on data, so that the use of intelligence generates more and more opportunities, with new business models, in a world where all devices will be connected, producing and consuming data and generating wealth. In this way, it is seen that the Federal Public Administration tends, in the short and medium term, to use solutions that involve Artificial Intelligence technologies to deliver public services of better quality and efficiency to the population.

The prospect is that, in the short term, there will be a need for the Federal Public Administration to carry out highly material contracts involving Artificial Intelligence technologies, which must be duly monitored by the TCU. In this way, it is essential to know the main risks and benefits brought by Artificial Intelligence technologies, as well as the new challenges imposed for external control.

II. Main Content

Determine the Audit Objectives

The objective of the work was to carry out a study about the current stage of the use of AI technologies by several organizations of the Federal Public Administration (FPA), under many aspects.

The first objective refers to learning about the main technologies used in the development of AI solutions, and evaluating their current stage, as well as the perspectives regarding their use in public agencies in addition to identifying the main risks associated with the matter.

The second objective refers to learning about the current stage of development and the proposal content for a Brazilian Artificial Intelligence Strategy, as well as identifying the main risks and opportunities associated with its implementation.

The scope of this work comprised the use of AI in the organizations of the three branches of the Republic, but only the aspects related to AI as machine learning technologies have been assessed. Solutions developed by organizations in the private sector were out of scope.

Finally, the audit aimed to verify how the control of AI solutions within the FPA may be done, and how AI technologies may serve as support to the control activity.

Define Audit Methods

The following audit questions have been made:

a. What are the main AI technologies and the benefits to the FPA when deciding to apply them?

b. What is the current scenario when it comes to using AI in the FPA?

c. What are the challenges and risks that should be considered by managers when deploying AI solutions in the FPA?

d. What is the current stage of implementing the Brazilian AI Strategy and the risks and opportunities regarding that implementation?

e. What methodologies may be applied by the audit teams to audit AI solutions in the FPA?

f. How can AI be used to assist with the control activity?

Audit Steps, Methods, and Results

First, it should be noted that the work has been done according to the Auditing Standards of the Federal Court of Accounts (NAT).

This audit has been made through teleworking modality, according to the better practices defined by the TCU for its virtual audits using the tools already available.

In order to refine the planning matrix prepared by the audit team, there have been interviews with experts, and suggestions for audit matters have been collected, as well as the identification of relevant organizations regarding the use of AI in the country.

The audit was based on interviews with managers aiming to learn about the main cases related to the aforementioned technology use and features in the public sector. In addition, it has been verified the existence of works carried out by several international organizations, and they comprise many of the objectives of this study, serving as an important source for this work. Literature related to the topic has also been used, as well as academic sources, according to those in the section named References.

Furthermore, the audit team has created an electronic questionnaire sent to 293 organizations as part of the scope, aiming to gather details on how AI is used to solve business problems and add value to services provided to citizens.

The team has also interviewed experts in the field, both in the academic and the private sectors, in order to find assistance in preparing the information found in this study.

III. Analysis and Inspiration

This survey was proposed with the general objective of learning about the current use of Artificial Intelligence technologies in the Brazilian Public Administration, to evaluate the risks and opportunities of the Brazilian Artificial Intelligence Strategy, and verifying the main aspects related to the use of AI technologies in control activities.

Regarding the use of AI in the Federal Public Administration, the work covered the main benefits of using AI solutions in the Public Sector, among them the reduction of expenses as a result of greater operational efficiency, the improvement of services available to citizens, and the availability of new services, in addition to advantages of using data for decision-making by government entities.

In this context, the study has found out that in Brazil, despite the country presenting competitive differentials that encourage greater implementation and AI solutions, such as the great availability and representativeness of data, the Brazilian Public Sector still lacks better policies and investments to improve its infrastructure, human capital, and innovation capacity.

In order to obtain a comprehensive overview of the current state of AI use in the FPA organizations, a survey was conducted by means of an electronic questionnaire that was answered by 263 organizations. Among the main findings from the analysis of the answers is the fact that most institutions are still not planning to use AI solutions in their production processes or in the provision of services to citizens. Among the main obstacles reported for the adoption of the technology, the low number of available employees with sufficient training for implementation was reported.

The survey also showed that most of the solutions are developed by the agencies' own teams, and 48% of the organizations surveyed have not carried out any internal training regarding the development and implementation of AI technologies.

We also found out that the Judiciary Branch is ahead in relation to the use of AI solutions, which has benefited and contributed mainly to decreasing the case backlog of that Branch. The Judiciary is also implementing strategies and infrastructures that allow the sharing of data and the exchange of solutions among its various bodies.

Among the main benefits pointed out with the use of AI, the organizations surveyed reported the automation of repetitive processes with increased productivity, the reduction of time and costs in the execution of processes, the optimization in the allocation of human resources, increased reliability and transparency of processes, and greater timeliness in responding to customers.

In relation to the Brazilian Artificial Intelligence Strategy (EBIA), the way some nations have dealt with the subject, both nationally and regionally, with the objective of becoming leaders in the segment was presented.

The assessment of the Brazilian Artificial Intelligence Strategy (EBIA), based on the Guidelines for the Control of Public Policies prepared by the TCU, revealed flaws in its preparation that could compromise its proper implementation, monitoring, and achievement of objectives. The following findings were identified in relation to EBIA:

a. EBIA objectives are not specific, measurable, attainable and time-bound;

b. there is no clear definition of the initial reference stage ("ground zero", "initial situation") and no definition of indicators and targets to support assessment of EBIA's performance and results;

c. formal flaws in the presentation of the logic model and how EBIA impacts problems/ opportunities, projecting its results and long-term impacts;

d. the governance and management framework needed to implement the policy are not formally institutionalized in EBIA;

e. EBIA's monitoring and evaluation framework is not formally defined in the norm.

The Study addressed the use of AI in the control activity, identifying cases of use of the solution in Public Administration organizations. The work also addressed the control of AI solutions by the Public Authorities, addressing the main references already produced on the subject. Because of the various specific characteristics that AI solutions present in relation to other computer systems, we understand that a specific training plan should be prepared on the subject for the TCU's technical staff, especially the Information Technology Audit Secretariat.

IV. Questions

1. How to define the best solutions for sharing data from jurisdictional units in the audit activity?

2. How to assess the quality of the data made available by the jurisdictional units so that reliable conclusions can be drawn based on these data?

3. How to evaluate the consequences of data quality in the implementation of AI solutions by the jurisdictional units?

(Written by the Audit Department for Information Technology, Federal Court of Accounts, Brazil)

06 Case Study of Continuous Supervision of Social Benefits in Brazil

I. Introduction

Since 2015, the Continuous Oversight of Benefits (FCB) process has been an essential instrument for the Tribunal de Contas da União (TCU) in auditing the granting, payment, and maintenance of social benefits. Operating in a scenario that sees over 30 million payments made monthly and managing databases with hundreds of millions of records, the FCB has been crucial in promoting transparency and accountability. This process covers the three major social policies in Brazil: pension, social assistance, and labor benefits. According to the latest figures, the pension policy currently serves approximately 40 million beneficiaries, social assistance supports around 20 million families, and labor benefits reach nearly 8 million workers, underscoring the vast scope of TCU's oversight.

The integration of advanced information technologies and automated processes into the FCB framework has significantly enhanced TCU's ability to detect irregularities and optimize internal controls. By conducting comprehensive quality assessments and developing risk-based typologies, the FCB process has already uncovered hundreds of thousands of instances of improperly disbursed benefits. This approach has led to adjustments in internal procedures, reducing undue payments. Furthermore, investments in infrastructure, such as the creation of LabContas — a centralized repository for audit-related databases — enable monthly reviews that provide a complete view of payment flows. In an era of rapid technological evolution, these measures ensure that TCU remains at the forefront of public accountability and fiscal responsibility across Brazil's key social benefit policies.

II. Main Content

Data Acquisition

The FCB process obtains its data through formal agreements with the entities under its jurisdiction and leverages the robust infrastructure provided by the LabContas system. These established agreements ensure that auditors receive the required datasets on a regular basis. LabContas serves as a centralized repository that collects and stores complex data from multiple sources, making it available for routine audit tasks. This arrangement guarantees that the data is acquired in a timely, consistent manner and supports the systematic execution of audit procedures. By integrating these mechanisms, the process handles large volumes of detailed information without interruption, thus ensuring that data-driven analyses can be performed efficiently.

Data Quality

Public databases in Brazil exhibit varying levels of quality, which can directly impact the reliability of audit outcomes. To address this, the FCB process emphasizes the selection, validation, and maintenance of high-quality data. Objective criteria are established based on international standards, such as those outlined in DAMA DMBok and ISO, to ensure consistency and accuracy. Regular assessments are performed to verify data accuracy and completeness, which is essential for minimizing false positives and enhancing detection capabilities. Additionally, when significant quality issues are detected during these evaluations, the process triggers the creation of specific registration typologies. These typologies serve as systematic records that detail the nature of the quality problems, providing a clear basis for corrective actions and further analysis.

Analysis

In the analysis phase, the process utilizes the comprehensive datasets to develop detailed typologies that reveal a high volume of audit findings. Rather than relying on statistical samples, the methodology examines entire populations, ensuring that every potential irregularity is captured. This exhaustive approach significantly enhances detection

power and provides a complete picture of the payment processes under review. Data is systematically processed and reviewed to identify patterns that indicate weaknesses in internal controls or instances of improper disbursement. The results of these analyses not only highlight areas for improvement within the audited entities but also drive the refinement of internal procedures. This focused analysis directly contributes to the ongoing enhancement of audit practices and supports the broader goal of reducing undue payments.

Verification of Findings

The verification phase is dedicated to confirming the accuracy of the findings identified during the analysis stage, with a strong focus on reducing false positives. This step is particularly important for minimizing the negative impact of inaccurate claims on vulnerable populations. Auditors collaborate closely with the entities under oversight, engaging in a thorough review of each flagged case to ensure that the findings are correct and substantiated. The process employs cross-checking methods that compare data from multiple sources, thereby validating the initial findings and confirming their reliability. By working in tandem with the jurisdictional entities, auditors ensure that any discrepancies are promptly addressed, which not only reinforces the credibility of the audit but also assists the entities in making necessary adjustments to their internal controls.

Continuous Monitoring

Continuous monitoring is implemented as an integral part of the FCB process to maintain the effectiveness of internal controls over time. This approach involves evaluating the developed typologies in successive audit cycles, which demonstrates the ongoing impact of external oversight on the work processes of managers. Regular monitoring allows auditors to track trends, assess the effectiveness of control measures, and quickly identify emerging risks or issues. Through these consecutive cycles, any changes in data quality or process performance are detected early, allowing for timely interventions and adjustments. This systematic monitoring not only reinforces the overall oversight framework but also ensures that improvements are sustained, providing a consistent check on the effectiveness of the entities' internal controls and contributing to long-term fiscal responsibility.

III. Analysis and Inspiration

The FCB process offers a robust framework for progressively enhancing detection methods within the auditing of social benefits. Initially, the process ensures secure data access and confirms that the foundational datasets are both complete and reliable. It then advances to conducting substantive tests and evaluating data quality using objective criteria aligned with international standards to minimize false positives. When significant quality issues are detected, auditors create business typologies that classify and prioritize risks, resulting in data-driven findings that guide further investigations. Additionally, advanced data analysis techniques, including data mining and AI-driven algorithms, can be integrated to broaden the detection of risks and improper payments. This structured approach allows auditors to pinpoint irregularities at various stages—from simple access issues to complex deficiencies in data quality and potential fraud—thereby enhancing overall oversight and accountability.

IV. Questions

Despite the significant achievements of the FCB process, several critical questions remain to be addressed: How can emerging technologies such as Large Language Models (LLMs) and Generative AI (IAG) be effectively integrated into audit processes to further enhance the detection of irregularities while ensuring transparency? What strategies can be employed to incorporate Explainable AI into our systems so that the decision-making processes of complex algorithms remain clear and understandable to all stakeholders? In what ways can the integration of emerging technologies into audit processes be aligned with existing regulatory requirements and data privacy standards to maintain legal and ethical integrity? What specific training and capacity-building initiatives should be implemented to ensure that audit teams are fully equipped to manage, interpret, and act upon insights generated by advanced data mining and AI tools?

(Written by Audit Department for Social Benefits, Brazil's Court of Accounts)

07 Case study: Data Analysis on Mining Tailings through Satellite Imagery Coquimbo Region

I. Introduction

The mining tailings found in the Coquimbo region correspond to 51,05 % of national deposits, according to the information provided on the website of the National Geology and Mining Service (SERNAGEOMIN). The report also implies a high number of abandoned and inactive tailings deposits present a high level of environmental contamination. According to the study from the University of Chile (2018); “Evaluation of the acute effects of continuous exposure to particulate matter from mining tailings on the respiratory health of schoolchildren in the city of Chañaral”, which presented several respiratory and neurological diseases in the surrounding population due to water contamination with heavy metals, high concentration of mercury in soils and vast areas saturated by particulate matter.

In 2010, an assessment study of a contaminated site conducted by the National Environmental Center (CENMA) of the University of Chile detected the presence of mercury in 18 tailings within the urban radius of Andacollo. The analyses revealed several points with concentrations above 10 mg/kg, and the highest sample measured was 30.34 mg/kg.

Furthermore, the Final Report No. 168 of 2021 of the Comptroller General Republic of Chile (CGR) found, among other relevant aspects, that both the Ministry of the Environment of Chile and SERNAGEOMIN have not implemented actions related to the management of human health risk. In this context, lacking instruments that govern and contain management strategies for reducing or eliminating the significant risks caused by such structures, the adverse effects on human populations and the environment will continue.

Within this framework, an audit was conducted to take into consideration the social and environmental impacts of mining tailings and the risks detected during the planning stage related to deficient oversight, lack of coordination, and legal gaps in current legislation, among others.

In addition, it is relevant to mention that the results of the regional anti-corruption survey carried out in 2020 by the CGR of Chile, reported that 34% of respondents in the Coquimbo region believe that corruption is concentrated in the audits carried out by public services. The following situation will be addressed in the audit under study since one of the objectives will be to verify SERNAGEOMIN's compliance with its auditing functions, as well as to analyze the measures adopted by the Service based on the audits carried out in recent years.

Regarding the scope of intensifying the use of massive data for the audit of the CGR of Chile, the present audit had the support of the Unit of Massive Analyses for Audit, the latter will provide the inputs to counteract the compliance under the active regulations. The UMAA using georeferencing will allow obtaining cartographic information on the position of mining tailings of diverse origins in the communes of Illapel, Andacollo, and La Higuera.

Through this audit, the CGR of Chile seeks to contribute to the Implementation and fulfillment of the 17 Sustainable Development Goals (SDGs), approved by the United Nations General Assembly in its 2030 Agenda for the eradication of poverty, protection of the planet, and the prosperity of all humanity.

This review is framed within the framework of the 16th SDG, Peace, Justice, and Solid Institutions.

II. Main Content

Determining the Audit's scope

Scope

Conduct and audit the actions and measures adopted by competent institutions in relation to the management of mining tailings deposits located in the municipalities of Andacollo, La Higuera, and Illapel.

Specific objectives

- Verify that the audited institutions have ensured that the tailings comply with applicable and established technical conditions, according to their competencies.
- Corroborate whether the agencies of environmental jurisdiction over the matter have coordinated adequately and promptly.
- Examine the accounts related to the collection and payments of fines associated with the audits carried out. To verify that they comply with the principle of legality and completeness and are duly accredited.

Define Auditing Methods

The National Geology and Mining Service will be audited regarding the evaluation, construction, and closure of tailings deposit projects in the Coquimbo region.

Normative:

- The general regulatory framework is reviewed in the areas of mining, water, and environmental impact assessment.
- Review specific regulations: mining and water.

Risk assessment:

- The institution does not carry out audits of active tailings deposits.
- The service does not comply with the collection of penalties.
- The service approves tailing deposits that do not comply with the established requirements.

To counteract compliance in accordance with the regulations established in the files and the associated risk assessments, georeferencing will allow obtaining cartographic information on the position of mining tailings from different sources.

The following is an explanation of the data analysis performed:

Methodology for Data Analysis

Tailings identification was based on the analysis of free satellite images from the Copernicus Sentinel 2B satellite[1] of the European Space Agency (ESA), which offered a resolution of 10 meters (minimum viewing area).

The platform used for the analysis was Google Earth Engine, a mosaic of images configured from January 1 to December 31, 2021[2].

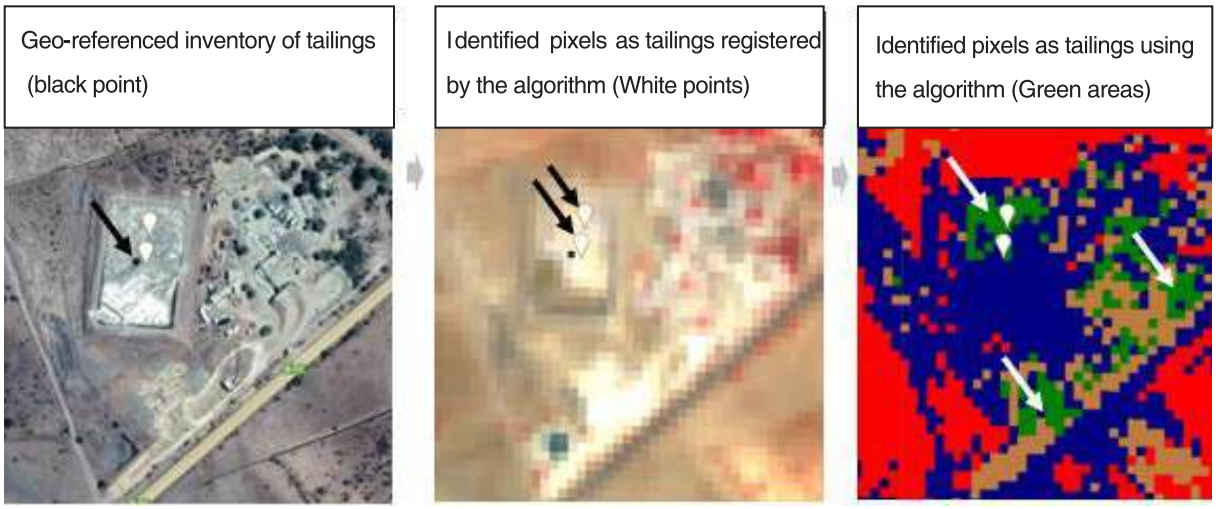
The Machine Learning technique used for pixel classification was called Random Forest, a supervised classification algorithm, and the samples were defined from the georeferencing of the 389 tailings identified in the SERNAGEOMIN Data Base.

In the supervised identification process, the RandomForest algorithm identified pixels with similar values which defined these pixels with specific reflectance for tailings in identified bands. The following figure shows the procedure graphically.

[1] The RGB layout used two combinations of bands: Conventional False Color (B8, B4 and B3), and a combination for identification of differences in rock (B12, B4 and B2).
[2] The available images selected were those with a maximum cloudiness of 10%.

Figure N°1

Pixel Identification using the Random Forest supervised classification algorithm.



Source: Own elaboration.

The resulting images (third image in figure N°1) were vectorized and subsequently filtered to obtain only those areas larger than 2,000 meters.

Finally, a visual review of the resulting areas was performed to determine the possible findings inside and outside the areas.

Possible tailings identified should ideally be confirmed by field visits.

In addition, there may potentially be more tailings than those initially reported, given the limitations of the imagining and technique used.

Data Analysis results

The analysis of the data identified 114 possible tailings that are not inventoried in the 2020 registry, which can be viewed on a map generated by Google, with the following link: https://www.google.com/maps/d/u/0/edit?mid=1MbAD3udF5e2AiQwtdfuaguFoqBaOhzg&usp=sharing

The registered tailings, the possible identified tailings, the area identified through the algorithm, and the communal limits can be visualized through the map of the findings.

To facilitate confirmation in the field and to prevent the environmental impact that results from poor tailing management. The findings have been prioritized into three groups, according to their distance from population centers and their North and East coordinates.

They are listed below (Datum WGS84 Huso 19S):

1. **Possible tailings located within 2,500 meters of a population center:**
   31 cases are concentrated in the localities of Andacollo, Vicuña, and La Serena.
2. **Possible Tailings located within 5000 meters of a population center:**
   13 cases are primarily located in the localities of Rio Hurtado and Punitaqui,
3. **Possible Tailings located more than 5000 meters away from a population center.**
   70 cases are in several localities, especially Los Vilos.

However, it is important to note that the algorithm of this data analysis obtained from a Pre-Audit process will be adjusted based on the observations from the field tests, to optimize the control points of the risk assessment process, for example, the analysis revealed that, given the characteristics of the study, there were challenges corresponding to "leach stockpiles" and "irrigation pools".

Steps, Methods, and Audit Results

As part of the inspection procedures, various tests have been carried out with the aim of validating compliance with administrative and mining regulations by the audited Entity. These tests implied carrying out various validation activities where the contrasting documents are included and verifying in the field the material presence of what is indicated in the documentation.

By virtue of the results of the audit, it is evident that the Entity does not ensure that all the tailings deposits are included in its cadastre, in particular, a tailing was identified in the

commune of La Serena not registered by the Entity, with spill risk.

III. Analysis and Projection

The present case highlights the transformation of the auditing paradigm in the CGR of Chile utilizing unstructured data and satellite images to contrast compliance with the normative standards established on the official documentation. The preliminary results of the audit process evidence that by using big data techniques, the audit's scope becomes more effective, the resources are used more efficiently, and the audit becomes more fruitful.

IV. Questions

1. To refine the results of algorithms for mine tailing identification. Which indicator and/or data should be included in the analyses?

08 Case Study: Asset variations analysis of Public Officials in Chile

I. Introduction

The Assets Evolution and Declarations Department (DEPyD by its acronym in spanish) of the Inspection Division of the Supreme Audit Institution of Chile has been developing a new product to analyse the growth of public servant's assets in search of anomalies. It has elaborated advanced tools to strengthen the analysis related to the assets and wealth of public officials, as a support tool for SAI Chile auditors, in compliance with Law 20.880, on Probity in Public Function and Prevention of Conflicts of Interest. This legislation requires a public servant to declare his or her assets and liabilities, under certain conditions established by the law.

In addition, this effort is in line with Sustainable Development Goal 16 (SDG 16): "Promote peaceful and inclusive societies for sustainable development, ensure access to justice for all, and build effective, accountable and inclusive institutions at all levels". This case study focuses on the use of patrimonial analysis methods and technological tools to identify inconsistencies, promote transparency, and combat fraud.

II. Main Content

Objectives

The primary goal is to identify potential red flags in the evolution of public servant's wealth. Thanks to the automation of processes, those analysis can be speeded up. A final report is made after extracting and analysing the acquisitions of public officials and their families (vehicles, real estate, investments, etc.) and, if required by law, their declared data. Other

general objectives are:

- **Increase transparency:** A public version of the declarations is available to citizens to ensure trust and transparency.
- **Prevention of Conflicts of Interest:** By identifying potential overlaps between public responsibilities and private interests, the analysis helps mitigate risks.

Methodology

This is an internal product oriented to detect and eventually prevent fraud, and also can be generated at the request of other offices or departments conducting audits. The process is based on a thorough analysis of individual public officials, through mass data extraction from different sources, following a series of agreements made by SAI Chile with some key public entities.

All these processes are automated by multiple codes written in Python programming language, which are responsible for the extraction and calculations to create a final spreadsheet with detailed information on the individual's wealth.

Then, another semi-automated Python code creates a report in word processor format to improve efficiency by saving time and help analysts with the redaction the final product.

The codes and algorithms involved in the process consider several steps:

1. Verification of Official Data:

Ensuring that the national individual's identification number (RUT) of the subject is accurate, to proceed with the consequent analysis.

2. Data Extraction (four main sources):

2.1 Internal Revenue Service (SII):

One of the Python codes automatically extracts several key pieces of information, in the form of vectors, about the financial and asset movements of the investigated public officials. These include the purchase and sale of vehicles, both new and used, real estate transactions, whether purchased on credit or cash, as well as sales of real estate, are also examined to identify any discrepancies. Sources of income are detailed along with payments from companies for salaries, professional fees, and retirement benefits. Financial investments such as stock purchases, foreign currency purchases, and mutual fund activity (both investments and redemptions) are carefully analyzed to uncover patterns that may indicate potential inconsistencies or unusual asset growth.

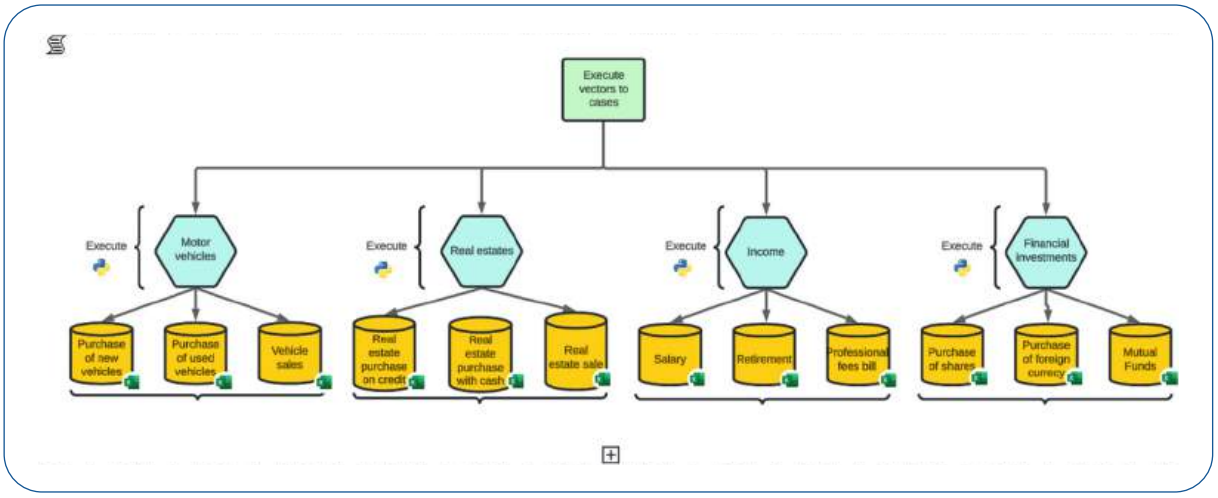


Figure 1: Vectors extracted from Internal Revenue Service
Source: Prepared by the author

2.2 Business Participation:

Public officials may participate in the creation of commercial companies, which can be sources of potential conflicts of interest. In this section we can see the information of the company name, individual identification number, date of its creation, date of the start of commercial activities, names and identification numbers of those who conform the company and their participation percentages.

The process can take companies identification numbers and run a SQL code that allows to check if there is any purchase order associated with these companies and to which public institution is related. It can also get the date of the order, the amounts and the modality of the official announcement and allocation of funds (public competition, direct allocation, bidding), among others.

3. Family networks:

Through the Civil Registration and Identification Service (SRCeI) data, the public agency responsible for maintaining records of births, marriages, deaths, and other civil status information, is used a Python code to extract information about family networks through a series of iterations. Then, using a R language developed code, the relationships up to the level of iterations required are built. This process allows to identify the most direct family network of the subject. Once this data is obtained, the process cross-reference the family network through the Internal Revenue Service (SII) to also analyse their assets, in a similar way to what was done with the public official, looking the identify potential patterns or anomalies across the family.

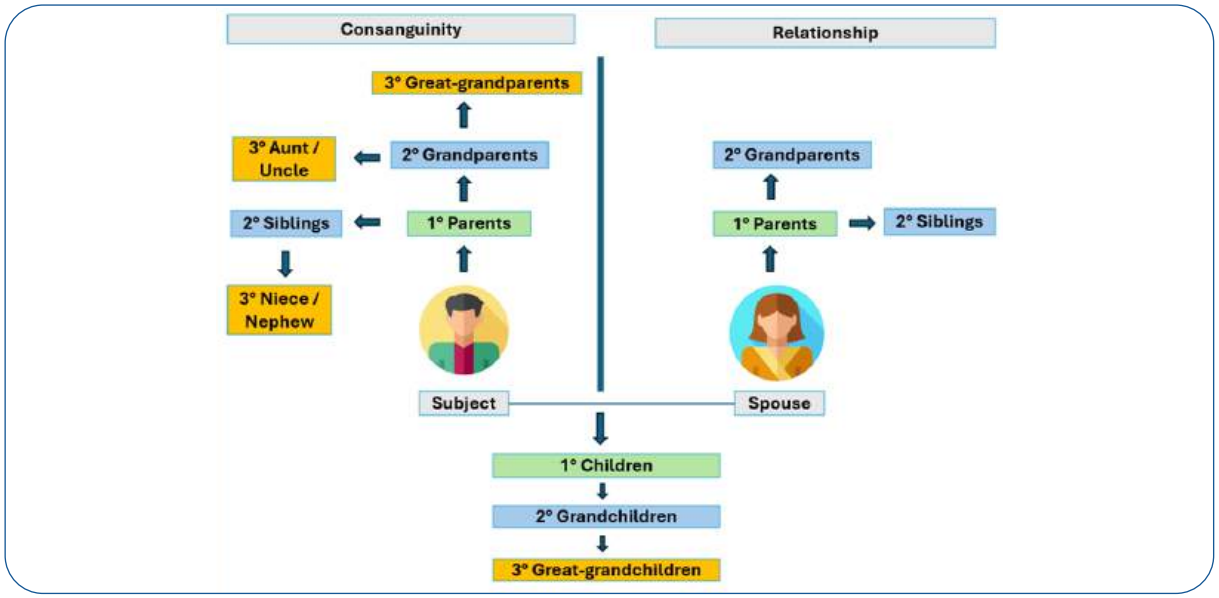


Figure 2: Family Networks
Source: Prepared by the author

4. Obtaining Details of Interest and Asset Declarations:

A Python code automatically extracts the necessary information from all declarations made by the public official. This process summarizes the most relevant data and prepares it for inclusion in the report, ensuring that critical insights are easily accessible for analysis and presentation.

III. Results

The result is a comprehensive profile of the subject and his or her family network, incorporating all the data gathered from the sources previously described. This includes detailed information from asset declarations and Internal Revenue Service (SII) records. In addition, the report integrates information on potential inheritances that may affect the subject's wealth status. Only public officials who require further investigation are profiled.

The report is constantly evolving, with plans to incorporate additional variables through new agreements between SAI Chile and public agencies and expanded data-sharing initiatives. During the current year, a total of **445** public officials and **2.225** of theirs family members has been analyzed, ever since the first developments in March 2024, resulting in **232** suspicious cases.

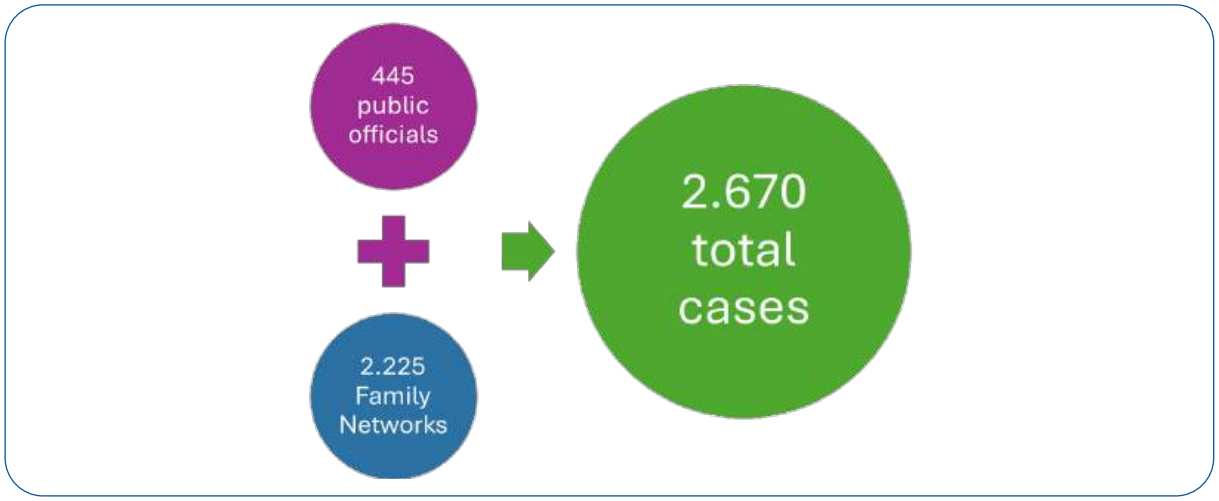
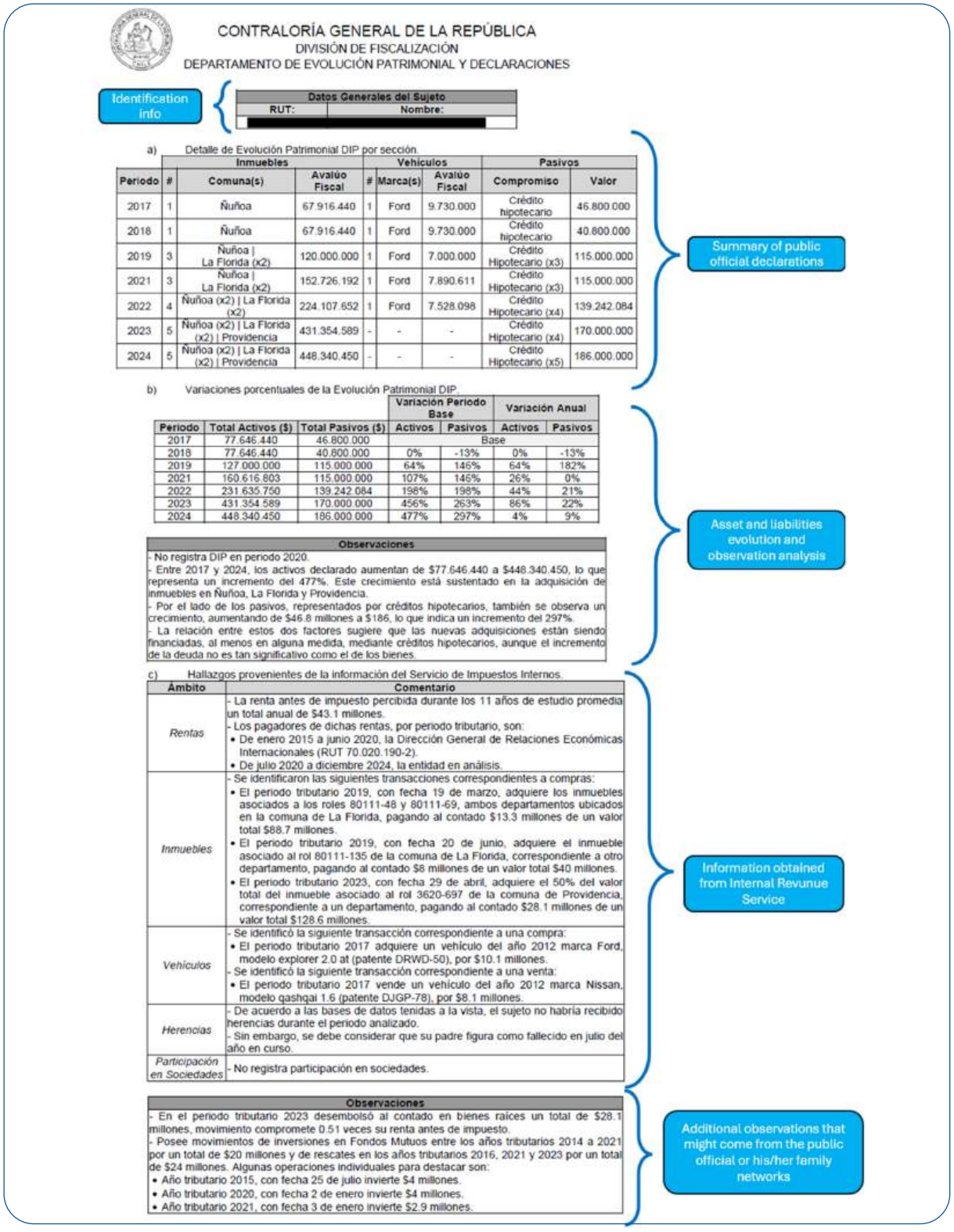


Figure 3: Final report after automated processes plus analyst observations.
Source: Prepared by the author



CONTRALORÍA GENERAL DE LA REPÚBLICA
DIVISIÓN DE FISCALIZACIÓN
DEPARTAMENTO DE EVOLUCIÓN PATRIMONIAL Y DECLARACIONES

Identification info

Datos Generales del Sujeto	
RUT:	Nombre:

a) Detalle de Evolución Patrimonial DIP por sección.

	Inmuebles			Vehículos			Pasivos	
Periodo	#	Comuna(s)	Avalúo Fiscal	#	Marca(s)	Avalúo Fiscal	Compromiso	Valor
2017	1	Ñuñoa	67.916.440	1	Ford	9.730.000	Crédito hipotecario	46.800.000
2018	1	Ñuñoa	67.916.440	1	Ford	9.730.000	Crédito hipotecario	40.800.000
2019	3	Ñuñoa | La Florida (x2)	120.000.000	1	Ford	7.000.000	Crédito Hipotecario (x3)	115.000.000
2021	3	Ñuñoa | La Florida (x2)	152.726.192	1	Ford	7.890.611	Crédito Hipotecario (x3)	115.000.000
2022	4	Ñuñoa (x2) | La Florida (x2)	224.107.652	1	Ford	7.528.098	Crédito Hipotecario (x4)	139.242.084
2023	5	Ñuñoa (x2) | La Florida (x2) | Providencia	431.354.589	-	-	-	Crédito Hipotecario (x4)	170.000.000
2024	5	Ñuñoa (x2) | La Florida (x2) | Providencia	448.340.450	-	-	-	Crédito Hipotecario (x5)	186.000.000

Summary of public official declarations

b) Variaciones porcentuales de la Evolución Patrimonial DIP.

			Variación Periodo Base		Variación Anual	
Periodo	Total Activos ($)	Total Pasivos ($)	Activos	Pasivos	Activos	Pasivos
2017	77.646.440	46.800.000	Base			
2018	77.646.440	40.800.000	0%	-13%	0%	-13%
2019	127.000.000	115.000.000	64%	146%	64%	182%
2021	160.616.803	115.000.000	107%	146%	26%	0%
2022	231.635.750	139.242.084	198%	198%	44%	21%
2023	431.354.589	170.000.000	456%	263%	86%	22%
2024	448.340.450	186.000.000	477%	297%	4%	9%

Observaciones
- No registra DIP en periodo 2020. - Entre 2017 y 2024, los activos declarado aumentan de $77.646.440 a $448.340.450, lo que representa un incremento del 477%. Este crecimiento está sustentado en la adquisición de inmuebles en Ñuñoa, La Florida y Providencia. - Por el lado de los pasivos, representados por créditos hipotecarios, también se observa un crecimiento, aumentando de $46.8 millones a $186, lo que indica un incremento del 297%. - La relación entre estos dos factores sugiere que las nuevas adquisiciones están siendo financiadas, al menos en alguna medida, mediante créditos hipotecarios, aunque el incremento de la deuda no es tan significativo como el de los bienes.

Asset and liabilities evolution and observation analysis

c) Hallazgos provenientes de la información del Servicio de Impuestos Internos.

Ámbito	Comentario
Rentas	- La renta antes de impuesto percibida durante los 11 años de estudio promedia un total anual de $43.1 millones. - Los pagadores de dichas rentas, por periodo tributario, son: • De enero 2015 a junio 2020, la Dirección General de Relaciones Económicas Internacionales (RUT 70.020.190-2). • De julio 2020 a diciembre 2024, la entidad en análisis.
Inmuebles	- Se identificaron las siguientes transacciones correspondientes a compras: • El periodo tributario 2019, con fecha 19 de marzo, adquiere los inmuebles asociados a los roles 80111-48 y 80111-69, ambos departamentos ubicados en la comuna de La Florida, pagando al contado $13.3 millones de un valor total $88.7 millones. • El periodo tributario 2019, con fecha 20 de junio, adquiere el inmueble asociado al rol 80111-135 de la comuna de La Florida, correspondiente a otro departamento, pagando al contado $8 millones de un valor total $40 millones. • El periodo tributario 2023, con fecha 29 de abril, adquiere el 50% del valor total del inmueble asociado al rol 3620-697 de la comuna de Providencia, correspondiente a un departamento, pagando al contado $28.1 millones de un valor total $128.6 millones.
Vehículos	- Se identificó la siguiente transacción correspondiente a una compra: • El periodo tributario 2017 adquiere un vehículo del año 2012 marca Ford, modelo explorer 2.0 at (patente DRWD-50), por $10.1 millones. - Se identificó la siguiente transacción correspondiente a una venta: • El periodo tributario 2017 vende un vehículo del año 2012 marca Nissan, modelo qashqai 1.6 (patente DJGP-78), por $8.1 millones.
Herencias	- De acuerdo a las bases de datos tenidas a la vista, el sujeto no habría recibido herencias durante el periodo analizado. - Sin embargo, se debe considerar que su padre figura como fallecido en julio del año en curso.
Participación en Sociedades	- No registra participación en sociedades.

Information obtained from Internal Revunue Service

Observaciones
- En el periodo tributario 2023 desembolsó al contado en bienes raíces un total de $28.1 millones, movimiento compromete 0.51 veces su renta antes de impuesto. - Posee movimientos de inversiones en Fondos Mutuos entre los años tributarios 2014 a 2021 por un total de $20 millones y de rescates en los años tributarios 2016, 2021 y 2023 por un total de $24 millones. Algunas operaciones individuales para destacar son: • Año tributario 2015, con fecha 25 de julio invierte $4 millones. • Año tributario 2020, con fecha 2 de enero invierte $4 millones. • Año tributario 2021, con fecha 3 de enero invierte $2.9 millones.

Additional observations that might come from the public official or his/her family networks

Source: Prepared by the author

IV. Analysis and Inspiration

This case study illustrates how the use of massive data analysis technologies and systematic methodologies strengthens public oversight. In recent years, a series of press reports have highlighted the increase in the wealth of various public officials, raising public concern and calling for greater scrutiny by regulatory bodies. These situations reinforce the need for technological tools that can detect anomalous patterns in the acquisition of goods and the management of resources.

In addition, the process is constantly being improved to incorporate more advanced analytical tools, including emerging technologies such as machine learning, which could facilitate real-time monitoring of anomalous asset movements.

V. Questions

1. What additional strategies could be implemented to improve real-time data integration?

2. What are the main ethical and cybersecurity challenges in using mass data in asset audits?

3. How can automated systems be further refined to help analysts make faster and more accurate decisions in differentiating legitimate anomalies from potential fraud?

(Written by the Department of Assets Evolution and Declarations, Office of the Comptroller General of the Republic of Chile)

09 Case Study of Mask Sales Data Analysis During the Pandemic in China

I. Introduction

The infection prevention and control of COVID-19 is the top priority in China,which concerns the safety and health of the people and the overall economic and social development.In the first days of the pandemic,there were acute shortages of medical supplies,especially medical protective clothing,medical face masks and goggles,which are necessary for the treatment of critically ill patients.

To implement the guiding principles from General Secretary Xi Jinping's major speeches on the prevention and control of COVID-19 and the decisions and arrangements of the CPC Central Committee,the National Audit Office of the People's Republic of China (CNAO) organized a special audit of financial funds and donations for the prevention and control of COVID-19 in the first time.The Department of Electronic Data Audit conducted data analysis on the medical supply shortage to find out where the problem was and how to deal with it and improved the medical supplies.It has proved the role of auditing in promoting the sound operation of society and economy,and demonstrated the concept of "putting people and their lives first ".

II. Main Content

Determine the Audit Objectives

Medical protective supplies are urgently needed to treat critically ill patients and protect doctors and nurses.Therefore,it was necessary to master the status of the entire industry chain of key medical supplies production in the shortest possible time,identify the key

stage that restricted capacity,and find the reasons behind,so as to provide the National Health Commission,National Development and Reform Commission,the Ministry of Industry and Information Technology,and the General Administration of Customs and other relevant departments with a comprehensive view of the production of medical supplies,point out pain points and blockages,and propose practical and feasible measures to promote the capacity of key medical supplies.

Define Audit Methods

The purchase of equipment and raw materials,product sales,and transportation involved in business operations are represented in the taxation and customs data, therefore,it is possible to build an industry chain model of medical protective supplies based on the taxation and customs data.By focusing on optimizing the supply-demand matching of production factors,improving key raw materials production capacity,and maintaining the market order,the auditors found out and revealed problems in the operation and import &export management of medical mask and protective clothing manufacturers,and put forward corresponding suggestions.

Based on the above ideas,the auditors collected and sorted out data of taxation, customs,industry and commerce,electricity,medical device production licenses, and the internet,to conduct analysis on the supply-demand matching of production factors of medical supplies,capacity of key raw materials production,and the operation of market order.

Audit Steps,Methods,and Results

1.Supply-demand matching of production factors of key medical supplies.

Limited by logistics,labor force and industrial distribution,China's medical supplies industry is unevenly distributed.Most of the medical mask manufacturers and raw material suppliers are not in the same area.During the lockdown,cross-region logistics was restricted,which led to material shortage and the production suspension of medical mask

manufacturers,damaging efficient operation of the entire industry chain.The analysis steps were as follows:

(1) Study the production and supply in the medical mask industry chain and find out the weaknesses.For example,the main raw materials,production techniques and representative enterprises of medical masks,manufacturers of production equipment in each stage,and key components.

(2) Conduct analysis on medical mask manufacturers,important raw materials, geographical distribution of component suppliers,purchase-sale relation,and market share with the comprehensive utilization of data of medical products administration,industry and commerce administration,taxation,and data from other departments.

(3) Conduct analysis on the work and production resumption of medical mask manufacturers,the supply of raw materials and component for production equipment in low-risk areas by comprehensive utilization of data of COVID-19, electricity sector and others to find out the existing difficulties.

2. The key raw materials production capacity and market operation order.

After transaction trace analysis of the industry and correlation analysis of enterprises purchasing and selling important medical supplies production equipment,the auditors found that it was difficult to import the key components of some production equipment,and the work and production resumption of domestic suppliers was not as sound as expected,resulting in insufficient supply of complete set of equipment in the market,which limited the further expansion of capacity of related manufacturers.The analysis steps were as follows:

(1) Conduct analysis on the main suppliers of production enterprises and the purchase of complete sets of production equipment by enterprises with relative qualifications with the comprehensive utilization of data of industry and commerce administration,taxation,and

data from other departments.

(2) Conduct analysis on the purchase of complete sets of equipment and key components by relevant production equipment manufacturers at home and abroad in previous years with the comprehensive utilization of data of taxation and data from other departments.

(3) Conduct analysis on the work and production resumption of domestic suppliers of complete sets of production equipment and key components with the comprehensive utilization of data of taxation and data from other departments.

Analysis Results

The data analysis indicated that the distribution of medical mask manufacturers and raw material suppliers shows an obvious regional concentration pattern;the work and production resumption of melt blown fabric equipment manufacturers and component manufacturers suffers from a lack of coordination,which is not conducive to shorten the supply cycle.In addition,some companies were engaged in price gouging.The auditors suggested that differentiated policies should be implemented in different places to help enterprises resume work and production as soon as possible,improve the supply capacity for key medical supplies,strengthen the control of prices of melt blown fabric,and ensure sound market orderliness.

The highlight of the case is the utilization of business data from multiple industries in precisely identifying the difficulties in the entire industry chain of mask production in different places,which provided specific references to improve the capacity of medical supplies production and maintain market orderliness,fully demonstrating the accuracy,timeliness,and efficiency of big data audit.

IV. Questions

1. How to determine the content and scope of data required for analysis quickly when it comes to audit accountability and an audit objective?

2. In order to find out the difficulties in the production of medical supplies quickly, what indicators and data should be examined,or what data should be introduced to correlation analysis?

(Written by Department of Electronic Data Audit,National Audit Office of the People's Republic of China)

10 A Risk-based Approach to Inspections of Asbestos Work

Niels Dyhrberg-Nørregaard, SAI Denmark

I. Introduction

Since 1990, the use of all forms of asbestos in construction has been prohibited in Denmark; however, asbestos persists in numerous buildings, thereby constituting an ongoing occupational health hazard. According to Statistics Denmark, one in three roofs in Denmark is estimated to contain asbestos. Furthermore, asbestos may be present in other unregistered locations within buildings. During demolition or maintenance activities, workers are at risk of inhaling carcinogenic asbestos dust. The Danish Cancer Society claim that approximately 300 Danes die annually due to past exposure to asbestos dust, underscoring the critical importance of effective oversight of asbestos work.

In 2019, the Danish Parliament adopted a new Working Environment Act, partly based on recommendations from an expert committee and a broad political agreement from April 2019. One recommendation from the expert committee stipulated that The Working Environment Authorities (WEA) should develop novel and improved risk models to more accurately identify companies with the most significant working environment issues for inspection.

According to an expert committee on the working environment, the enhancement of risk models necessitates WEA's capacity to acquire data from other governmental bodies and public registers, which can then be integrated with the Authority's own information. The expert committee's recommendations were incorporated into the Act and further elaborated upon in its legislative history.

II. Main Content

Determine the Audit Objectives

The purpose of this audit is to investigate whether the WEA has a sufficient basis for assessing where the risk of illegal asbestos work is greatest. In this case description we focus on one distinctive part of the audit: whether the WEA can improve a risk-based selection of asbestos work for inspection by utilizing data from other authorities.

Define Audit Methods

The audit is based on merging data from 3 different IT systems, that is:

- Data extraction from the WEA's case management systems
- Data extraction on asbestos in buildings from the Building and Housing Register (BBR)
- Data extraction on construction waste from 13 selected municipalities.

All data covers the period 2020-2023

(1) From the WEA, we have received multiple datasets.

We have received data on inspections and registered violations of the Working Environment Act at those inspections. We have also received registration of all complaints relating to asbestos work. In addition, we have received extracts of information from asbestos notifications and construction site notifications.

According to section 12 of the Danish asbestos regulation, the employer must report asbestos work to the Danish Working Environment Authority regardless of the size of the construction site. However, notification is not required for asbestos work with a "low dust level". Asbestos notifications are therefore the Danish Working Environment Authority's primary source for where in the country asbestos work is being carried out. When the Danish Working Environment Authority supervises asbestos work, they select the construction site based on either asbestos notifications or complaints.

(2) From the Danish Valuation Agency, we have received 2 extracts of data from the BBR.

We have received data on all building permits concerning demolition, where at least one of the buildings is registered as containing asbestos. The municipalities report to the BBR when they have granted building permits. On the building permit registration, there is information about the building's materials, the period for the construction work, and what type of construction work permission has been applied for - e.g. remodeling, demolition or partial demolition. The building permit is continuously updated, with, among other things, the start date of construction and the date of completion of construction. There are 10,549 building cases concerning demolition in the period 2020-2023.

Furthermore, we have received master data on all buildings that at some point during the investigation period have contained asbestos. The data also indicates if the building no longer contains asbestos during the period. The data contains information about the materials of buildings, including information on asbestos in roofing and exterior walls. If a change in the building's materials has occurred during the period, let's say that an asbestos roof has been removed, the change and the date of the change are shown in the dataset. We have restructured, enriched and assessed the quality of the data. The extract of master data from the BBR contains 921,005 buildings in the period 2020-2023.

(3) From 13 municipalities, we have received data on construction waste containing asbestos.

Clients and companies are obliged to report construction waste over 1 ton to the municipality, including what type of waste is to be disposed of, e.g. asbestos. According to Section 7 of the regulation on the handling of waste and materials from construction and demolition work, the client is obligated to notify the municipality of the construction work at least 2 weeks before the work commences. According to Annex 1 of the regulation, the report must include information on the date, the client's name and address, the property's

address and cadastral designation, the expected start and end dates of the construction work, the year of construction, and the expected quantities and types of waste for the entire project.

We have obtained data on report of construction waste containing asbestos from the municipal case management system Bygningsaffald.dk (ConstructionWaste.dk). The 13 municipalities included in the audit have voluntarily extracted reports involving asbestos. We have received the data in a report format, which we are restructuring into rows and columns in Excel using the DataSnipper add-in

We have structured the data so that it can be merged. The municipalities have been selected to vary in terms of size, geography and type of building (multi-story buildings vs. single-story houses)

Audit Steps, Methods, and Results

1.Processing Data and validation

1) Testing the Reliability of Information on Asbestos in BBR.

We have tested the reliability of the BBR's information on roofing materials by comparing the BBR's master data with the locations where the WEA has identified illegal work with asbestos in roofs. We have only tested the reliability of the BBR's information on asbestos in roofs – and not in exterior walls – as this is where information in the BBR has the greatest value. Asbestos is present in approximately 30% of roofs and only in approximately 3% of exterior walls in the country's buildings.

Through our search for text strings, we found a total of 404 reactions that concern asbestos-containing roofing material. For the addresses where illegal work with asbestos in roofs has been found, we have matched the address in the BBR register and examined whether, at the time of the reaction, there was information about asbestos in the roof in the BBR. For 89% of the addresses where illegal asbestos work in roofs was found, asbestos

in the roof was also registered in the BBR. Therefore, we consider data on asbestos in BBR to be fairly reliable.

2) Validation of Information on Locations

The analyses are based on merging the datasets we have received from the respective authorities. We have merged data on various address information, including coordinates, road points, and full addresses. The data quality varies across the datasets, which is a challenge as the addresses are the key to merging all the datasets of the study. This is reflected, for example, in the WEA's data on inspections, where the address of the inspection is registered in free-text fields. The data quality therefore depends on the style of individual case worker's descriptions.

In particular, data on asbestos notifications is a challenge. The WEA does not systematically register information from the asbestos notifications beyond the name of the company reporting the work. In addition, the case worker often describes the type of work, the address of the work site, and/or the company's address in the title field and/or case note field. These are free-text fields with case worker descriptions of work procedures and addresses. We have found the addresses where the construction site is physically located by a semantic search of the free-text fields in the asbestos notifications' case note and case titles. For around 10% of the asbestos notification, however, we could locate the address through a construction site code assigned to the notification.

To increase data quality we have run a validation test of all the addresses in all the datasets. The validation is done by sending requests to an API at the Danish Address Register (DAR). The purpose of the validation is to ensure that the address information is correct and consistent by retrieving the information from an authorized source. The validation handles both active and historical addresses as well as requests where more than one suggestion for the correct address is returned. When we retrieve address information from DAR, we also get geo-coordinates, road points, and an address ID that belongs to the address.

We have extracted and cleaned the addresses for asbestos work in the asbestos notifications. After data cleaning, we can link approximately 93% of the 7,049 asbestos notifications to an address. In total, 6,577 unique asbestos notifications are thus included in the study, distributed across 5,870 locations. In our analysis, we have adjusted the results to make sure that the loss of addresses from 7% of the asbestos notifications does not affect the results when we merge asbestos notifications with addresses from, for example, construction waste and the BBR.

2. Method and Analysis

1) Demolition Registrations in BBR as a Source for Finding Illegal Asbestos Work

We have compared how many building permits of demolitions of buildings with asbestos in the BBR, that has been reported to the WEA in the period 2020-2023. We have done this by merging the locations of asbestos notifications with the locations of the building permits. We have merged locations based on address, road point and coordinate. In this way, we maximize the number of location matches, as different addresses may be associated with the same building while these addresses will typically have the same road point and coordinate.

When we merge the 2 datasets based on location, we find 690 matches where both an asbestos notification and a demolition case have been registered. We have then added a criterion that the asbestos notification and demolition case must be in the same period. The asbestos notification sooner than 30 days before the start date of the demolition and before the end date of the demolition. This reduces the number of matches to 626.

We have limited the population of building permits to buildings over 50 m2 to ensure, as far as possible, that the demolition is carried out by companies rather than by private individuals. This reduces the number of matches to 589.

Furthermore, we have corrected the results in a way that takes into account that we have

not been able to identify addresses on 7% of the asbestos notifications.

2) Notifications of Construction Waste as a Source for Finding Illegal Asbestos Work

We have compared how many of the notifications of construction waste containing asbestos from the 13 selected municipalities, there has been reported asbestos work to the WEA in the period 2020-2023. We have done this by merging data on construction waste with data on asbestos notifications for the municipalities. The WEA has received 1,470 asbestos notifications distributed across 1,298 addresses across the municipalities of the study.

We have merged notifications of construction waste with asbestos notifications based on addresses, road points and coordinates. Once again, we have done this to maximize the number of matches between asbestos notifications and notifications of construction waste.

When we merge the 2 datasets based on location, we find 748 matches of an asbestos notification and a notification of construction waste. Once again, we have established the criterion that the asbestos notification and the notification of construction waste must be registered in the same period. The asbestos notification must be sooner than 30 days after the end date on the notification of construction waste. This reduces the number of matches to 584

We have further limited the analysis so that only notifications of construction waste for the addresses where the construction work is performed by a company are included. We have done this by filtering all notifications of construction waste where a company appears in the "contractor" field.

After the delimitation, the population of notifications of construction waste containing asbestos has been reduced from 6,925 to 3,606. In other words, 48% of the notifications are excluded because the asbestos work may be carried out by private individuals. In consequence the number of matches between asbestos notifications and notifications of

construction waste is reduced to 491.

Furthermore, we have corrected the results in a way that takes into account that we have not been able to identify addresses on 7% of the asbestos notifications.

Analysis Results

We find that the WEA has received asbestos notifications in 7% of the cases in the BBR where buildings that are likely to contain asbestos are demolished. All in all, we have been able to identify 9,199 addresses where buildings that likely contain asbestos are being demolished. At 8,583 of those addresses – corresponding to 93% – asbestos work had not been reported to the WEA.

We also find that in only 14% of the instances where construction waste containing asbestos has been reported to the municipality, asbestos work on the location has been reported to the WEA. In 86% of the cases – corresponding to 3,103 addresses in the 13 selected municipalities – the WEA has not been aware that asbestos work has been carried out.

The results demonstrate that information on building permits and construction waste can increase the WEA's knowledge of asbestos work significantly. Therefore, we conclude that the WEA can strengthen the risk-based selection of asbestos work by using data from the BBR and the municipalities.

11 Case Study of Bank Loan Data Analysis

I. Introduction

Banks play an important role in enhancing the economic development of any country. As the economy grows, so does the demand for bank loans. The banking sector is constantly facing challenges in the loan application processing and management, such as identifying and managing the loan risk, reducing the loan processing time, and improving the loan decision's accuracy. In order to ensure that banks can better allocate resources and provide timely and effective support to the real economy, the Accountability State Authority (ASA) conducted a data analysis of the bank loan information. The aim of the analysis was to determine the key factors affecting the loan application processing and management, as well as to identify methods to improve the loan efficiency and curb their related risk.

II. Main Content

Determine the Audit Objectives

The objective of the audit was to:

1. Evaluating the bank's loan application processing and management .
2. Identifying the key factors affecting the loan application processing and management.
3. Proposing methods to improve the loan efficiency and reduce their related risks.

Define Audit Methods

To achieve the audit's objectives, auditors use different data analysis techniques such as data mining, machine learning (Artificial Intelligence model), and statistical analysis through collecting the loan application data and loan performance data from the various banks in addition to the external data sources such as credit offices and governmental agencies.

Audit Steps, Methods, and Results

1. The loan application processing analysis: the first step was to analyze the loan application processing data in order to understand the loan application process and identify any bottlenecks. The analysis steps were implemented as follows:
   - Collecting the loan application data from the banks and external data sources.
   - Data cleaning and pre-processing related to the loan application data.
   - Using the data mining techniques to identify the patterns and relationships in the loan application data.
   - Identifying the key factors affecting the loan application processing period and its approval rate.
2. The loan management analysis: the second step was to analyze the loan performance data to understand the loan management processes and identify the areas for improvement. The analysis steps were implemented as follows:
   - Collecting the loan performance data from the banks and external data sources.
   - Data cleaning and pre-processing related to the loan performance data.
   - Using the machine learning techniques in order to analyze the loan performance data.
   - Identifying the key factors affecting the loan default rate and the loan repayment rate.

Analysis Results

The analysis results showed that the loan application processing time was affected by several factors such as the loan's type, and amount, as well as the applicant's credit history. The loan approval rate was also affected by these factors in addition to the loan type and amount as well as the presented collaterals.

The results also showed that the loan default rate was affected by several factors such as the loan amount, the interest rate, the loan period, and the borrower's credit history. The loan repayment rate was also affected by these factors as well as the loan period, the interest rate, and the presence of collateral.

III. Analysis and Inspiration

The case study highlights the importance of using data analysis techniques to understand the loan application and loan management processes in the banking sector. The results of the analysis could be used for the decision makers information, the loan efficiency improvement and the loan risks limitation. Additionally, the results can be used to develop more effective loan application and loan management processes, as well as to implement data-based risk management strategies. The case study provides inspiration for other organizations to adopt data analysis techniques to better understand their own processes and identify the areas for improvement.

IV. Questions

1. What factors led to the growth in the demand for bank loans?

2. What are the steps that banks could take to mitigate the lending associated risks?

3. How can banks assess the creditworthiness of business owners and determine the loan amounts they are eligible to obtain?

4. What is the technology's role in streamlining the loan application process as well as improving the overall loan experience?

5. What are the implications of the current economic environment on the businesses seeking bank loans and on the banks?

6. How can banks help business owners to better understand the terms and conditions of their loans, including the repayment schedules, interest rates, and fees?

12 Case Study of Youth Vocational Rehabilitation

I. Introduction

This case study comes from a recently published performance audit report Youth not in employment, education or training - identification, outreach, and rehabilitation (Audit report no. 1/2025). Among other audit objectives, our goal was to assess the effectiveness of youth vocational rehabilitation in advancing youth employment and education.

II. Main Content

One of the main objectives of the audit was to estimate the causal effect of a relatively new vocational rehabilitation treatment, Nuotti coaching. In operation since 2019, Nuotti coaching is a "low-threshold treatment", that is, no medical diagnosis is needed to apply to Nuotti coaching. In 2022, the government spent approximately 23 million euros on Nuotti coaching services, excluding rehabilitation allowance (at the minimum, 32 euros per working day).[1]

The treatment is targeted at 16–29-year-olds, whose functional ability has weakened significantly. In Nuotti coaching, each participant has his/her own personal coach, with whom they cooperate to achieve their goals. The objectives of coaching are decided together with the participant, and they might include – for example – gaining employment, or attending school, or participating in labour market training. Coaching includes 20 1-hour meetings in a period of five months (or less, as single meetings may be combined into 2-hour sessions). Coaches can also assist participants in their daily activities e.g. with the social or

[1] The level of rehabilitation allowance is above last-resort social assistance but below basic unemployment allowance.

medical services or with the employment office. Coaches themselves are usually seasoned social workers or health care professionals.

For the case study, we combined individual register data from Kela (the Social Insurance Institution of Finland) and Statistics Finland. Treated young persons participated in Nuotti coaching in 2019. Controls were a random sample of long-term social assistance recipients from 2018. Social assistance is the last-resort government subsidy for low-income households. Either group had no previous rehabilitation experience. From Statistics Finland we obtained a 10-year panel data set (2013–2022) on the sampled individuals, containing various variables, such as age, gender, main activity, no. of working days, wage income, etc. The anonymised microdata was analysed using Statistics Finland's FIONA remote access system.

Our baseline sample was restricted to those young persons who had finished their comprehensive schooling by 2013, that is, they were at least 16 years old in 2013 (i.e. they were born in 1990–1997). For the baseline sample we selected those Nuotti participants who also received social assistance in 2018 (all Nuotti participants were not social assistance recipients in 2018).

Our baseline sample therefore had 636 Nuotti participants (with a mean age of 23 in 2018) and 1 156 controls (mean age 24). Sample size was 17 920. The unrestricted total sample size was 30 180. We had a three-year follow-up period, 2020–2022. The estimations were carried out separately for men and women.

III. Analysis and Inspiration

Participation in Nuotti coaching is not randomised; treatment allocation is based on written applications. First, the youth decide whether to apply, and then Kela decides who gets treated. Therefore, we have obvious selection bias. Nevertheless, to obtain a credible counterfactual and to establish causal inference, we need parallel pretreatment trends (see Fig. 1 below).

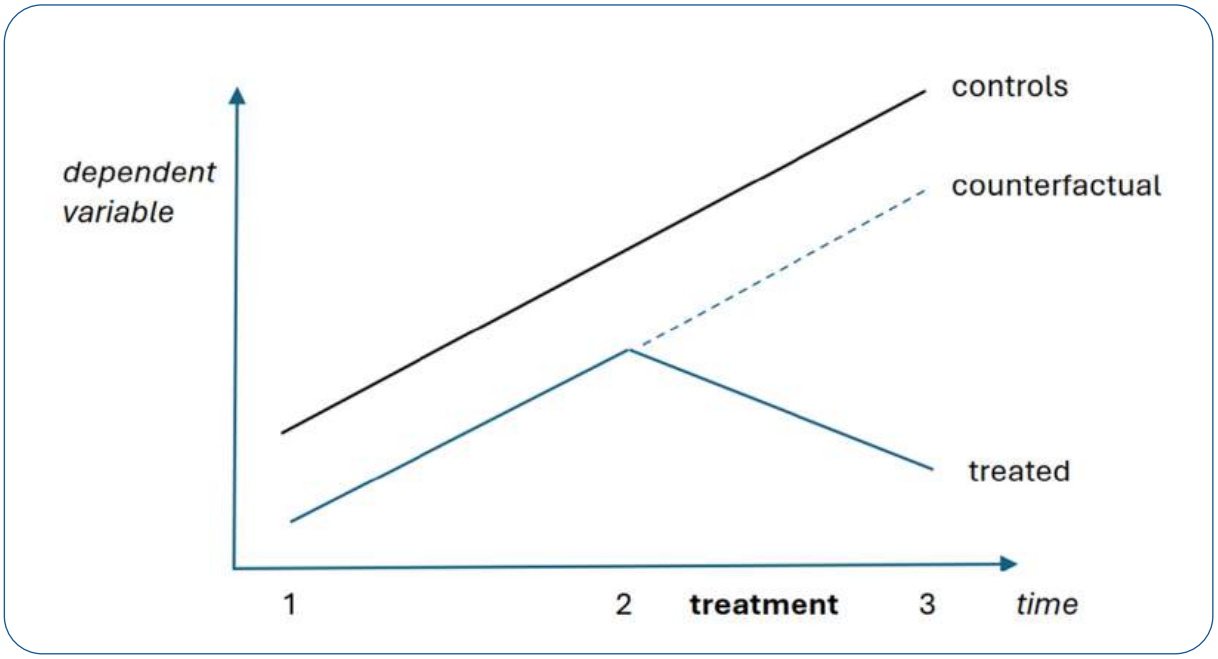


Figure 1. Identifying assumption in difference-in-differences models.

Parallel trends are unlikely in this case because of self-selection. In practice, parallel pretreatment trends are quite rare, and often arise only in natural experiments, where unexpected exogeneous circumstances determine treatment allocation. We tested the existence of parallel trends using the estat ptrends command in Stata and most of our dependent variables exhibited nonparallel pretreatment trends.

To overcome this fact, we used the synthetic difference-in-differences (SDID) regression method suggested by Arkhangelsky et al. (2021).[2] In the SDID regression, we use unit and time weights to make pretreatment trends approximately parallel. SDID has been available as a user-written command for Stata since 2023.[3]

[2] Arkhangelsky D., Athey S., Hirshberg D.A., Imbens G. W. & and Wager S. (2021): Synthetic Difference-in-Differences. American Economic Review, Vol. 111, No. 12 (Dec. 2021), 4088–4118.
[3] The command can be installed in Stata by typing "ssc install sdid". For details, see: Clarke D., Pailañir D., Athey S. & Imbens G. (2023): Synthetic Difference-in-Differences Estimation. Institute of Labor Economics, Discussion Papers No. 15907. January 2023.

We used employment status (0/1), student status (0/1), number of working days per year, and wage income per year as dependent variables. The SDID estimator accounts for unobservable individual fixed effects (such as innate ability) and time fixed effects (like economic fluctuations). Treated and control groups are quite similar in observable characteristics, but possible sources of bias are unobservable time-varying characteristics, such as person's motivation, substance abuse, etc. Because of their inherent nature, these are impossible to control for.

With the SDID estimator, we found no effects on employment status, working days per year, or wage income, neither for men nor women. The only positive effect of Nuotti coaching we found was an increase in men's student status (Fig. 2). The share of male students increased by 3.9 percentage points (p = 0.034) after the treatment when compared to the controls. We tested the robustness of these results using various sensitivity tests (by relaxing the baseline sample restrictions), and the results were similar. Women's student status was unaffected.

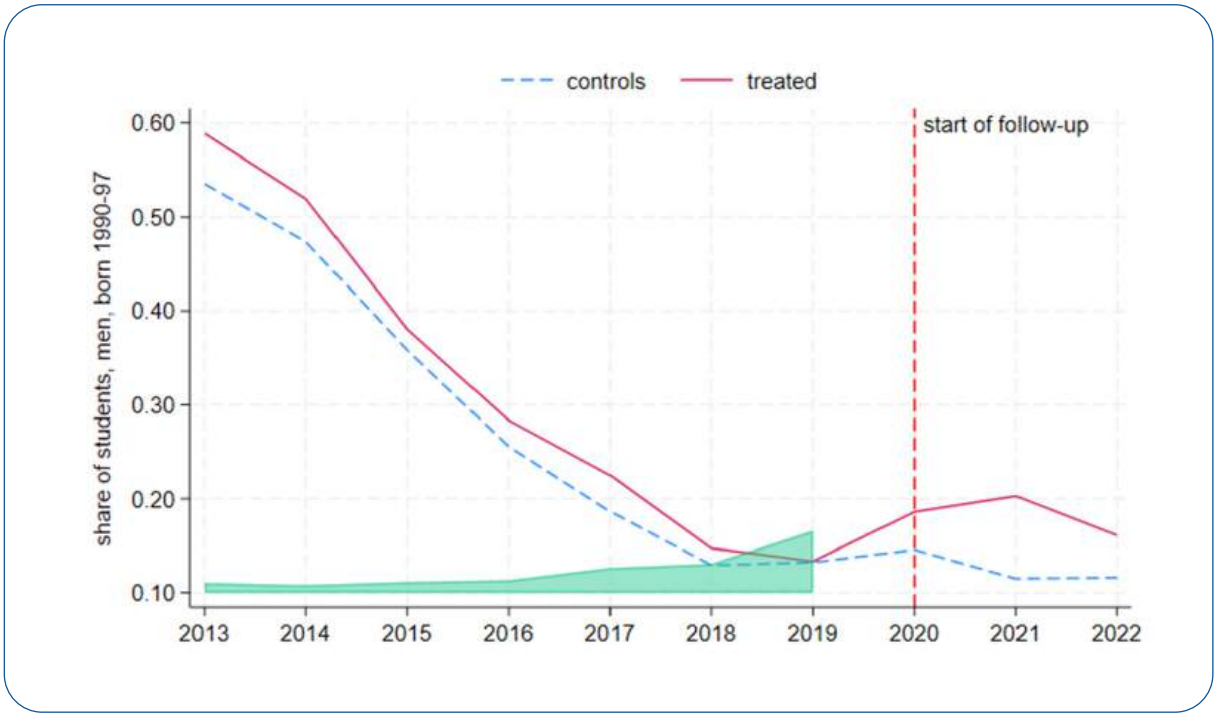


Figure 2. The effect of Nuotti coaching on men's student status.

IV. Questions

Our regression results indicate that youth vocational rehabilitation is not totally without merits, although positive effects were quite scarce. In hindsight, a longer follow-up period would have been useful, as standard secondary education usually lasts for three years. Men's increased educational attainment could lead to completed vocational degrees and subsequently into employment and higher wages, but it could take up to 5–6 years, or even longer, for these effects to emerge in register data.

Why did Nuotti coaching improve only men's student status but not women's student status? We do not have definitive answers. From the background data we know that treated women were better educated than treated men even before the treatment started. In 2018, 37.8% of the treated men had at least secondary education, while 41.7% of the treated women had at least secondary education – a difference of approximately four percentage points. Our results suggest that coaching has helped men to catch up the pre-existing educational difference.

(Written by Ville Vehkasalo, Audit Unit, National Audit Office of Finland)

13-17 BPK's Experiences in Big Data Audit

I. Introduction

Rapid technological advancements have made Big Data crucial for decision-making in various industries. Amid this transformative shift, the Audit Board of Indonesia (BPK) has embraced Big Data not just as a tool but as a cornerstone of its auditing practices. This move signifies more than adaptation—it represents a groundbreaking effort to ensure audits achieve greater precision, efficiency, and relevance in tackling increasingly complex challenges.

Big Data serves as the foundation upon which the building of Artificial Intelligence (AI) can be constructed and advanced. The vast reservoirs of data collected and analyzed provide the essential raw material that enables AI systems to learn, adapt, and evolve. By harnessing Big Data, BPK can derive profound insights and develop sophisticated AI models that drive actionable intelligence. This synergy between Big Data and AI is pivotal in enhancing audit methodologies, enabling auditors to predict patterns, identify anomalies, and make informed decisions with unprecedented accuracy and speed. Consequently, Big Data is not merely an ancillary tool, but an invaluable component propelling the integration and application of AI in auditing and beyond.

Since 2020, SAI Indonesia (BPK) has established the BIDICS Platform to handle vast amounts of interconnected data. By 2021, BPK started using Robotic Process Automation (RPA) and Machine Learning (ML) to improve efficiency. RPA collects unstructured data, converts it into structured data, and allows ML to provide auditors with better insights. Through the lens of five distinct use cases, BPK will share the experiences in applying

Big Data Audit not only using Big Data in audit but also auditing the Big Data Analytics implementation in auditee.

13 Case Study 1: Scraping Government Procurement Portals Using RPA

Overview

Robotic Process Automation (RPA) involves using software robots to automate repetitive tasks. BPK embed the RPA into the BIDICS Platform. Every year auditor in BPK requires a list of government procurement in each audit entities they've audited. Years ago, the auditors have to wait several days just to receive such data from the auditee. Since the procurement announcements are available for public, BPK initiated to scrape the data everyday using a never sleep machine. BPK uses RPA to scrape data from government procurement portals, providing auditors with up-to-date information for analysis. The goal is to reduce the waiting time for receiving the procurement data.

Methodology

RPA navigates procurement portals, grabs the unstructured data, extracts the data, and stores it in a structured format, thus reducing time and effort for data collection. This allows auditors to focus on analysis. RPA not only does convert unstructured data into structured data but also classify it into four types of procurement anomalies: the winner is not awarded a contract, the winner did not offer the cheapest price, offered price is higher than the owner's estimate, and the winner is blacklisted.

Results

BIDICS, through the RPA, provides auditors with initial signs of irregularities in government procurement. This information is used during confirmation processes as part of audit correspondence, following audit standards.

14 Case Study 2: Detecting Bid-Rigging in Numerous Government Procurement

Overview

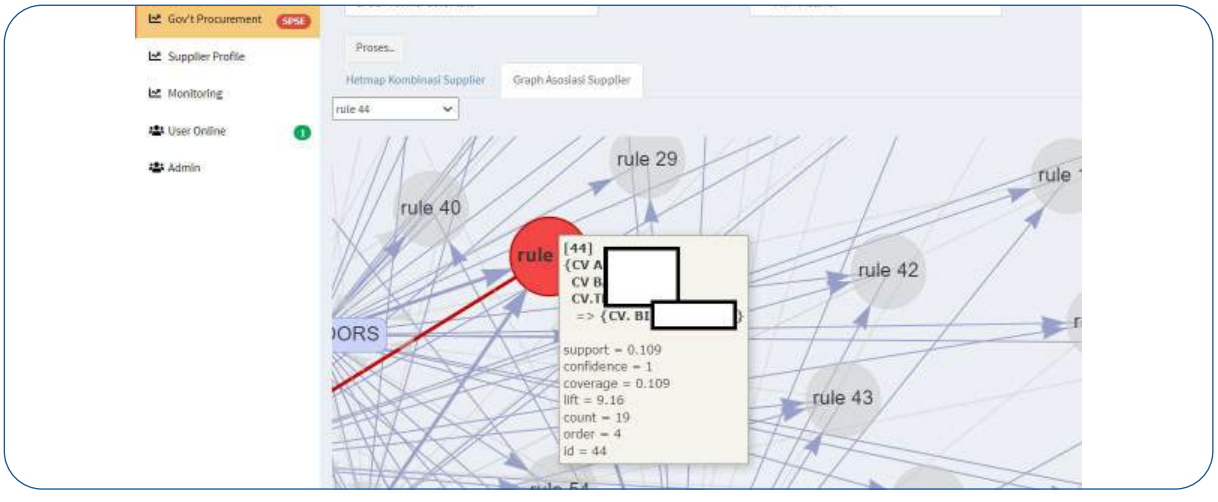
Bid-rigging is a form of fraud where competing parties collude to manipulate the outcome of a bidding process. Detecting bid-rigging in government procurement is challenging due to the complexity and volume of data involved. Big Data Analytics technique, particularly association rule mining, has proven effective in identifying suspicious patterns indicative of bid-rigging.

Methodology

BPK employs RPA-collected data to detect possible bid rigging in government procurement. Using the CRISP-DM methodology with the Apriori Algorithm in BIDICS Platform, BPK analyses large datasets to find togetherness of bidders in several procurements, revealing patterns of bidder collusion possibility. For instance, if a group of bidders win contracts alternately in certain procurements, it may suggest bid rigging.

Results

The BIDICS identified several suspicious patterns, leading to further investigations that confirmed bid-rigging activities. This case illustrates the power of AI in uncovering complex fraud schemes that traditional methods might miss.



15 Case Study 3: Detecting Contract Splitting to Avoid Competitive Bidding Using Text Similarity

Overview

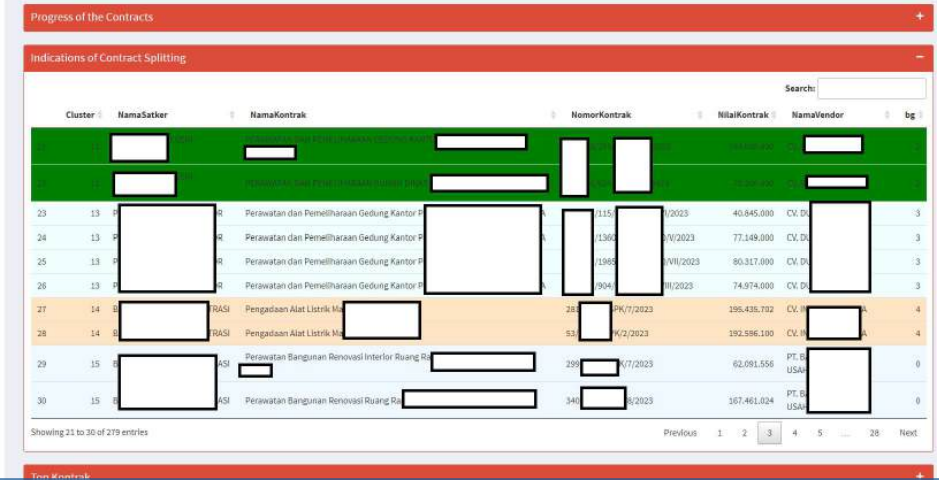
Contract splitting or dividing a large contract into smaller ones to avoid competitive bidding thresholds is another common fraud in government procurement. Big Data Analytics techniques, such as text similarity analysis, can detect such practices by comparing contract descriptions and identifying similarities.

Methodology

Text similarity analysis involves natural language processing (NLP) algorithms to compare textual data. In this case, contract descriptions are analyzed to identify overlapping content, which may suggest deliberate splitting to circumvent competitive bidding processes.

Results

BIDICS is able to cluster several instances where contract descriptions were nearly identical, despite being awarded separately. Subsequent investigations revealed intentional contract splitting, leading to audit findings and corrective actions to prevent future occurrences. The following picture illustrates that several contracts with similar names, when combined, exceed 200 million Rupiahs, which is the threshold for direct procurement from the same suppliers.



16 Case Study 4: Using K-Means Clustering as a Sampling Method to Decide a Site Visit Location

Overview

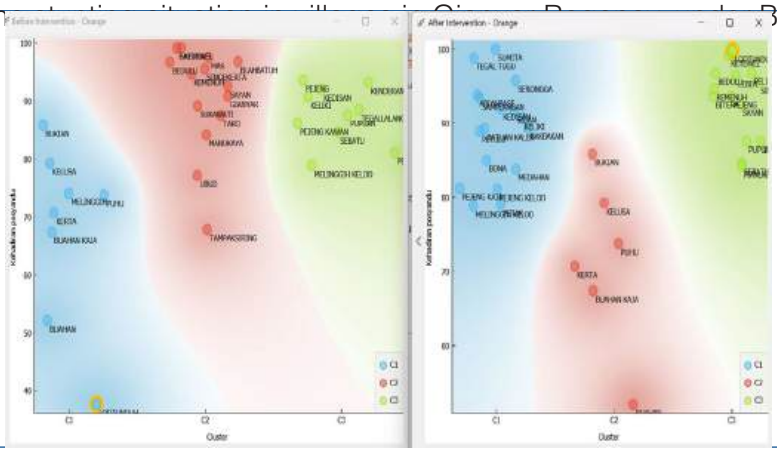
Indonesia consists of 34 Provinces and 508 Local Governments. In 2024, BPK conducting a performance audit on government responses on stunting. Due to the limitation of workforce and budget, it is not possible to visit all local governments. In order to address this circumstance, the audit team proposed to use Big Data Analytics techniques to define which villages should be visited.

Methodology

Supported by the Big Data Analytics Lab, the Audit team used K-Means clustering to select villages for field audits. The algorithm categorized villages as high, medium, or low risk based on sanitation, toddler height, public health services, and nutrition. Each regency applied the algorithm to ensure affordability and clarity of results.

Result

By visualizing the clusters generated by the K-Means algorithm, the audit team can make more informed decisions about visiting villages. The following graph illustrates the data visualization of th[illegible] Bali Province.



BPK used this audit as a case study during the International Training on Big Data in July 2024 at the

Bali International Training Center. The training lasted one week and included 12 SAIs.

17 Case Study 5: Auditing the Implementation of Machine

Overview

In the past five years, Indonesia has experienced an increase in cases of online gambling. The government is focusing on blocking websites that contain information about online gambling. Ministry of Informatics and Communication (MIC) plays a main role to regulate the Internet traffic including the content. MIC has set up infrastructure to block illegal content including pornographic, hate speech, and online gambling. One of infrastructure has been equipped with machine learning model to classify the web content.

In 2024, BPK conducted performance audit on how government addressing the issue of illegal content. The audit team has to deal with the implementation of machine learning. The team assessed whether machine learning has effectively satisfied the government need in an efficient manner or not.

Methodology

MIC uses a combination of three algorithm to classify the web content into pre-defined label such as porn, gambling, hate speech, and hoax. MIC uses Big Data Analytics technique that applied several algorithms to predict the category of the web content. Since these three algorithms are gray-boxes, where the auditor knows the formula but cannot observe its functioning, the auditors have used the "Audit Around Machine Learning" approach.

Result

The audit team has proposed recommendations to improve the performance of Machine Learning including the preservation of data set history, history of accuracy of each retraining process, and monitoring the Machine Learning process.

1. Data Quality and Preparation

High-quality data is crucial for effective Big Data Analytics in auditing. Ensuring data accuracy, completeness, and relevance is the first step towards a successful Big Data Audit. Collecting and preprocessing data from reliable sources enhances the Big Data Audit performance.

2. Continuous Monitoring and Updating

Big Data as a massive source of audit evidence must be continuously monitored and updated to adapt to evolving patterns and trends. Regular audits and evaluations of Big Data ensure they remain effective and accurate over time.

3. Collaboration between Auditors and Data Scientists

Effective collaboration between auditors and data team is essential for leveraging Big Data Audit. Auditors provide domain expertise, while data team bring technical skills to develop insight from Big Data.

The integration of Big Data Analytics into the audit process represents a significant advancement in the field, offering enhanced insights and more robust audit outcomes. By focusing on data quality and preparation, continuous monitoring and updating, and fostering collaboration between auditors and data scientists, organizations can harness the full potential of Big Data to improve the accuracy, efficiency, and effectiveness of their audits. These practices not only ensure a more thorough evaluation but also position the organization to better respond to future challenges and opportunities.

18 Case study on Data Analytics- WORKING GROUP ON BIG DATA

Analytics on IFMIS data by SAI Pakistan for Auditing Budget Utilization

I. Introduction

The Financial Accounting & Budgeting System (FABS) of the Government of Pakistan is an IFMIS in which budgeting, payment processing and accounting functions of federal and provincial governments are performed. The ERP-based FABS is managed under the office of the Controller General of Accounts. Payroll and pension of more than 4 million employees and retirees, and thousands of daily payments claims of suppliers to government departments across the country are processed through the FABS. The system suffers risk of fraudulent payments, payment of ineligible allowances, less deduction of due taxes from supplier claims, and such other wrongful payments. The system also has record of timeliness or otherwise of budget releases, re-allocations and re-appropriations of budget. The SAI Pakistan has established a data-warehouse platform to pull-in data of audited entities for analysis, inclusive of data from IFMIS. The SAI has also acquired data-analytics tools, inclusive of NLP capabilities. To audit discrepancies in payments, fraudulent activities or anomalies in budgeting processes, the SAI has established an API linkage with IFMIS. A pool of auditors has been trained in data analytics. After the requisite linkages, infrastructure and capacity building the SAI has initiated data analysis to assess quality of budget utilization through its IT Wing.

II. Main Content

Objectives, Process and Methods

The initial objective of audit was to analyze payroll, pension, and supplier payment data for any discrepancies. Subsequent analysis would be extended to assess quality of budget

releases and its impact on payment processing and procurements. The focus of ongoing analysis is to identify key trends in the area of employee related expenses, red flags in pay and pension roll and suggest possible remediation measures.

Through API the data from FABS was pulled into the audit data-warehouse. Various scenarios of potentially fraudulent activities and discrepancies in payments in terms of violation of payroll, pension and supplier payment rules, gathered from past inspection reports form the basis for data analysis. The analytical tools of QlikSense software are being used for analysis. The use of NLP has not been to the desired level due to gaps in learning curve of auditors who have recently undergone capacity-building. The pace is expected to pick up as the learning curve grows.

Based on anomalies identified through the analytics, the audit teams will go into the respective audited entities for physical scrutiny of manual record and follow transaction trails in record of inventories, stocks and stores, bank data as per requirement.

Analysis, Results and Prospects

Various discrepancies and red flags have been identified which will be further probed during field audit visits. Such discrepancies include issues like pay from multiple personnel numbers being credited into same bank account, pay and pensions of certain individuals being excessively higher than the normal range, personnel of certain departments getting allowances that are not allowed for the particular post or the particular departmental staff. The audit is still ongoing with data analytics in process, based on findings of which the field audit activity will start. Some key discrepancies have been identified, establishing the utility of linkages with IFMIS and other databases of audited entities, and of identification of risk areas for a more efficient and focused field audit. Various discrepancies in quality of data in the IFMIS have also been identified. For example, date of recruitment of employees when correlated with date of birth data yields absurd scenarios which cannot be practically possible. Feedback on these discrepancies will help improve quality of IFMIS, and in turn of audit. Going beyond identification of such discrepancies SAI reports will advocate

continued expansion of IFMIS. Of late IFMIS has been expanded to capture domestic debt data through interface with Central Directorate of National Savings. Such expansion will enable SAI to maximize financial and compliance audits through promoting internal controls of IFMIS to identify and redress discrepancies, and shift strategic focus of SAI to audit of performance outcomes of government's programs in social development sector. The very initial analysis also reinforces the need for data-exchange protocols among audit office and the audited entities for improved quality of data. The present exercise has primarily relied on correlations and insights into the large quantum of data through data-analytics tools.

Question

1. How to extend analysis beyond PFM and Budget Utilization analysis to cover non-financial data analytics?

2. How to overcome barriers to data-exchange, and build capacities for NLP-based analytics?chain of key medical supplies production in the shortest possible time,identify the key

19 Developing a Data-Driven Framework for Risk Assessment of Portuguese Public Procurement Contracts

I. Introduction

Public procurement is a key pillar of public service delivery, representing on average approximately 12% of GDP across OECD countries. Threats and risks affecting the public procurement function can have significant consequences on the quality and quantity of public services that governments can provide. Supreme Audit Institutions (SAIs) are key actors in the public procurement system. By publishing audit or other control reports and sharing findings with the public and its stakeholders, SAIs promote transparency. This fosters public confidence in government procurement processes, safeguarding it from threats and ensuring its efficiency, effectiveness and integrity.

In recent years, SAIs and other oversight bodies have increasingly explored the use of data to enhance their oversight activities. Modern SAIs use data analytics to monitor procurement activity in real time, detect possible irregularities and recommend preventive or corrective measures.

Recognizing the need to adapt and innovate, Portugal's Court of Auditors (Tribunal de Contas) has engaged on a broad digital transformation, illustrated most recently in its 2023-2025 Strategic Plan for Digital Transformation (Digital Strategy). The plan underscores the Court's ambitions and needs to strengthen risk assessment in public procurement and embrace the opportunities and challenges posed by big data, digitalization and artificial intelligence.

This Case Study explores opportunities and challenges concerning an initiative of the

Court to strengthen its use of data and advanced analytics for assessing risks in public procurement, namely in the development of a data-driven model that can be used to identify and quantify a relevant number of risks associated with public procurement carried out by entities subject to the Court's control. This framework was developed in the frame of a EU funded project (SG REFORM) which was developed by a project team from Organization for Economic Co-operation and Development (OECD), NOVA IMS University and the Court itself.

II. Main Content

1. Defining objectives

The Court's overall goal for this project was to develop (and, once finished, adopt into current use) a framework for data-driven risk assessments as a means of moving towards a more automated, AI-driven form of risk assessment, within its control and oversight activities.

As such, the primary objective was to establish a public procurement risk matrix in order to identify, prioritize and measure those risks that could be measured exclusively through data obtained by the Court: either data received from several public agencies that would be willing to collaborate with the Court, or the data that the Court already receives and manages.

Numerous methodologies and studies have been developed to define and explain various risks that can be assessed from public procurement data. This academic literature was initially considered by the project team. (for example, several papers by authors such as Mihály Fazekas, Alberto Vannucci, Manuel J. García Rodríguez, among many others).

However, it is important to note that the Risk Matrix proposed by the Court faced had to address several specific and quite restrictive criteria, namely:

- The risks had to be accepted and recognized by the Court's auditors as realistic, relevant and, as much as possible, easy to understand and assess.

- On the other hand, the risks should be aligned as much as possible with the Court's current scope of activities and strategic control objectives, within the scope of public procurement.
- The risks should cover, to the extent possible, all phases of the public procurement cycle.
- Finally, the quantification of these risks had to be obtainable from the data available to the Court, in an automatic manner.

Thus, the objective was to derive this customized, internal Risk Matrix, and then proceed by developing a data-driven model that would, by using all the data that could be made available to the Court, automatically identify and quantify all the relevant risks of public procurement contracts.

2. Defining methods

The methodology followed by the Court was, in essence, quite typical of any data-science projects: understand the business aspects of the problem → acquire and verify data necessary to solve the problem → model the solution → deploy and evaluate the solution[1] ().

The development and refinement of the Risk Matrix corresponded, in gross terms, to the "business understanding" phase of the project's methodology. This phase required quite an extensive time frame, as the project team started with standard academic literature review on public procurement risks but – as already pointed out –also interviewed and discussed with the Court's auditors and professionals their own views on how to identify and assess public procurement risks from a data driven perspective. These interviews, questionnaires and discussions can take quite a long time to realize, due to the normal agenda and calendar restrictions that affect all the interested parties. Additionally, questionnaires were also carried out with the most relevant stakeholders.

Another aspect that had to be carefully planned and defined, before the execution phase,

[1] Which is heavily inspired from both Microsoft's Data Science Lifecycle and the well-known CRISP-DM process model.

is the establishment of data exchange protocols with external public agencies. Although the Court already receives and stores a plethora of data, including quite an extensive amount that exclusively pertains to public procurement, it was very clear that some major sources of data were missing , and could only be obtained through partnership protocols established with external agencies – this was the case, for example, with the national agency that regulates public procurement in Portugal (IMPIC) and made available their BASE database records (basically, almost all of the public procurement contracts established within Portugal since 2018).to this project purpose.

The final aspect that needed to be defined was the data-driven modeling to be conducted on all the data collected by the project. After evaluating the categories of risks, the nature of the data, and the desired outcomes, the project team concluded that a mixed approach would be the most effective choice. They developed three distinct types of risk indicators.

- Risks based on an “expert rule-system” (risks identified by unambiguous decision trees, based on experience and judgment of experts on public procurement).
- Risks based on statistical inference (risks identified in accordance with the probability of specific values or results).
- Risks based on machine-learning models (risks identified through the use of supervised learning executed over previous public procurement data).

3. Milestones and project actions

The initial steps were carried out almost simultaneously, in order to save time and be more efficient: the Court of Auditors started by establishing data exchange protocols with several public agencies, at the same time that the internal brainstorming and subsequent refining of the Risk Indicators Matrix was developed.

The Risk Indicators Matrix started being developed from the start of the project, in the beginning of 2022, but only stabilized into the final set towards the second part of 2023. At this time, the Risk Matrix includes 37 risks classified by 9 indicator groups: Financing (1), Procurement procedure (9), Contracting requirements (1) Evaluation of bidders, tenders and

award procedure (8), Contract award and execution (10), Payment and financial obligations (1), Conflict of interest, fraud and corruption (3), Intervening parties (issues with suppliers or contracting authorities (3) and non-compliance with of competition laws (1).

Concurrently with the development and evolution of the risks to be identified from the data, the Court took active steps to obtain and consolidate this public procurement data. At present time, the project is receiving and using the following data:

- [external agency] IMPIC's public procurement database records: 18 datasets / 24 million rows
- [external agency] Information portal on various topics of management of public resources of the Portuguese State, namely the European Union funds database: 2 datasets / 790 thousand rows
- [external agency] AdC's competition agency's database: 1 dataset / approx. 400 rows
- [internal system] Court's public procurement database: 2 datasets / 2.8 million rows
- [internal system] Court's complaints database: 1 dataset / 1200 rows
- [internal system] Court's Prior Control database: 1 dataset / 33 thousand rows

Finally, the three types of models described above – rules-based, inference-based and ML model-based – were programmed and are executed on an integrated and custom-built Python application. This Python application has been developed based on well-known and robust, open-source libraries such as Numpy, Pandas, Polars (among others) and the Torch and PyTorch Lightning libraries for the Convolutional Neural Network used for development of a ML-based collusion risk indicator.

At the moment, and although the Court has put together a SQL Server-based "simplified" Data Lake for storing and managing the diverse source data collected and used within the project, the Python scripts are still being fed flat-file data (in csv format) in order to run the calculations and output Arrow results files. These results, indicating whether any specific contract meets or fails a specific Risk Indicator (resulting in a binary value of one or zero) are then used for a star-schema data model that makes possible the data visualization

results (dashboards) made available in the Power BI service.

The Court has already started to further integrate this technological stack, as it envisages a smooth and seamless transfer between receiving the original source data and outputting final results into dashboards available to internal users. This, however, requires some additional work, which the Court is currently undertaking internally, in order to place this tool into "production-ready" state and available to all users within the Court.

III. Analysis and Inspiration

The development of this new tool is still ongoing, so it is not currently possible to share the results from having applied it to the day-to-day work of the Court of Auditors. Nonetheless, by sharing some of the developmental work already achieved (in conferences, report published by OECD, and other events), we have noticed a growing interest in the tool and on how data-driven risk assessment for public procurement contracts can aid auditors of the Portuguese Court of Auditors, as well as from other SAIs.

IV. Questions

The questions raised by the work carried out in this project, being of particular concern to Portugal's Court of Auditors, focused on the following themes and aspects:

a. Data exchange mechanisms. How can data be interchanged with greater efficiency, taking less time and effort whilst also improving the effectiveness of cross-agency data use? How to make public procurement data more interoperable, so as to maximize the effectiveness of data-driven models like ours? How can we promote secondary use of data that does not imply undue use of resources by the agencies responsible for this data?

b. Data Governance / Data Quality. How can we improve the general level of Data Governance – in particular, Data Quality – in the different stakeholders involved in data-driven projects? Is it sufficient to serve as an example, a "champion" of Data Governance, by improving our levels of governance and quality, in respect to our own data? Are there collaborative activities and endeavors that will help promote and increment Data Governance across the entire spectrum of public administration?

c. Validity of Results from Data-Driven Methodologies. How can the results that projects like this one provide, be checked against the real-life, day-to-day operations of SAIs dealing with public procurement control and auditing? How can validation procedures be developed that are easy to execute, trusted by the SAI's experienced auditors and augment the model's internal algorithms?

(Written by Centre for Innovation, Technology and Methodologies, Portugal's Court of Auditors)

20 Ensuring Local Jobs for Regional University Graduates - An Analysis

I. Introduction

Board of Audit and Inspection (BAI)'s Audit on Demographic Change I (Local regions) predicted that non-capital regions of Korea are headed towards becoming 'super-aged societies' due to ongoing super-low fertility levels, resulting fall in school-going population and migration of youth to the capital region. This situation is expected to only get worse as more and more regional universities outside the capital region begin to have a shortage of new students, which eventually will lead to their shutdown.

The Ministry of Education, in recognition of this issue, has made structural changes like the reduction of university enrollments, as well as policies to enhance the competitiveness of universities, such as restructuring and specializing departments that are in high demand in local communities while reducing admission quotas. As a representative initiative, the Ministry has invested a total of KRW 2,889.2 billion over five years (2014-2018) to promote university specialization projects.

The purpose of these projects is to enhance the competitiveness of universities by specializing departments that are related to local industries and demands, and to promote co-prosperity with local communities.

The Future Strategy Audit Division of BAI conducted an audit to identify and solve issues in the implementation of these specialization policies through data analysis of various institutions' databases.

Through this audit, the role of audits in promoting proper policy implementation was demonstrated, such as making sure the government's projects achieved their intended purposes within the budget, the execution process was efficient, and the intended outcomes were achieved.

II. Main Content

Determining Audit Objectives

The Ministry of Education received a massive budget for investment in the specialization project. However, there has been no comprehensive analysis of whether the project was employed for those departments that are in high demand in local communities and if yes, whether they have resulted in enhanced local job opportunities for the students of these universities.

Therefore, there was a need to examine whether the Ministry's specialization project and restructuring of admission requirements, planned in accordance with local community demands, achieved its intended results.

To conduct this audit, a comprehensive analysis of the issues regarding the specialization project was necessary by utilizing big data from various institutions and establishing improvement measures.

Defining Audit Methods

Using data such as the National Business Survey from the Statistics Korea and the Graduate Employment Statistics from the Korea Educational Development Institute, the main industries in each of the eight regions were identified first, after which the departments with high employment rates by major industries were identified by region. This defined the departments which are in high demand in the region (region-specific departments).

Based on this, a comparison and analysis of whether these departments were selected for

the specialization project and whether region-specific departments were considered in the easing of admission requirements was done.

Furthermore, focused interviews were conducted with six selected universities, categorized by region and type, to identify their concerns and suggestions on policies easing admission requirements, university specialization projects, innovation support projects, and job opportunity programs.

Audit Steps, Methods, and Results

The following is an analysis that used the BARON (Best Audit & Inspection System for Rule-based Observation Network), the BAI's audit data analysis system.

1. Creation of local community needs and region-specific departments

To comprehensively verify the effectiveness of the Ministry of Education's specialization projects and restructuring of admission requirements, it is necessary to first identify the major industries in each region and analyze the departments with high demand in the local community. The analysis steps are as follows:

(1) Identification of major industries by region

Based on the 77 industries of the "Korean Standard Industrial Classification," the major industries were classified by considering the average number of employees in each region for five years.

(2) Analysis of departments with high local demand

The employment rate of graduates in the 77 industries nationwide was analyzed, and departments that are highly linked to the major industries in the region and have high demand from the local community were identified.

2. Analysis of region-specific departments and specialized departments

To evaluate whether region-specific departments with high demand in the local community were prioritized for selection when implementing the specialization project, an analysis of the

linkage between region-specific departments and specialized departments was conducted for the eight regions.

The analysis revealed that out of the total of 2,025 specialized departments selected and supported, only 839 (41%) were region-specific departments.

3. Analysis of easing admission requirements for region-specific departments and specialized departments

The specialization project links universities with local communities through specialization in comparatively advantageous fields. Therefore, it is desirable to minimize the easing of admission requirements for the supported departments. The analysis examined whether easing of admission requirements and structural adjustment policies were efficiently implemented and managed by analyzing the status of easing admission requirements of region-specific departments and specialized departments.

(1) Analysis of easing admission requirements for region-specific departments

Using the "University Admission Quota Statistics," the easing of admission requirements for region-specific departments in the eight regions, including the Seoul metropolitan area, was compared and analyzed against the regional average, revealing that the admission requirements of 27 out of 87 region-specific departments nationwide were eased more than the regional average.

(2) Analysis of easing of admission requirements for specialized departments

Using the "University Admission Quota Statistics," the easing of admission requirements for specialized departments in the eight regions, including the Seoul metropolitan area, was compared and analyzed against the regional average, revealing that the total admission requirements of specialized departments were eased more than the regional average in three regions.

Analysis Results

The purpose of the specialization project is to strengthen the connection between local communities and universities by focusing on supporting departments with high local demand. However, the Ministry of Education allocated the least amount of budget (11%)

to the 'regional strategy' category, which promotes specialization linked to the project's purpose and local strategic industries, in comparison to other two categories: 'university autonomy' and 'national support.'

Additionally, by supporting departments that schools applied for without properly researching and considering local demand, the initiative has not been able to provide sufficient support for departments with high local demand.

III. Analysis and Inspiration

The highlight of this case is that it accurately identified the problems of implementing the policy of specialization by using various big data resources from multiple organizations, such as national business data, employment data of higher education institution graduates, data on the status of specialization projects, and university admissions requirement data, in order to analyze the issue from multiple perspectives. The audit also provided alternatives for improvement.

IV. Questions

1. How to determine the content and scope of data required for analysis quickly when it comes to audit accountability and an audit objective?

A. The topic of analysis must be established first and studied from various perspectives for a thorough understanding. Then, items and checklists that need to be reviewed for actual use during the audit must be identified step-by-step. After that, a list of data suitable for the analysis's purpose should be created and the availability of each data set verified.

(Written by Division of Digital Audit Policy and Management, Board of Audit and Inspection)

21 Application of the Propensity Score Matching to Assess the Impact of Field Tax Audits on Auditees` Performance

I. Introduction

Since 2016, control and supervision activities have been reformed in the Russian Federation in order to reduce the administrative burden on business. Authorities have adopted a riskoriented approach to their control and supervisory activities, the choice of objects, topics and issues for inspection have been based on risk indicators and a pre-audit analysis.

In 2022, the Accounts Chamber of the Russian Federation analyzed the work of the Federal Tax Service of the Russian Federation (FTS of Russia), particularly their organization and implementation of field tax inspections and its impact on audited companies. To implement this assessment, the Propensity Score Matching method (PSM) was used.

The study used data on field tax audits from 2017 to 2021 and the financial performance indicators of the companies for each year of this period.

II. The main part

Purpose of the audit

The purpose of the analysis was to assess the impact of field tax audits on the company's performance (revenues, expenses, revenues) depending on the size of the company, industry and duration of this influence.

Audit Methods

There are several steps of assessment

1) We formed two datasets: one with performance indicators of companies that were audited in one of the years during the period 2017-2021 and the second with performance indicators of companies that were not audited for 2017-2021.

2) Using machine learning methods and financial statements, for each company we assessed the probability of a tax audit.

3) Using the Propensity Score Matching method, we built a control group that consists of companies that were not audited and that had the closest Score to the auditees.

4) We compared the performance of companies from treatment and control groups from the start of the audit and for 2-3 years afterwards.

Research stages, methods, results

At the first stage, initial datasets were formed for companies that have passed field tax audits in 2017-2021, and for companies that have not been audited. After the initial data cleaning procedure, the datasets amounted to 29,413 companies with checks and 1,348,954 companies without checks.

For the study we selected the following data on companies: region of registration, type of economic activity, organizational and legal form, form of ownership, size of the company[1] (large business, medium-sized enterprises, small enterprises, micro-enterprises); tax treatment, average number of employees, data on balance sheets for 2017-2021:

- Assets (current and net assets)
- Capital and reserves
- Non-current liabilities

[1] In the Russian Federation, micro-enterprises include companies with annual revenues of up to 120 million rubles and up to 16 employees; small enterprises include companies with annual revenues up to 800 million rubles and up to 101 employees; medium-sized enterprises include companies with annual revenues of up to 2 billion rubles and a staff of 101 to 250 people; large enterprises are all other companies.

- Current liabilities
- Revenue
- Cost of sales
- Commercial costs
- Management costs,
- Income from participation in other companies
- Other incomes
- Other expenses
- Interest receivable
- Interest payable
- Profit/ (Loss) before Taxation
- Net income or net loss

At the second step, the data was pre-processed: we excluded duplicates, deleted entries with gaps; some of indicators were taken as logarithms, which made it possible to bring the distribution of features to a normal form; we also created several auxiliary features and dummy variables. Datasets contain data only for companies for which tax audits ended in 2019 (this was done because such companies are characterized by the most complete cycle of events: in 2017 they were not exactly audited, in 2018 - either the beginning of the audit, or the beginning of preparatory verification procedures, 2019 is the year of completion of the verification, 2020-2021 is the subsequent years after the verification).

After pre-processing and data filtering procedures, the dataset contains 5,813 companies with inspections in 2019 and 397,254 companies without inspections.

At the third stage, we used a number of machine learning methods (logit model, Random Forest) to divide companies into two groups: “subjected to verification” and “not subjected to verification”. All companies, audited and unaudited, were assigned "probability of audit".

At the fourth step, matching was carried out according to the Propensity Score Index. For each company from the dataset with audits, a company from the dataset without audits

was selected which had the closest value of the predicted audit probabilities. The selection of the corresponding company without audit was such that we obtained one-to-one correspondence between companies from the dataset with audits and companies from the dataset without audits. The company size criterion was also taken into account during the selection of companies.

At the final stage, we compared the average performance of companies with and without audits, one year before the audit, in the year of the audit, in the year of the end of the audit, and within two years after the audit.

Results of the study

The study showed that field tax audits do impact the main financial indicators of the audited companies in comparison with similar companies that have not been audited. For example, a year before the start of the audit, on average, the revenue of companies that were subsequently subjected to audits is 13.4% more than similar companies without audits, and the revenue of companies that have been audited becomes 12.5% less than the revenue of companies without audits, and then 15.4% less two years after end of the check.

The numerical value of impact on performance depends on the size of companies, namely, large and medium-sized companies are much less sensitive to inspections than small and micro enterprises. The effect of tax audits strongly depends on the industry in which companies operate. For example, the least impact of audit on companies is observed in manufacturing, while companies in the trade sector are more affected by audits.

In addition to financial indicators, we analyzed other indicators such as the number of employees, and it turned out that the audit had practically no effect on the number of employees of the audited companies.

We note that the dynamics of totals of indicators for entire groups of audited and nonaudited companies turned out to be unidirectional (either growing or falling in the same way for both

groups). This may mean that the change in the difference between the indicators of audited and non-audited companies is achieved due to a faster (outpacing) growth of indicators of non-audited companies or due to a slower decrease in the corresponding indicators in case of their fall. This may indicate that field audits may influence not only the audited companies themselves, but also indirectly, due to various restrictions (including the finite number of companies on the market), on non-audited companies.

III. Analysis and Inspiration

We built a machine learning model to construct control groups with very high indicator of model evaluation metrics; in this study, the receiver operating characteristic was used (Receiver Operating Characteristic curve Area Under Curve – ROC AUC), the ROC AUC was above 0.93. This model can also be used in other cases when it is required to build control groups according to the Propensity Score Matching method, and use it when it is necessary to evaluate the impact of some treatment on the experimental group of objects in comparison with initially similar objects, but which were not treated.

IV. Questions

1. How large should the initial sample of companies for the control group be compared to the treatment group to apply this Propensity Score Matching method?

2. For what other audit tasks can the Propensity Score Matching method be used?

3. Can the model be modified to assess separately the impact on the treatment group and the impact on all other companies (industry of economic activity) that were not affected by the treatment?

(Written by Revenue Audit Department, Accounting Chamber of the Russian Federation)

22 Recursive Microsimulation Model & Long-term Poverty Level Forecast

Written by Research and Methodology Department,
Accounts Chamber of the Russian Federation

I. Introduction

Poverty reduction is one of the main national goals in Russia. Rapid economic growth in the 2000s reduced the very high levels of poverty left over from the 1990s, but since the 2008 global financial crisis, the average growth rate of the Russian economy has been much lower than in 2000-2007. Despite a moderate increase in household incomes after 2008, the level of poverty in Russia remains quite high.

The forecast of the Russian government for economic growth up to 2036 assumes that real GDP in 2036 is 1.7 times higher than in 2018, and real disposable income is 1.5 times higher. The pension reform started in 2019 proposes to increase the retirement age from 55 to 60 for women and from 60 to 65 for men. To evaluate the poverty level until 2030, taking into account the pension reform, we developed a recursive micro-simulation model. We use this model not only to predict the poverty level but also to assess the impact of certain social benefits introduction on the poor people, in particular on families with children. All simulations were carried out at the beginning of 2021.

II. Main Content

Audit Objectives

In 2018-2019, it was planned that the poverty rate would be halved by 2030. Economic growth is the main source of the poverty reduction. Therefore, it is necessary to assess whether the planned economic growth is sufficient to halve the poverty rate as it was expected. It is necessary also to simulate the introduction of social benefits for various groups of poor people and assess the poverty rate after introduction of these benefits. This

would help to evaluate the impact of various government measures to combat poverty.

Audit Methods

To evaluate the poverty rate until 2030 we constructed the recursive microsimulation model based on the Survey of Population Income and Participation in Social Programs database for 2017. This model has been constructed in such a way that:

1) It forecasts the households' level of income according to the predicted economy growth rates;

2) It takes into account the demographic forecasts for the next decade and reweights the sample in accordance with these forecasts;

3) It simulates the increase of the retirement age in Russia during the nearest years by imputing the labor income for those pensioners in the sample who would continue to work if the retirement age was increased.

Thus, we obtain the distribution of income of Russian population for every year up to 2030 so we can evaluate all necessary macroeconomic characteristics of the population income (including the poverty rate), simulate the introduction of social benefits for various groups of poor people and assess the poverty rate after introduction of these benefits. This allows us to evaluate the impact of various government measures to combat poverty.

Audit Steps and Results

We operate with the 2019 price level, so all estimates and forecasts are formulated in 2019 prices. This allows us to exclude inflation from our considerations.

We consider baseline and pessimistic scenarios of Russia's economic development based on the forecasts of various economic institutions. They differ in the rate of economic growth (3% and 1.5% respectively starting from 2023) and the pace of economic recovery in 2021-2022 after the COVID pandemic.

The following steps have been made recursively for each year 2020-2030 and for each scenario.

1. The first step is the determination of the demographic weights in the Survey of Population Income and Participation in Social Programs database (Russian abbreviation VNDN). We

use the algorithm of reweighting provided by Rosstat (Russian Federal Statistical Service) to re-calculate all sample weights in VNDN database in accordance with the total population in each gender and age group according to Rosstat demographic projections.

2. The second step is the imputing the labor income for those pensioners in the sample who would continue to work if the retirement age was increased. Here we use the following method.

a) Using the probit model we evaluate the probability of having a job for each prepension age people. Various socio-demographic characteristics of people are used as predictors.

b) For each person who can keep a job after raising the retirement age (for each person with a probability of keeping a job more than 1/2), we impute to them a labor income equal to the labor income of their working "nearest neighbor", which is found using the Mahalanobis metrics in the space of all socio-demographic characteristics contained in VNDN database.

3. The third step is the calculation of new income distribution and poverty rate according to the new labor incomes and new sample weights. Here we assumed that labor income grows in proportion to GDP growth with a coefficient of 0.8 chosen from the government's economic projections. For each year and for each scenario we obtain the corresponding income distribution and the corresponding poverty rate as the proportion of people who are below the subsistence level.

4. The last step is to evaluate some government measures aimed at combating poverty, namely, several types of social benefits for the poor: "flat" or fixed benefits for all lowincome families with children aged 8 to 18 (the same for all such families in the amount of one subsistence minimum for each child), "compensatory" benefits (specific for each such family in the 1.5 of amount of the child income deficit, but not less than 1000 roubles), and similar benefits for all low-income households ("flat" benefits for all low-income households in the amount of one subsistence minimum and "compensatory" benefits in the amount of the household income deficit).

For each scenario we calculate the poverty rate in each year (and particularly in 2030) without the social benefits at all and with each kind of social benefits considered above. For each kind of social benefits we also calculate the total volume of such benefits. This allows

us to evaluate the federal budget expenditure necessary to implement these social benefits.

Analysis Results

We obtain the following results: if all additional social benefits for poverty are not introduced, the poverty rate for 2030 would vary from 9.2% to 7.2%. This means that the national target of halving the poverty rate by 2030 (starting at 13% in 2017) would not be met, and the new targeted social benefits for poor people are needed.

The simulation of extension of social benefits to families with children aged 8 to 18 shows that "flat" benefits for poor families with children will reduce the poverty rate in 2030 to 4.24% in the baseline scenario and to 5.24% in the pessimistic scenario, and "compensatory" benefits for families with 8-18 years old children will reduce the poverty rate to 4.83% and 6.03% in the baseline and pessimistic scenarios respectively.

The simulation of introduction of social benefits for all poor families shows that that "flat" benefits for all poor families reduces the poverty rate in 2030 to 3.35% or to 4.3% in the baseline and pessimistic scenarios respectively and the "compensatory" benefits will completely combat the poverty: the immediate decline of poverty rate equals 100%.

In other words, the simulation results show that the risks of not meeting the national target of halving poverty by 2030 are very high, and targeted social transfers for the poor are needed to halve poverty by 2030.

III. Analysis and Inspiration

The highlight of the case is the use of the detailed mathematical microsimulation model to predict the income level and the poverty rate of Russian population for the next decade. Also this case demonstrates that this model can be used to analyze certain government measures aimed to combat poverty.

This model utilizes various mathematical methods, including some methods of mathematical statistics and econometrics.

IV. Questions

1. How one can differentiate the incomes growth rate for different branches of economy and different regions of Russia to use this data in the model?

2. How one can modify the model to get the more accurate one?

3. In order to combat the poverty what other targeted government measures could be introduced?

23 Case Study of Identifying Entities Overly Dependent on Government Funding

I. Introduction

Subsidies and government purchases are the biggest government budget expenses so auditing the efficiency of these expenses is important. We collected data for the 2021-2023 years and conducted the descriptive analysis of government funding. The key result of this analysis is identifying the entities which are overly dependent on government funding.

II. Main Content

Determine the Audit Objectives

The objective was to conduct descriptive analysis of subsidized entities, government procurement contractors and their financial statements.

Define Audit Methods

The audit method was descriptive analysis.

Audit Steps, Methods, and Results

1. Data on subsidy contracts, government procurement contracts and financial statements of entities over 2021-2023 was collected.
2. The dataset on subsidy contracts contains information about contracts between government and entities: id and name of an entity, subsidy amount and year. According to such contract, the government provides funding to an entity in return for meeting some goals specified in a contract. A contract may be revised, so only the latest version of a contract was kept. After that the dataset was grouped by an entity and a year.

3. The dataset on government procurement contracts contains id and name of an entity, price and year. The dataset was grouped by an entity and a year.
4. The dataset on financial statements contains id and name of an entity, assets, revenue, net profit and year.
5. These datasets were merged by an entity and a year.
6. A criterion was developed to identify entities overly dependent on government funding: an entity should sign at least 2 subsidy contracts and at least 2 government contracts each year during 2021-2023. Using this criterion, the descriptive analysis was conducted.

Analysis Results

The list with the entities overly dependent on government funding was sent to other departments to facilitate planning the performance audit next year.

III. Analysis and Inspiration

This case illustrates the value of combining several data sources for audit purposes. Focusing on subsidy contracts alone may have led to incorrect conclusions about which entities depend on government funding most.

IV. Questions

1. What other criteria could be used to identify entities overly dependent on government funding?

2. What other datasets could be used to enrich the dataset described in the case?

(Prepared by Department of Research and Methodology, Accounts Chamber of Russia)

24 Case Study of Project "First Aid" Analysis

Package of economic measures to help employers, businesses and the self-employed persons during the pandemic

I. Introduction

Due to the coronavirus pandemic, many entrepreneurs in the Slovak Republic (SR) had to close their businesses or significantly limit their activities based on the decision of the Public Health Authority of the SR. Subsequently, employers began with collective dismissal of their employees and many self-employed people ended or suspended their business activities. As a response to this situation, on March 31, 2020, the Government of the SR approved a scheme of covid help to preserve employment by the Ministry of Labour, Social Affairs and Family. The project was called "First Aid" and its main goal was to provide employers with support to retain their employees and self-employed persons with support to maintain the operation of their activities.

The Supreme Audit Office of the Slovak Republic (SAO SR) organized a special audit, the aim of which was to check the timeliness and transparency of the provision of First Aid at four ministries - Ministry of Labour, Social Affairs and Family, Ministry of Economy, Ministry of Transport and Ministry of Culture.

The Strategy and Analysis Department at SAO SR summarized the data provided by the Ministry of Labour, Social Affairs and Family, Social Insurance and Financial Administration. Based on these data, data analysis was performed. The main goal of the data analysis was to verify the fulfillment of the selected conditions for the provision of subsidies from the First Aid schemes for the period from March 2020 to June 2022, including a warning about some inconsistencies in the data provided. The analysis also focused on the amount of money provided through this project within the regions (at the level of the SR districts) and

also whether there were duplicities in the simultaneous receiving of subsidies from different ministries.

II. Main Content

Determine the Analysis Objectives

The First Aid project was a response to the reduced operating conditions of businesses in Slovakia during the COVID-19 crisis. Due to the difficulty of the situation, the Ministry of Labour, Social Affairs and Family, in the implementation of this project, prioritized the provision of the help as quickly as possible over the rigorous control of the data and information of the applicants.

The purpose of the analysis was to draw attention to risks and problems that arose during the realization of the First Aid project. The results of this analysis could also be helpful in the future when introducing and implementing similar types of projects even in crisis situations, such as the COVID-19 crisis.

Define Analysis Methods

The analysis was carried out on the basis of the following data:

- Data collected for the purpose of providing the covid help by the Ministry of Labour, Social Affairs and Family,
- Tax related data by the Financial Administration of the Slovak Republic
- Employment data by the Social Insurance Company
- Data related to the provision of covid help by other ministries (Ministry of Economy, Ministry of Transport and Ministry of Culture)

All these data served as the basis for the creation of a summary database. The summary database was subsequently sorted out by the analysts and the analysis of fulfillment of the individual conditions for the provision of covid help for businesses was conducted.

Analysis Steps, Methods, and Results

The analysis steps were as follows:

(1) Study of conditions for providing covid help and find out the potential weaknesses. The covid help under the project First Aid was provided through six different measures. Thorough study of the individual conditions of the given measures served as a basis for the selection of conditions for a more detailed analysis.

(2) Checking the fulfillment of selected conditions:

a) Eligibility of the applicant (Since in the conditions of the given measures the latest date on which the company could have been founded in order to be entitled to draw the allowance was stated, comparison of the date of submission of the application for the allowance with the date of establishment of the company was carried out. Almost all detected discrepancies in this part of the analysis were defined by the ministry as incorrectly entered data in the information system, which was subsequently corrected by the ministry. Due to ineligibility, only a minimum of recipients (0.04%) had to return the allowance provided.)

b) Target group for measures for employers (In this part the number of employees with the number of supported employees was compared. Discrepancies were found in 5.8% of cases from the three introduced measures. During the communication with the Ministry of Labour, Social Affairs and Family, ministry stated that the error could have been on the side of the employer, who could have typed the value incorrectly when submitting the application, or on the side of the officer who processed the application. Also the number of employees unable to work was checked. In this part analysis showed that there were minimum number of such cases.)

c) Conduct analysis of tax conditions (As part of the tax conditions, analysts have been finding out whether the recipient is a tax debtor, whether he has fulfilled the registration obligation as a payer of income tax and whether he is a payer of VAT. The analysis shows that 1.1% of recipients are on the list of tax debtors, 3.3% are not income tax payers and up to 90.4% are not VAT payers, which confirms the fact that the largest group of

applicants were micro businesses.)

d) Conduct analysis connected to quality of data provided by Ministry of Labour, Social Affairs and Family

Three discrepancies were found out during the processing of data provided by the ministry:

- regarding company address – Based on the matching of addresses from the total number of 171,871 business entities, 27% were located at an address that appeared in the address list of recipients two or more times
- the date of payment of the financial contribution was earlier than the effective date of the agreement
- he paid amount of covid help was bigger than the requested amount.

As the reason for the errors concerning the date of payment and the paid amount, the Ministry of Labour, Social Affairs and Family stated that the data was wrongly typed into the system by labor office workers and also by applicants. They were subsequently corrected in the information system after mutual consultations with the SAO SR, and thus a substantial part of these irregularities were eliminated.

(3) Overview of the providing of covid help in the regions of Slovakia (Analysis shows that the largest recipients of “First Aid” can be found in the western part of Slovakia in the region of the capital city Bratislava and on the contrary, the least in the eastern part of Slovakia.)

(4) Simultaneous providing of subsidies from other ministries - Ministry of Economy, Ministry of Transport and Ministry of Culture (Duplicative drawing of subsidies from several ministries at the same time can also be uneconomical. The result of connecting data from individual ministries was that 4.2% of applicants drew subsidies simultaneously also from Ministry of Transport, 8.7% also from Ministry of Economy and 0.3% also from Ministry of Culture.

Analysis Results

The data analysis indicated that there were missing some important data in the database provided by the Ministry of Labour, Social Affairs and Family and also some data entry errors occurred. This could have resulted in the subsequent unauthorized provision of subsidies. In addition, the linking of the databases of subsidies from other ministries showed that some recipients of subsidies received help from several support schemes simultaneously. Another interesting finding is that many recipients resided at the same address, which can be a subject of audit in the future as well.

Strategy and Analysis Department suggested increasing the quality of data by automation of systems used and digitalization of the entire support scheme. The next step to increase the efficiency of support schemes is the linking of information systems of all ministries, but also mutual cooperation in the conception and designing of any support and subsidy schemes in order to prevent their use for the same or similar purposes.

The added value of this data analysis is also that all these findings that are result of the data analysis can be used when setting up new aid schemes in the future.

III. Analysis and Inspiration

The highlight of the case is the processing of huge amount of data from multiple sources in precise identification of difficulties in the implementing and providing of covid help, which provides an inspiration on how to improve the system and method of realization of government help in the future.

IV. Questions

1. How to ensure the connection of information systems of individual ministries?

2. In order to maintain the quality of data, consider the introduction of control mechanisms built into the information systems of ministries.

(Written by Strategy and Analysis Department, Supreme Audit Office of the Slovak Republic)

25 The Quality of Life in the Slovak Municipalities

I. Introduction

Measuring the quality of life is a complex process that plays a crucial role in assessing living conditions in municipalities. The quality of life depends on various factors, such as access to education and healthcare, as well as issues related to safety, unemployment, environment, etc. These factors directly affect residents' daily lives, making regular monitoring essential. The collected data allows local governments to identify areas that need improvement and create policies that enhance community development and well-being.

The Supreme Audit Office of the Slovak Republic (SAO SR) has been able to audit self-government of municipalities since 2006, when its audit mandate was expanded. To contribute to better handling of public affairs information and allow citizens to assess and compare the quality of life in municipalities from various perspectives, the analytical tool named MuMap (Municipalities on Map) was developed by the Strategy and Analysis Department at the SAO SR. Through various sections, users can explore data on indicators like average debt per capita, EU-funded project financing and audit findings. The app's slogan, 'Everyone can be a good auditor', reflects its philosophy of encouraging active citizen engagement in public affairs. The latest feature, 'Quality of Life', includes an interactive map comparing Slovak municipalities based on diverse quality-of-life indicators. This tool provides clear and informative insights, allowing anyone to access data about local conditions and identify areas for improvement.

II. Main Content

Determine the Analysis Objectives

A thorough examination of factors such as healthcare, demography, social care, education, safety, economic activity, culture, public engagement, and environmental conditions is essential for assessing the quality of life. Each of these elements plays a significant role in shaping the daily lives of residents. This analysis helps municipalities evaluate and compare their performance, providing insights into areas that may require improvement. Local authorities can track progress and learn from other regions facing similar challenges.

Define Analysis Methods

To analyse the quality of life in Slovak municipalities, a well-organized and methodical approach is necessary. This analysis draws on a variety of data sources, including demographic, economic, healthcare, educational, safety, environmental, public affairs participation, social care, and cultural information. By combining these diverse datasets, a comprehensive assessment of living conditions across municipalities is achieved.

Analysis Steps, Methods, and Results

The analysis involves several key steps to ensure effective data collection, processing, and visualization.

(1) Data Collection.

The quality of life in municipalities is evaluated through a comprehensive set of indicators across multiple sectors. These indicators reveal regional inequalities and highlight areas that need attention. However, it is important to note that the data used for this evaluation is limited by what was available and accessible. In some cases, we were unable to include certain indicators due to the lack of quality or reliable data. For example, many requests suggested that data on sports would be relevant, but we have not yet been able to obtain them. These limitations reflect the challenges in gathering comprehensive data for all areas of interest.

Quality of life is assessed based on these indicators in the following areas:

- Demography: Change in the population size.
- Economic Activity: Unemployment rate.
- Healthcare: Distance to the nearest hospital.
- Education: Kindergarten capacity, Distance to the nearest primary school, Distance to the nearest secondary school.
- Safety: Number of crimes per 1,000 inhabitants.
- Environment: Average degree of thread from extreme weather, Distance from the landfill.
- Interest in Public Affairs: Participation in local government elections.
- Social Care: Distance to the nearest social facility for seniors.
- Culture: Number of leisure facilities for children and youth, Number of libraries, Number of theatres, Number of cinemas, Number of museums and galleries.

In some cases, data availability influenced methodological decisions. For the two largest cities in Slovakia – Bratislava, the capital of the Slovak Republic, and Košice, the country's second-largest city – data were analysed at the level of individual city districts rather than in aggregate form. While this approach allowed for a more detailed comparison, it could be perceived as a technical limitation in assessing the overall quality of life in these two major cities.

(2) Data Processing.

The data is processed and linked with geographic information that defines the boundaries of individual municipalities.

During data processing, we encountered an issue involving name duplications among Slovak municipalities. Specifically, there are 205 municipalities in Slovakia that share the same name with at least one other municipality located in a different district. This issue was resolved by utilizing a unique identifier – municipality code, allowing us to distinguish between municipalities with identical names.

In contrast to other challenges, we avoided the issue of incompatibility between the technical solution and the data structure. This was achieved by adapting the technical solution to fit the data structure, and vice versa, adjusting the data structure according to the technical requirements. As a result, the technical solution and the data structure were seamlessly integrated and aligned.

To obtain the distance-based indicators, it was necessary to calculate the distances between municipalities and specific institutions, such as hospitals, schools, and other facilities. We used aerial distances, or 'great-circle' distances, which are calculated based on the geographic coordinates (latitude and longitude) of each municipality and the nearest municipality with the relevant institution.

In addition to processing the data, a scoring methodology is applied to rank municipalities based on their quality of life. The methodology follows these steps:

a) Each indicator is sorted from best to worst based on optimal values. For example, lower unemployment rates are considered better, while a longer distance from the landfill is preferable. It is also important to specify that all indicators are equally weighted.

b) For each indicator, a point rating is calculated – with a maximum of 3,000 points for the entire area. For example, if an area is represented by 3 indicators, each of them has a 1,000-point fund. The maximum number of points per indicator is divided by the number of unique values of the indicator – this creates one step. Then, the best value of the indicator is assigned the highest possible number of points, and the remaining values are assigned points according to the following procedure:

1st place = x1

2nd place = x2 = x1 - step

3rd place = x3 = x2 - step

....

n-th place = xn = x(n-1) - step

c) If an area is represented by multiple indicators, we combine them. If the area contains indicators with the same nature (i.e. for all of them, the lower the value, the better, or

conversely, the higher the value, the worse), the points for the indicators are summed, resulting in the point rating for the area. However, if the area consists of indicators with different nature (i.e. for some, the lower the value, the better, and for others, the higher the value, the better), the following procedure is applied:

The points for the indicators are summed, and the unique point values for the indicators are sorted from the highest to the lowest. As in point b), the point rating for the entire area (max 3,000) is then calculated.

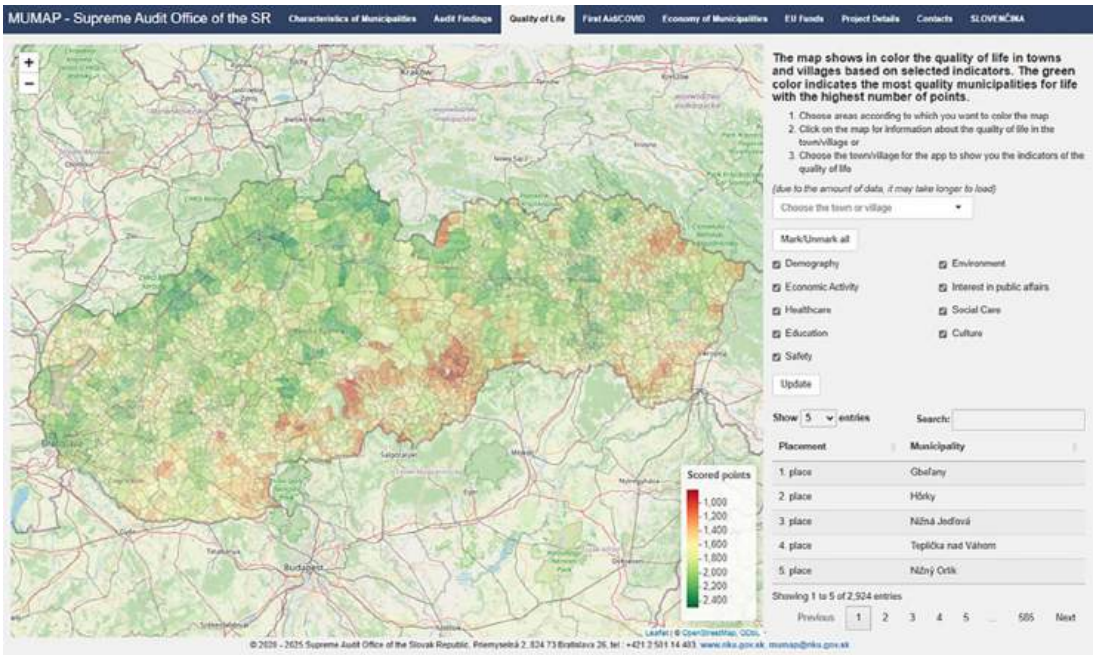
d) The final step of the calculation involves creating the ranking – the higher the point rating, the better the ranking position of the municipality.

(3) Data Visualization.

The processed data is visualized using R software, leveraging Shiny – a web framework for R and Python – and the Leaflet package, which offers powerful features for integrating with Shiny applications. The result is a map that provides an overview of the quality of life in individual municipalities in Slovakia, allowing for comparative analysis.

The data visualization itself was carried out using the Leaflet package, which enables the creation of dynamic and interactive maps. The individual municipalities were represented as polygons, and their colours reflected the quality of life score based on selected indicators. The user has the option to obtain detailed information about each indicators through interactive tooltips. The figure N°1 shows the user interface of the a

Figure N°1
Visual Representation of the User Interface.



Source: Own elaboration

One of the main advantages of the map is its interactivity, which allows users to customize the displayed information according to their preferences. The user interface includes dropdown menus and checkboxes. The dropdown menus let users select a municipality, and once chosen, the map zooms in on it and displays a marker. The checkboxes allow users to select specific areas. When the areas are selected or modified, the quality of life score is recalculated, and the areas within the municipalities are repainted based on the new score, after clicking the update button.

A data table was also added within the layout of the app, representing a ranking of municipalities. The user can easily interact with this table by using the search box.

Analysis Results

The interactive quality of life map provides an effective tool for visualizing complex data on the quality of life in Slovak municipalities. For policymakers, it can serve as support in deciding on regional development priorities. Researchers will appreciate the ability to analyse individual indicators and their interrelationships, while the general public will gain insight into the factors affecting the quality of life in their community.

III. Analysis and Inspiration

The analysis highlights the innovative use of diverse data sources to assess and compare the quality of life across Slovak municipalities. By integrating indicators from healthcare, education, environment, demographic, economic activity, safety, public affairs engagement, social care and culture, the project offers a comprehensive and actionable view of regional disparities. Inspired by the need for transparency and effective governance, this tool demonstrates how technology can transform complex datasets into practical solutions for community development.

IV. Questions

1. Can the ranking of municipalities based on quality of life metrics influence migration trends within the country?

2. In what ways could the insights provided by the quality of life analysis influence local governance strategies and decision-making processes and promote greater public engagement and participation in community development?

(Written by Department of Strategy and Analysis, Supreme Audit Office of the Slovak Republic)

26 Case Study on Data Analysis of Corporate Tax Expenditure Practices in Türkiye

I. Introduction

Tax expenditures are an indirect form of government spending, achieved through the tax system rather than direct expenditures. It includes various tax breaks, exemptions, deductions, credits, and deferrals granted to individuals, corporations or other entities, for the purpose of reducing their tax burden. Tax expenditures are a key fiscal policy tool and constitutes a considerable percentage in the macroeconomic indicators. According to the IMF , the ratio of tax expenditures to GDP can reach levels approaching % 15 in recent years. In 2022, total tax expenditure is 5,12 % of GDP of Türkiye. For this reason, Turkish Court of Accounts (TCA) has carried out financial and compliance audits on this field for years.

The Turkish Revenue Administration's digitalization efforts have led to the creation of digital platforms for collecting tax returns, financial statements, journal books, and other relevant taxpayer documents in digital format. This digitalization has encompassed all corporate taxpayers. As a result, TCA audit team, responsible for conducting financial and compliance audits of corporations, leveraged this digital data to holistically analyse corporate tax expenditure practices against existing tax laws and regulations

By utilizing data analysis techniques, TCA aimed to analyse a vastly greater number of transactions across a wider geographical area in a shorter timeframe, leading to more efficient and effective audits. Furthermore, the audit team improved compliance monitoring, refined risk assessment, detected fraud patterns, and reduced audit risk.

II. Main Content

Audit Objectives

While utilising data analytics as part of audits conducted on corporate tax expenditure practices, our audit objectives were focused on ensuring regulatory compliance and improving the accuracy, fraud detection and efficiency in the processing of these claims. Details of these main audit objectives are as follows:

1. Regulatory Compliance

- Verifying that corporate tax expenditure practices comply with applicable tax laws, regulations, and policies.
- Ensuring that all claims meet the eligibility criteria for corporate tax expenditures.
- Assessing whether tax authorities and claimants follow standard procedures for filing and approval.

2. Accuracy and Validity of Claims

- Verifying that all data fields on tax returns of corporate tax payers (e.g., income and expenses, exemptions claimed, supporting documentation) are accurate and complete.
- Verifying that all corporate tax expenditures granted are consistent with supporting documentation (e.g., invoices, tax returns, receipts).
- Identifying and correcting errors or discrepancies in recorded tax amounts.

3. Fraud and Anomaly Detection

- Identifying duplicate claims, fraudulent claims, or exaggerated exemptions.
- Assessing whether claimants are abusing tax exemption codes or applying for ineligible refunds.

4. Risk Assessment and Internal Controls

- Assessing whether tax authorities have adequate monitoring mechanisms to detect and investigate suspicious claims.
- Assessing whether automated systems and internal controls effectively reduce manual

errors and processing delays.

- Identifying risks in workflows and recommend process improvements.
- Evaluating the robustness of internal control systems and identifying possible areas of improvement.

5. Revenue Protection

- Ensuring that corporate tax expenditures are justified and do not result in unnecessary financial losses for the government or organizations.
- Measuring the total financial impact of improper corporate tax expenditures.

Steps & Data Analytics Method

A. Criteria for Data Analyses

Audit team defined the specific criteria based on the tax laws, rules and regulations to be applied for the tax expenditure claims, filling the tax returns and financial statements of corporate taxpayers. There were many tax incentives implemented through tax expenditures. Each practice includes different terms and conditions to be applied in tax returns or financial statements. Those terms and conditions served as the primary criteria for evaluation. In addition, internal controls of tax expenditure implementation were considered as criteria to conduct a systematic risk-based approach.

B. Collecting of Data

Corporate tax expenditure practices in Türkiye are managed by Turkish Revenue Administration (Revenue Administration) all over the country and involves approximately one million corporate tax payers. This results in a vast volume of data that must be analysed to ensure that corporate tax expenditures are applied in accordance with the established laws and regulations.

Revenue Administration implemented an “e-tax return” system in order to collect required data from taxpayers. This application provided TCA audit team with complete and accurate data before performing data analysis. Since the establishment of the system, around one

million corporate tax returns and relevant documents has been uploaded to this system annually.

Audit Team communicated with Revenue Administration to collect data and ensure the maintain of tax privacy. Parts of the tax returns and financial statements containing personal data were encrypted by assigning unique IDs to corporate taxpayers. Afterwards, audit team obtained digital documents, tax return and financial statements for each corporate taxpayer. The format of the data, name of the rows and columns were determined and standardized by the audit team. Finally, data received via cloud was uploaded to the TCA database in a predetermined format as a csv file.

C. Data Cleaning & Pre-processing

Before starting the audits, the data was cleaned and pre-processed to ensure accuracy, increase the reliability of the analysis and ensure that it was free from distortions caused by inconsistencies in the data set. For this purpose, 100 sample digital corporate tax returns were selected. Duplicate rows were detected, values such as dates and monetary amounts were standardized, and missing or incorrect fields in the data are identified.

Identifying outliers in corporate tax expenditure requests was also an important initial step. Since extreme values could indicate fraudulent activities or errors, the declarations with the highest values were separately considered before the analysis.

D. Data Analytics

After having collected the data from Revenue Administration, the audit team used data analytics to understand whether tax laws, rules and regulations were properly applied to tax returns and other relevant data. The data analyses were mainly conducted through a customized business intelligence tool of TCA called VERA. VERA allowed the audit team to use some algorithms and SQL queries in order to analyse the data set. There are also other tools and technics used as follows:

- Descriptive Analyses for Initial Insights

  Descriptive analyses were used to summarize, visualize, and identify patterns in data. In the auditing process, these techniques were used to detect anomalies,

assess compliance, and improve audit efficiency. For example, high-risk claims were identified by analysing claims history, refund or deduction amounts, and claim patterns. Techniques were used to identify businesses that share the same address, accountant, or email domain and are requesting excessive refunds, indicating possible collaboration. Additionally, tax professionals who repeatedly submit suspicious requests on behalf of multiple clients have been put under scrutiny as they may enable fraud. Finally mapping tax-related data to geographic locations, auditors uncovered hotspots of suspicious activity, and identified anomalies in tax claims across regions and potential fraud networks operating in specific areas.

- Data visualisation

  The audit team used visualisation techniques such as charts, graphs and heatmaps to identify patterns of claims and filings over time. By using this method, unusual fluctuations in reported income or deductions were easily detected. In addition, situations such as unusual spikes in reported income or deductions were detected.

- Statistics

  Taxpayers' reported income, expenses, and deductions were compared to industry averages. In addition, outliers in reported values (e.g., abnormally low or high tax payments) were detected.

- Trend Analysis

  The audit team used this technique to detect sudden drops in taxable income that may indicate underreporting. Also, tax payments of businesses and individuals were monitored over multiple years for inconsistencies.

- Cross-Validation with Third-Party Data

  Auditors matched taxpayer data with external sources such as bank records (for example cross-checked reported sales with POS data).

- Text Mining for Document Verification

  Many tax refund claims contain supporting documents such as financial statements, tax receipts, or contracts that can be analysed using text mining. By scanning these documents for inconsistencies, auditors detected discrepancies and reversals in account balances, and manipulations of income and expense items.

- Rule-Based Analytics

  Another effective approach for detecting fraudulent claims involved creating predefined rules and thresholds. These rules flagged refund requests that exceeded a reasonable percentage of taxpayer-reported tax payments and identified multiple exemptions granted to the same taxpayer or refunds processed without sufficient supporting documentation. For example, an unusually high percentage of processed refunds was taken as an indication that the approval process lacked adequate validation. Therefore, implementing rule-based analytics allowed auditors to set automatic alerts for high-risk claims, thereby increasing efficiency in fraud detection. VERA was often used for the ruled-based queries. Thanks to VERA, rules and thresholds were identified as algorithms and queries to obtain high risk /red flag outcomes

Analysis Results

As a result of our comprehensive analyses, we have reached significant findings regarding corporate tax expenditures practices, our key observations include:

- Decrease in corporate tax base due to non-compliance with financing expense restriction
- Decrease in corporate tax base due to exceeding the legislative limits regarding venture capital funds by corporate taxpayers
- Decrease in the tax base due to provisions exceeding the amount of trade receivables
- Exceeding the amounts of tax reductions specified in the legislation
- Different profit amounts in the corporate tax return and financial statements for the same corporation

Based on the data analysis results, it was observed that some tax expenditure practices have been incorrectly implemented. While some of these errors were observed nationwide, others were found to be concentrated in specific cities, regions or sectors. Additionally, it was determined that the errors are largely related to taxpayers filling out incomplete or incorrect tax returns or tax claims. Some of the errors, however, have been caused by deficiencies in internal controls over taxpayers' tax returns or financial statements.

III. Analysis and Inspiration

Big data analysis conducted on tax expenditures has also been a pioneering study for audits of other tax practices. In this case study, the collection of nationwide data, attention to data privacy, data cleaning, and the use of a customized business intelligence tool (VERA) of TCA are noteworthy. In this way, by developing analyses nationwide, an analytical understanding has been created regarding the geographical distribution of problematic issues and potential solutions.

IV. Questions

1. What are the main difficulties in collecting the data on tax expenditures analyses?

2. What kind of other analytical tools or technics may be useful in the analysis of tax expenditures?

3. Do you think data analysis can be beneficial in the auditing of all type of tax practices?

(Written By Data Analysis Department, Turkish Court of Accounts)

27 An example Performance audit on the use of public funds for achieving the goals of the Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and their Disposal

I. Introduction

The rapid economic development and scientific and technological progress of the last decades cause the growth of hazardous waste generation and accumulation in all countries of the world, including in Ukraine.

This determines the extreme importance of controlling hazardous waste, namely, its collection, storage, processing, utilization, removal, disposal, burial and transportation, including transboundary.

Aware of the growing threat to human health and the environment, 187 countries of the world have joined the Basel Convention, which is primarily aimed at protecting the environment and human life and health from exposure to hazardous waste. Ukraine, as a party to the Basel Convention, takes measures to achieve its main goals, but the growth of waste accumulation, including hazardous waste, continues to create significant risks of harming both the environment and the health of the population of Ukraine.

II. Basic content

Defining the purpose of the audit

The purpose of the audit: establishing the actual state of affairs and providing an assessment regarding:

- implementation of the recommendations of the Accounting Chamber based on the results of the previous state external financial control (audit);
- Ukraine's fulfillment of the requirements and achievement of the goals of the Basel Convention on the control of transboundary transportation of hazardous waste and its

disposal;

- legality, timeliness and completeness of management decisions in this area;
- productivity, economy and effective use of budget funds.

The audit covers 2017-2018 and the first half of 2019.

Audit methods

Methods of data collection used during the audit: analysis of regulatory and legal acts, administrative and official documents related to the subject of the audit, documents on the organization of work and performance of tasks assigned to the objects of audit, statistical, financial and budget reporting; carrying out external control measures at specified audit objects; survey, sending requests, examination and comparison, receiving oral and written explanations of officials of audit objects.

In particular, the objects of the audit were the Ministry of Ecology and Natural Resources of Ukraine (since September 3, 2019 - the Ministry of Energy and Environmental Protection of Ukraine), the Volyn, Odesa and Lviv customs offices of the State Fiscal Service of Ukraine. Requests for information were sent to the following bodies: State Fiscal Service of Ukraine, State Environmental Inspection of Ukraine, State Statistics Service of Ukraine, National Police of Ukraine (letter, regional state administrations.

Stages and results of the audit

1. The requirements of the Basel Convention regarding regulatory and legal support for its implementation are fulfilled by Ukraine as a whole. At the same time, the adopted legislative and by-laws need further improvement and coordination among themselves, in particular, regarding proper control over the transboundary movement of waste, its identification and removal

2. The Ministry of Ecology and Natural Resources of Ukraine is designated as the competent authority and designated center for the implementation of the Basel Convention. The state body took measures to fulfill the requirements of the Convention, but they turned out to be insufficiently effective for the proper implementation of state policy in this area

3. The State Environmental Inspection of Ukraine was not empowered to carry out

environmental control at checkpoints across the customs border of Ukraine. Normative regulation was also needed regarding the mandatory sealing of the container with hazardous waste upon entry into Ukraine, as well as the verification of the integrity of the seal at the destination on the territory of Ukraine and upon exit from the country. In addition, there is a need for regulatory regulation of issues of control of the actual weight of hazardous and other waste at customs posts.

4. The audit also testified to the need to strengthen control over the cross-border transportation of hazardous waste and compliance by business entities with the licensing conditions for conducting such activities, including by increasing the licensing requirements for the material and technical base available to business entities and its pre-licensing inspection.

Audit conclusions

The results of the conducted audit showed the need to increase the effectiveness of control over the transboundary movement of waste and to make it impossible to leave it uncontrolled on the territory of the country through which international transit is carried out.

This requires amendments to the provisions of the national legislation of neighboring countries to clearly define the responsibilities of the responsible bodies for waste control, the scope and method of such control, as well as the procedure for interaction between these bodies.

In addition, border guards must have the necessary equipment to accurately determine the amount and type of waste being transported.

The audit showed that the current control at the border crossing points was often formal, when transporting waste, only documents were mostly checked, and there was no physical inspection of compliance with the type and quantity specified in these documents.

III. Analysis and inspiration

The conducted audit will contribute to the elimination of risks, in particular regarding the occurrence of unauthorized landfills of hazardous waste, as the regulations in force in this area provide an opportunity for business entities to direct extremely hazardous waste for

burial in the ground or dumping on the ground (for example, at a landfill) and other storage. One of the main such reasons was the insufficient exchange of data between the state bodies authorized to carry out control in this area, in particular regarding the volumes of import, export and transit of hazardous waste.

Therefore, this indicated that the authorized body for fulfilling the requirements of the Basel Convention was not sufficiently informed about the implementation of operations in the field of waste management.

IV. Question

1) How to quickly determine the compliance and consistency of the data of state bodies authorized to carry out control, in particular, regarding the volumes of import, export and transit of hazardous waste?

2) How to quickly obtain relevant data on the volume of waste accumulation from external sources for comparison with official data?

28 Applying data analysis to the national insurance database in auditing of the State Audit Office of Viet Nam

I. Introduction

National Insurance Database stores information on social insurance, health insurance, unemployment insurance, and information on health and social security which is approved by the competent authority to recognize and ensure the insurance rights and obligations of citizens. The National Insurance Database is the Government's database, which is built uniformly across the country, shared by agencies, organizations and individuals in order to provide information on insurance in accurately and timely manner to serve the state management, meet the requirements of socio-economic development, agencies, organizations and individuals.

The National Insurance Database includes the following information: (i) basic personal data; (ii) citizens' contact information; (iii) household information group; (iv) information group on social insurance; (iv) group of information on health insurance; (v) group of information on unemployment insurance; (vi) group of information about the employer; (vi) group information on basic medical; (vi) group of information on social security. Thus, the national insurance database is a very important source of comparative information in auditing of the State Audit Office of Viet Nam (SAV). Therefore, SAV has gradually applied exploitation and analysis of national database on insurance in auditing.

This case study will update the latest audit results related to the application of exploitation and analysis of National Insurance Database in auditing activities up to the end of 2024.

II. Main Content

Determine the Audit Objectives

The Vietnamese government has always prioritized policies to support the poor, improve people's living standards, and create favorable conditions for the poor and disadvantaged people through credit support policies. In which, credit support is often associated with each household, their wealth and poverty status, and working status of beneficiaries of national incentives. The audit objective associated with these credit support policies is to ensure that the right beneficiaries are entitled to the credit incentives as: each household receives incentives once (there is no case as both of husband and wife enjoying incentives); preventing people with good incomes to enjoy credit incentives for the poor; preventing people who already have jobs to continue to enjoy job-searching incentives and special incentives for students. The audited entity is the Vietnam Bank for Social Policies.

In 2024, the State Audit of Viet Nam will continue to expand the audit content associated with the comparison of social insurance expenditure data associated with: (i) Expenditure for employees who want to enjoy the one-time social insurance policy (when they decide not to work and will no longer pay social insurance, so they will receive a portion of the social insurance they paid in the previous period); (ii) Expenditure for employees to receive a subsidy while unemployed; (iii) Expenditure for agents to collect social insurance for the first year with a preferential rate higher than normal, when there is a case where the employee returns to pay social insurance according to regulations. All three of these cases are associated with the risk of the employee returning to work earlier, so the subsidy and preferential rates will have to be reduced accordingly. This means that if the above expenditure is paid in advance, after verification, any overpayments resulting from benefits or preferential rates received in advance must be repaid to the social insurance fund.

Define Audit Methods

The auditing method for this content is to collect and analyze the contents of individuals entitled to credit policy preferences and to collect basic information of person, household,

job from the National Insurance Database to point out cases containing a risk of illegal lending and conduct a detailed check on the files of the above subjects.

For social insurance expenditures in 2024, the State Audit Office of Viet Nam will analyze data on social insurance benefits enjoyed by employees and agents in monthly reports throughout 2024.

Audit Steps, Methods, and Results

1. Collecting accurate data about beneficiaries of credit incentives from the national database on social insurance, (which is detailed down to each employee, according to the agent at the end of the month with specific benefits).

Solving this matter meets with the following difficulties: (i) Information about the above subjects (mostly the poor) stored at Vietnam Bank for Social Policies is inaccurate and incomplete (Missing data on ID card, ID number); (ii) Only access to the subjects of the audited content; (iii) Information security must be ensured during data accessing; (iv) Technical requirements must be ensured because the national data of the above subjects is very large. (v) Collection data will become more detailed, down to the level of each month, social insurance payers, and employment status.

Implementation steps include:

(1) Obtain identification information (ID card, date of birth, gender, ...) of beneficiaries of credit incentives at Vietnam Bank for Social Policies

(2) Coordinate with Vietnam Social insurance to review and obtain correct information of the above subjects, in which, attention must be paid to the missing information.

(3) Clean and standardize the information of the national database on social insurance

(4) Clean and standardize monthly social insurance expenditure data for 2024.

2. Conduct data analysis

(1) Conduct household data and household lending analysis;

(2) Conduct employment data and employment concessional lending analysis;

(3) Conduct data analysis to test on current job and income levels with cases of debt exemption due to disappearance, unemployed, poor, etc. with income data reported (reflected in salary).

(4) Conduct analysis and check the status of employees with the level of social insurance support received, compare the status of employees on a monthly basis to find out any unreasonable levels of benefits.

(5) Conduct analysis of data on percentage and duration of preferential rates for social insurance collection agents for collecting on behalf of new social insurance cases with the status of employees.

3. Transfer analysis data to each audit team for detailed checking

(1) Ensure audit teams only access to information related to the scope of work they are in charge of.

(2) Compare with paper records on the labor status of employees, data for collection agents, and employees with electronic data.

4. Conduct support to analyze new arising problems

The data analysis expert team will support the audit teams to detect arising problems, including: analyzing the new audit findings, recalculating the analysis results on the basis of the work results of the audit team, providing technical support if needed.

Analysis Results

Analysis of social insurance data in the above case has shown mistakes of lending to poor households such as: both receiving poor households and receiving incentives for households to escape poverty (typically both of husband and wife borrow). Lending to the poor and to the unemployed is still exploited by some individuals with high incomes and public servants who are dishonest in their declarations to receive credit incentives. The data

on the recruitment of employees of the enterprise also shows the enterprise, thereby further determining whether the enterprise is really facing difficulties to enjoy the Government's preferential credit policy. The auditors recommend that the Vietnam Bank for Social Policies should also study the overall review to avoid the above errors.

Regarding social insurance expenditures in 2024, SAV has proposed to recover all expenditures that are not in accordance with the expenditure ratio and benefit period through review, analysis and comparison of data associated with findings through data analysis nationwide.

III. Analysis and Inspiration

The highlight of this case is the use of the national database on social insurance, which has a lot of useful information about person, household, business, employment status, medical examination, so the ability to compare data sources of other audited entities is different. However, the detail and plenty of the above data requires the use of information to ensure information security and be strictly controlled to avoid any abuse and exploitation without authority.

For social insurance expenditures, the application of SAV's analytical techniques is more detailed than previous analytical methods when the subject of the comparison review is conducted in detail for each month.

IV. Questions

1. How to set up and use the national database of social insurance quickly and effectively in association with existing audit content but must ensure information security?

2. Research to expand the analysis and use of the national database of social insurance to other audit contents?

3. How to continue expanding the capacity to analyze social insurance data to a real-time level and closely link it with shortening the annual audit into reports with shorter time to improve the efficiency and effectiveness of fund use?