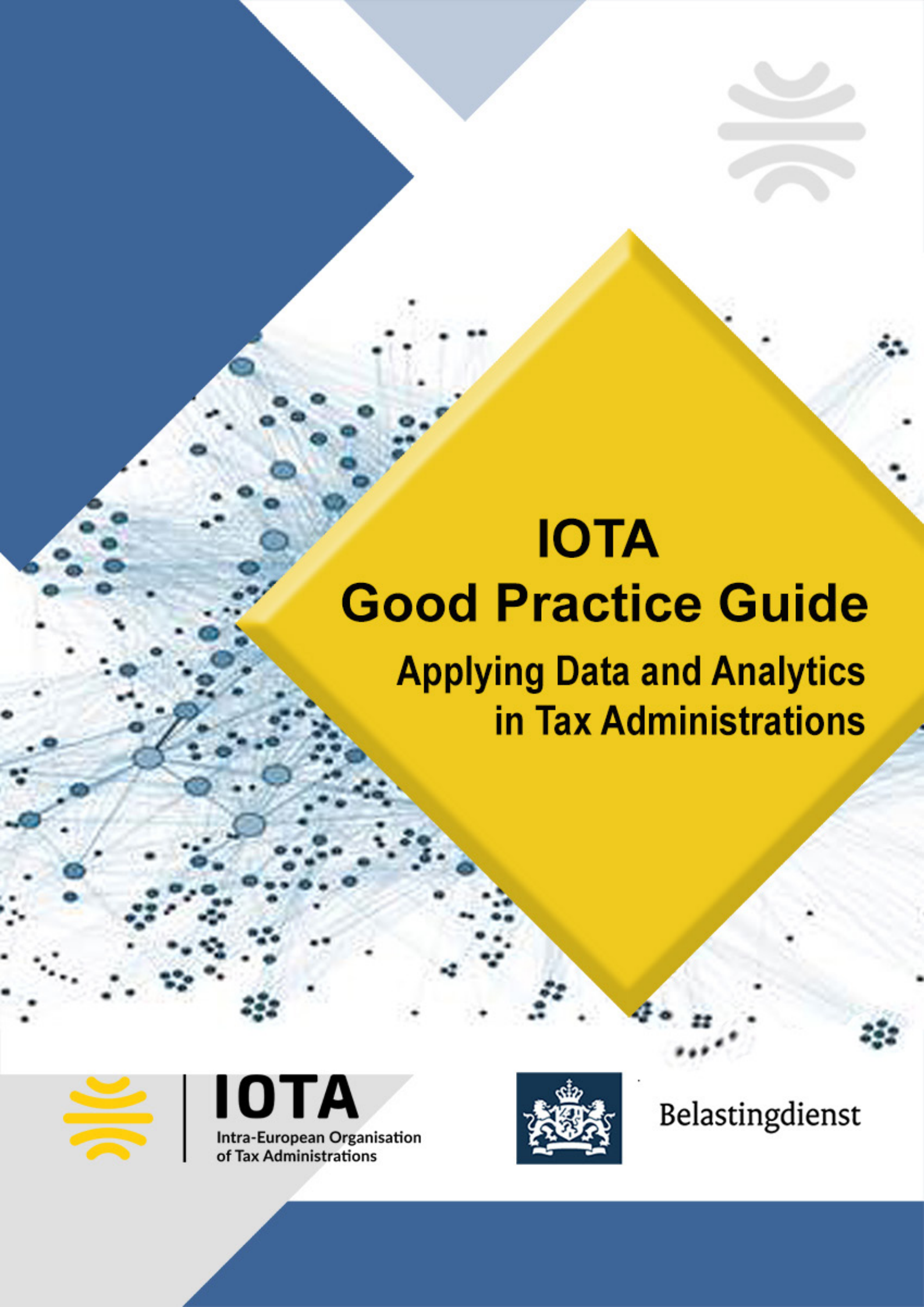


IOTA
Good Practice Guide
Applying Data and Analytics in Tax Administrations

IOTA
Intra-European Organisation of Tax Administrations

Belastingdienst

IOTA Good Practice Guide

Applying Data and Analytics in Tax Administrations

This document is a printable version of IOTA's Good Practice Guide (GPG) about applying data and analytics in tax administrations. The full GPG is published as a series of web pages on the secured IOTA website, which can be accessed with an approved IOTA account via the URL below.

https://www.iota-tax.org/applying-data-and-analytics-tax-administrations

In this document, summaries of the online content are presented, built up along the following chapters.

0. Introduction

The evolution of big data is dramatically reshaping the way businesses operate. Technologies emerge rapidly and the world expects tax administrations to join the revolution. In the Good Practice Guide, a wide variety of good practices in applying data and analytics within tax administrations is collected. Along the edges of five different topics, more than ten European countries have shared their valuable insights.

These insights about data and analytics were collected through a questionnaire (which can be downloaded on the GPG website). Roughly 30 countries took the opportunity to submit their answers regarding this project. Many of these novice and advanced countries in data and analytics expressed their interest in being further involved in this project and to actively share experiences, developments and future possibilities of data and analytics in tax administrations. Therefore, a task force has been formed and supported the Project Team for the valuable continuation of the project.
Also, the online IOTA GPG discussion platform, and the IOTA workshop 'Analysis and Efficient Use of Big Data - A Challenge for Tax Administrations' in Utrecht, the Netherlands in October 2016 provided useful input the project. From this input, five topics were selected that appeared to be most occurring and challenging. These topics are data management, predictive modelling, social network analysis, data visualisation and random audits. Additional examples from different areas within data and analytics can be found under 'Other topics'.

Since data and analytics is quite a technical subject, we strived to gather not only advanced examples and insights, but also good practices that help novice countries to develop their data and analytics capabilities. With each given example under the topics, the level of difficulty is displayed (basic, intermediate or advanced).

All examples on this website were assembled by a task force consisting of data experts from different IOTA member countries. Therefore, IOTA and the NTCA are very grateful to the following people for their valuable contributions to the final result.

Country/organisation	Name	Title
Hungary	Ádám Tibor Fajkusz	Task force member
Norway	Thomas Myhrvold-Hanssen	Task force member
Switzerland	Jean-Luc Wichoud	Task force member
The Netherlands	Shidrukh Khazen	Task force member
The Netherlands	Sander Lubbers	Interaction designer
The Netherlands	Roel Niessen	Project leader
IOTA	Mostafa Amini	Project coordinator
IOTA	Péter Póth	Web master

1. Data management

Many tax administrations experience data management as their biggest challenge, as they try to transform into a data-driven organization. Having access to the right data will enable organizations to draw the right conclusions and build valuable solutions. But what is data management? Asking this question to ten different people will give ten different answers. Data management includes topics like data preparation, data quality, data governance, data storage, data access and data security. One can easily drown in the numerous frameworks and approaches for managing data. Therefore, this chapter only focuses on some specific topics around data management, which altogether are far from exhaustive for data management as a whole. Nevertheless, the given insights are worth sharing and discussing.

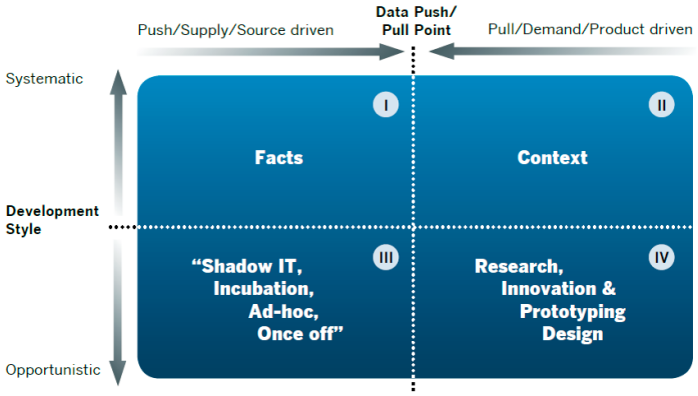
Data quadrant model (by Ronald Damhof)

Level: basic

Contact person: Bastiaan Veldkamp, The Netherlands

Problem:

A common pitfall in many organisations nowadays is that a lot of focus is put on creating innovative information products, and less on requirements for putting these in production. The strong focus on innovative information products is caused by a lack of insight of the different kinds of data preparation that exist and their different requirements.



Solution:

The Netherlands Tax and Customs Administration (NTCA) however strives to optimise how ad-hoc models and data that are created by their data scientists find their way into the structured production process. In an interview with Ronald Damhof we will show how his Data Quadrant Model helped the NTCA to organise its thinking around different types of data preparation. It starts with a basic assumption that data deployment starts with raw materials and ends up in some sort of product. And in the process of getting raw materials to end products, logistics and manufacturing is required. It also starts with the basic assumption that reliability and flexibility are both wanted, but are mutual exclusive. And lastly, it starts with the basic notion of the 'push pull point', which stems from the logistic- and manufacturing literature we were taught in high school about push systems and pull systems.

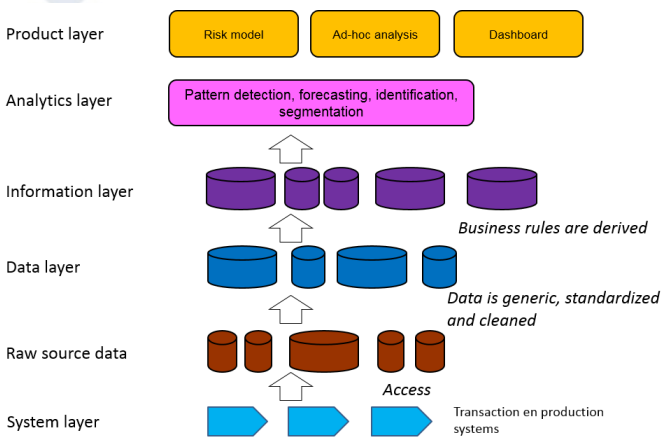
Data layer architecture

Level: basic

Contact person: Bastiaan Veldkamp, The Netherlands

Problem:

A common pitfall in many organisations nowadays is that a lot of focus is put on creating innovative information products, and less on requirements for putting these in production. The strong focus on innovative information products is also caused by a lack of insight on how to create a good method of data preparation.



Solution:

The second model that has been developed by the Netherlands Tax and Customs Administration's own data and analytics department is its own layered data architecture approach that has been instrumental in their success. In this method, the data is prepared in layers. In each layer, we build on the previous layer and we give each layer a colour for functional purposes. In the first layer that we build, our "blue layer", we standardize, cleanse and validate the data. In this layer, we also apply technical rules. In the next layer, our "purple layer", we build on the blue layer and enrich our data by creating business rules.

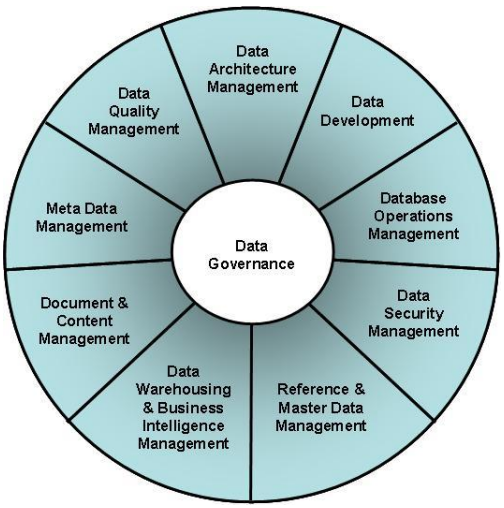
This method of data preparation helped the organisation to focus on creating insights that add value, to have a strong focus on data quality ("blue") and business validation ("purple"), a clear data lineage and a common language and shared approach.

The DAMA framework

Level: basic
Contact person: Bastiaan Veldkamp, The Netherlands

Problem:

The Netherlands Tax and Customs Administration (NTCA) has had quite some challenges when it comes to assessing their data management function.



Solution:

Therefore, the DAMA-DMBOK framework has helped NTCA to address this problem. This is a widely-accepted framework for data management that standardises terminology, processes, roles and deliverables. It covers a full range of topics including data development, data architecture, data quality, data security, master data, and meta data management. This model helped our organization to assess our Data Management function and to suggest and guide initiatives to improve data management.

Standard audit file for tax (SAF-T)

Level: intermediate
Contact person: Ricardo Santos Vítor, Portugal

Problem:

Identifying risks or quantifying possible errors in large sets of accounting data is a hard job for a tax auditor, especially when every business or accounting software uses its personal filing method.

Solution:

Standard Audit File for Tax (SAF-T) is an international standard for electronic exchange of reliable accounting data from organizations to a national tax authority or external auditors. For the Portuguese tax administration, adopting SAF-T was the starting point to gather big volumes of information about invoicing, transport documents, etc. By analysing these new data, the Portuguese were able to more successfully detect and prevent tax fraud.

2. Predictive modelling

Many tax administration think that predictive modelling is complicated and requires a high level of expertise to start with. A good practice it is to start small and easy and to grow iteratively.

Predictive modelling is an optimization process, data science and big data offers new opportunity to improve the efficiency of risk modelling.

Predictive modelling is the traditional area application of advanced analytics and big data in tax administration, because it is a great tool to increase revenue, detect fraud and support reorganization.

Predictive modelling is like cooking

You need some ingredients before starting with predictive modelling. These ingredients are your risk indicators. It could be raw material like business rules or prepared ingredients like the output of a clustering model or results of a network analysis.

With that, the state of the art is to create a tasty recipe, much better is to improve the procedure by iteration, like a chief who test a new dish and improve his recipe through the time. Mixing unusual ingredients needs some technics and is risky, and could create unexpected results. Of course, new advanced analytics and machine learning methods offer a great opportunity to find the ultimate recipe by optimizing the combination of the ingredients.

Risk indicators – Business rules

Level: basic
Contact person: Jean-Luc Wichoud, Switzerland

Problem:

How to capture good ideas and apply it in a systematic way in a risk model.

Solution:

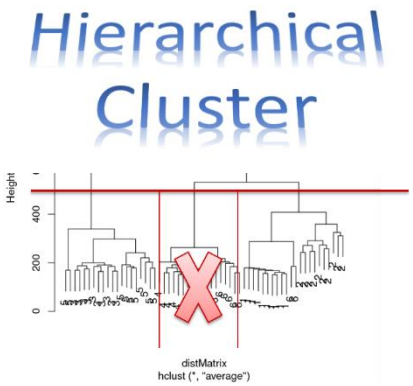
Using rules defined by the business as risk indicators is the old fashion and easiest approach, used for more than 15 years in tax administrations. It is one of the best ways to capture business knowledge in a risk model. In the attachment on the GPG website, you find an example from Switzerland, showing the integration and the evaluation of business ideas as new risk indicators and the challenge to find efficient rules fitting with the aims of the tax administration.

Risk indicators – Hierarchical clustering on declaration data

Level: intermediate
Contact person: Jean-Luc Wichoud, Switzerland

Problem:

We always try to find indicator covering a large variety of taxpayer for our predictive model to tend to cover all the risks and find other ways to enrich it.



Solution:

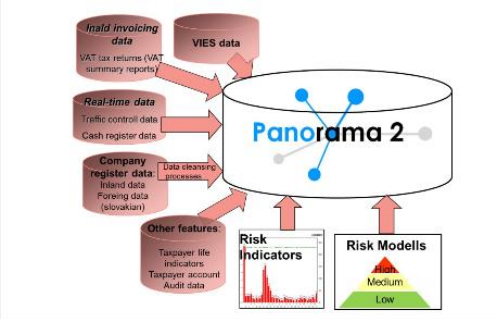
A risk indicator could be more complex, it could be an output of an advanced analytics model. In the attachment on the GPG website, you find an example from Switzerland, where a risk indicator is calculated using an unsupervised method (hierarchical cluster). It gives a more advanced example of a risk indicator used as a component to select taxpayers at risk.

Risk indicators – Network analysis with Panorama

Level: advanced
Contact person: Ádám Fajkusz, Hungary

Problem:

Quite often and in particular in the case of VAT-fraud, risk is linked to a network of taxpayers. How could we extract information from a network to define a risk indicator?



Solution:

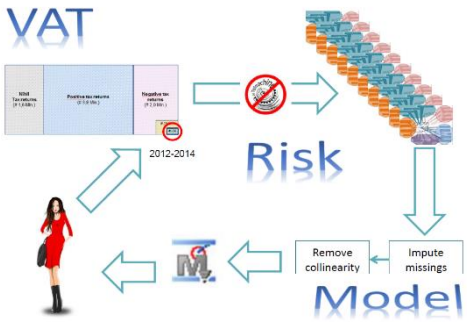
Other advanced techniques could be used for predictive modelling. For instance, network analysis is very useful to detect fraudulent organizations. The Hungarian Tax Administration uses the tool Panorama to create networks that help detecting VAT fraud.

Combine and optimize – VAT risk model

Level: intermediate
Contact person: Lisanne van der Breggen, The Netherlands

Problem:

How to build a predictive model to integrate risk indicators and optimize the selection of VAT-declaration with refund for check?

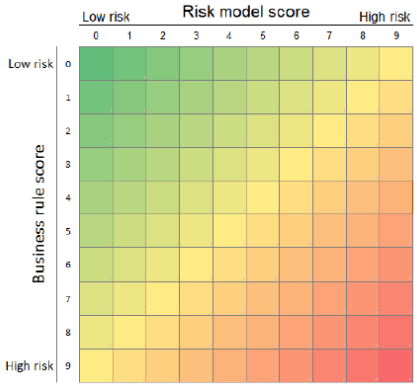


Solution:

The predictive model presented is focused on VAT declaration with refund, it is a risky area in many countries. The Netherlands has implemented an iterative approach to create the model. Risk indicators are the input, a training sample is extracted, the model is optimized on the sample and compared with another one, so iteratively the model is improved. The advantage of this iterative approach is a real optimization of the output of your model.

Combine and optimize – Risk matrix

Level: advanced
Contact person: Roel Niessen, The Netherlands



Problem:

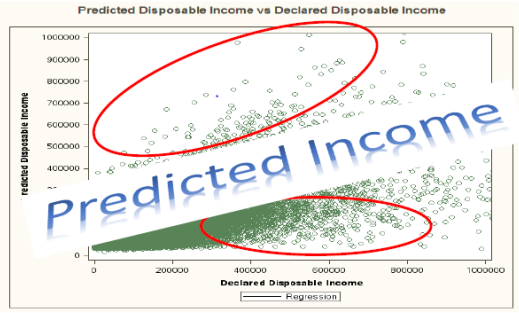
How to combine and optimize different types of risk indicators, in particular with risk indicators of different kinds?

Solution:

In the attachment on the GPG website, the approach applies for income tax by the NTCA. The aim is to combine a rule based model with a pattern based model by combing ranking of each model in a matrix. The goal is to capture the best of the both models.

Combine and optimize – Predicting income using Big Data and regression

Level: advanced
Contact person: Jacqueline Kirrane, Ireland



Problem:

How to combine and optimize Big Data/3rd party information with your predictive model to have a 360° view of the taxpayer.

Solution:

The Irish revenue has a model using regression to estimate the income based on Big Data and 3rd party information. The challenge is to combine and optimize to obtain a good predicted income. It is a smart way to use Big Data integrated by regression in a predictive model. It is a great challenge to manager the data quality with a vast number of information sources.

3. Social network analysis

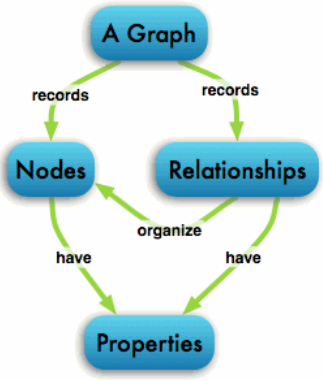
Social network analysis (SNA) is the visualising and studying of relations between people, organizations, IP addresses, and other connected entities. These entities are displayed as nodes in a network, where relations between these entities are represented by lines between the nodes. SNA provides both a visual and a mathematical analysis of relationships. As an example, social network analysis can be useful for detecting international, fraudulent networks or preventing these to occur.

Basic good practices in SNA

Level: basic
Contact person: Rachel O'Carroll, Ireland

Problem:

Social network analysis is that about social media like Twitter and Facebook? No, not in this context at least. But then what is social network analysis and how could it be useful for tax administrations?



Solution:

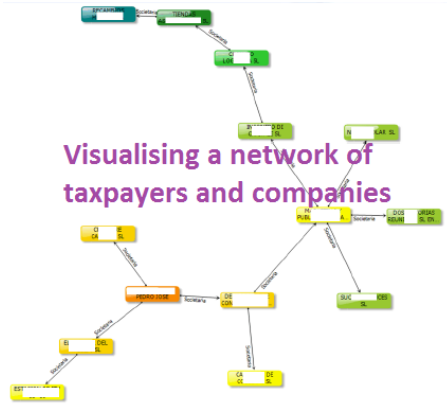
This presentation from the Irish Tax Administration shows what SNA actually is. They talk about the possible benefits of SNA, useful tools for visualising networks (like graph databases), relevant data sources for creating useful social networks, and how SNA could specifically be of good use for tax administrations.

Social network analysis for fraud detection

Level: intermediate
Contact person: Carlos Andres, Spain

Problem:

Traditional, case oriented analytical techniques do not allow tax employees to audit large networks of companies and people. Or at least the traditional techniques would take a long time for one particular investigation to pay off. Also, detecting fraudulent behaviour in a large international network of companies and people can be rather complex.



Solution:

By building and visualising social networks using open source tools, the Spanish Tax Administration can now easily derive information for any network, no matter how large.

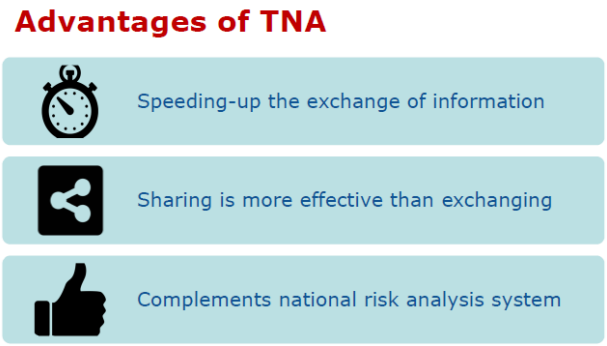
Transaction Network Analysis in the EU

Level: advanced

Contact person: Cristian Largeanu, European Commission

Problem:

Cross-border 'carousel fraud' is estimated to cost EU member states around €50 billion in lost VAT revenues.



Solution:

Since 2013 the European Commission has been working on applying transaction network analysis to detect this kind of tax fraud. By applying automated data mining of data provided by companies in their VAT returns, networks are detected where fraudulent (trading) transactions take place. The EC have conducted pilots, market scanning and working groups to discover the best practices for applying this technique. Not only the techniques used for social network analysis, but also the way different countries collaborate in such a project are interesting to learn from as an individual tax administration.

4. Data visualisation

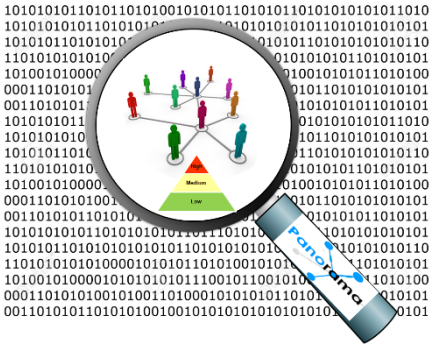
Efficient analysis of large data sets is done more and more by tax administrations. However, experiences show that finding the proper way to communicate results is just as important as the analysis of data. Presenting the data in a graphical format is challenging. An effective way to deal with these challenges would be to use different data visualisation tools. Therefore, tax administrations are now using several tools depending on their data complexity or target audience. These tools can be applied to for example identify trends or for finding correlations among the data. In this chapter, you can find some examples of data visualization that are currently used by tax administrations.

Visualising complex taxpayer relationships and risk management results at the same time

Level: basic

Contact person: Ádám Fajkusz, Hungary

Problem:



Processing, analysing and communicating data from a variety of data sources has been a reasonable challenge for tax administrations. Therefore, it was necessary to develop a system that enables a detailed, transparent and meaningful representation of the complex relationships of taxpayers on the one hand and the results of the tax and customs risk assessment on the other.

Solution:

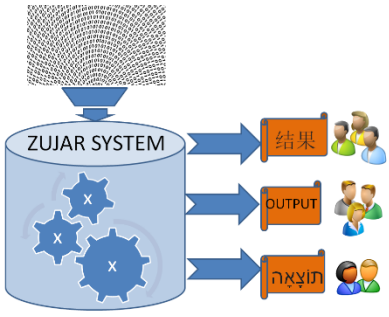
The development of the so-called Panorama visualization tool gave satisfactory answers to many of these challenges. On the one hand, it is designed to show different types of invoicing and hierarchy relations revealed by analysing taxpayer networks. On the other hand, it is also designed to visualise the rule based risk assessment results, offering useful information for various departments within the tax and customs administration.

Multifunctional system with data visualisation functions

Level: basic

Contact person: Javier López Fuertes, Spain

Problem:



Each department uses the same data sets for different purposes. For instance, audit data could be used by Human Resources Department to check employees working time. The risk management department could use the same data to evaluate their risk indicators and the management department can make statistical reports from this data to measure the effectiveness of audit activities. Creating separate applications that serves all these aims could is more expensive to maintain.

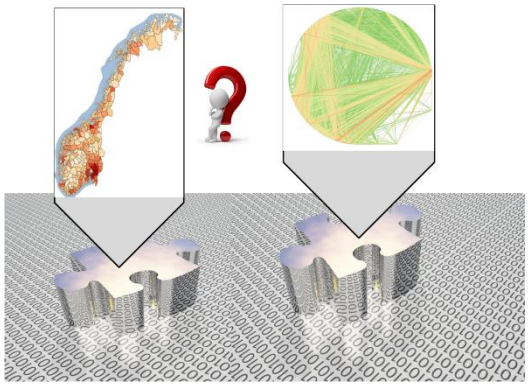
Solution:

In order to get the proper information promptly from the same data to the different departments, the Spanish Tax administration has developed a multifunctional system, called 'Zujar'. This tool enables several visualised reports for different purposes to be generated and forwarded in a much quicker way.

The choices of visualisation

Level: intermediate
Contact person: Nils Gaute Voll, Norway

Problem:



Visualization is often sold as a critical component whenever big data is discussed. The typical motivating example is often Anscombe's Quartet which shows how common statistical measures fail to identify differences between data sets. For this purpose, we discuss a case using maps to illustrate that a very simple low dimensional visualization might force us to make choices with no right answer. Following this example, we try to visualize a more complex case of clickstream data noting that it quickly becomes quite messy and that we need to make a lot of choices to make sense of it.

Solution:

The attached presentation on the GPG website concludes by observing that visualization is not a trivial matter and that when we visualize big datasets, seemingly trivial choices impact how we interpret the data. A highly skeptical attitude towards visualization of big data is thus recommended.

5. Random audits

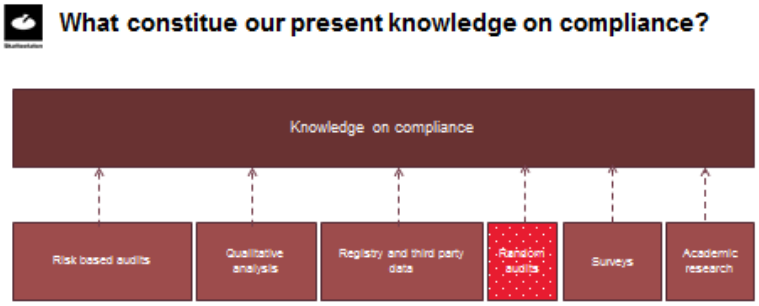
What happens if all our knowledge on tax payer compliance is skewed? The immediate answer is: we would be unable to identify potential risk areas. To mitigate the skewed but nevertheless in depth knowledge we gain from risk-based audits, a sample of randomized audits may prove very helpful in cases where knowledge on compliance cannot be achieved through other methods. The benefits from randomized audits typically evolve in the long run. Therefore, a random audit program calls for a justification beyond a short term, annual return. In this chapter, you find information on the concept of random samples, the pros and cons of randomized audits, and input on how to design a program.

The random audits program in the Norwegian Tax Administration

Level: basic
Contact person: Thomas Myhrvold-Hanssen, Norway

Problem:

How to design a random audit program?



Solution:

The program is meant to help the tax administration in building and structuring knowledge on the areas where we currently cannot assess through risk-based audits alone. The Norwegian Tax Administration has some prior experiences with random audits, however these audits have been targeted towards specific areas and questions, so far not been conducted in a systematic way.

In order to create a wide base for deciding whether a program for the systematic use of random audits should be implemented, and how this program best can be designed in order to help answer the existing knowledge requirements, we are currently undertaking a review of the experience of other tax administrations.

The importance of random samples

Level: intermediate
Contact person: Emma Gottesman, The Netherlands

Problem:

Our audits and compliance measures don't capture the whole scope of risk areas.

Solution:

Benefits of the random sample usually do not appear immediately, but take some time to materialize. They are of vital importance in the long run. The online version of this Good Practice Guide contains a paper that describes the importance of random samples for the tax administration in general. It also describes the way random samples are used in the NTCA.

6. Other topics

Aside from the five topics above, there are many more areas in tax administrations where analysis techniques can be applied. Also, the range of analytical models and techniques to detect non-payment or to gain more knowledge about taxpayer's needs and behaviour, is much wider than these five topics might suggest. In this chapter, a few more strategies and techniques around data and analytics are described. Even though these techniques might not (yet) be widely used, they could deliver a strong added value for your administration.

Channel strategy

Contact person: Thomas Myhrvold-Hanssen, Norway

Channel strategy is the creating of a plan to effectively reach your customers with your products and services. By applying channel strategies tax administrations aim to improve compliance and provide services that are accessible and understandable for everybody. Tax administrations consider their tax administrative requirements, clients preferences, channel characteristics and constraints such as costs and capabilities to be important factors for determining and finding the optimal channel mix.

Country example: Norway

The Norwegian Tax Administration has analysed the use of different channels to get more knowledge of taxpayer's needs. The information is used to build a new and better database of inquiries in order to improve their communication with tax payers. Their aim is to develop more professional cost efficient channels.

Debt management

Contact person: Andreas Voxberg, Sweden

Debt management is the creating of a strategy for debtors to manage their debt. Tax administrations develop these plans to reduce the probability of non-payment, to increase the speed of collection and to reduce costs. Aspects of this strategy can include a better insight in the behaviour of taxpayers, developing better ways to contact customers and improving the quality of taxpayer information on debt.

Country example: Sweden

The Swedish Tax Administration has developed one of their first models to target the risk of non-payment and accumulation of debts within the coming six months. What they need is to get cases to handle to prevent debts and further accumulation of debts by using different tools available to the STA. The model was developed in SAS Enterprise Guide and has been migrated to ODM. The model is a pure mining model with a fairly high complexity. Segmentation is done running different models on different segments.

Web scraping

Contact person: Mr David Dávila, Spain

Web scraping is a technology for taking out useful information from websites. Tax administrations could for instance use web scraping to detect in which countries big corporations are active and create invisible profits, or to quickly collect large volumes of data that might be useful for creating predictive models.

Country example: Spain

The Spanish Tax Administration has launched a project to detect economic activity in open networks. Results include the collection on the internet of real estate rental offers and the evaluation of sales in ecommerce virtual shops.

IOTA
Intra-European Organisation of Tax Administrations

Belastingdienst