Data science for tax administration

Pijnenburg, M.G.F.

1 Introduction

1.1 The Historical Link between Taxes and Data

Taxes and data have a close bond for centuries. The linking pin is called 'administration', see the relation below.

$$\text{taxes} \longleftarrow \text{administration} \longleftarrow \text{data}$$

To levy taxes an administration is required, for instance an administration of properties, transactions, or income. These administrations are basically just collections of data.

The link between taxes and data has emerged already at the start of civilization. Figure 1.1 displays an example of a Mesopotamian clay tablet. These tablets show commodities, like jars, cattle, and grain. Scholars now know that these tablets were mainly used for tax and administration purposes [12].

Also in the Roman times, the link between taxes and data is present. Most readers will recall the story about the birth of Jesus Christ, see Figure 1.2. At the time of Jesus' birth, Joseph and Maria had to travel from Nazareth to Bethlehem. Why? A census had been demanded by the Romans. And why did the Romans do all the efforts to organize a census in their great empire? Because of tax purposes. Jesus was born in the time that Augustus was the first emperor of Rome (reigning from 27 BC until his death in AD 14). Augustus changed the tax system such that each province was required to pay a flat poll tax on each adult [106]. For this reason it was important to know the exact number of adults. So also in the Roman times, one of the best examples of data collection, a census, is closely related to taxation.

In later times, data *analysis* emerged. If data has been gathered consistently for some time, it gives us the possibility to analyze the data and try to see patterns and

Figure 1.1: Clay tablet from Uruk in cuneiform writing describing commodities *source: Metropolitan Museum of Art, accession number 1988.433.3*.

learn things about the world around us. Much of human knowledge has been collected in this way.

Governments are among the early adopters of analyzing data to gain insights. An early example of statistical inference is the Trial of the Pyx [112], an annual procedure that is held in the United Kingdom since 1282 AD, see Figure 1.3. The procedure is held by Her Majesty's Treasury, the governmental department whose minister is also responsible for the British tax authority. The goal of the procedure, that still exists today, is to ensure that newly minted coins conform to the required standards, in particular whether the required amount of precious metals are used. The procedure is among the oldest statistical sampling methods.

Another example of the early involvement of governments in data analysis can still be seen in the origins of the word 'Statistics'. The word derives ultimately from the Latin *statisticum collegium* ('council of state') [81]. The German word 'Statistik', first introduced by Gottfried Achenwall in the mid eighteenth century, originally designated the analysis of data about the state. It acquired the meaning of the collection and analysis of data in the early 19th century [111]. According to the online encyclopedia Wikipedia, *Thus, the original principal purpose of Statistik was data to be used by governmental and (often centralized) administrative bodies*.

The link between taxes, administration, and data has stood the test of time as tax administrations are today still large, data processing organizations. Moreover it is inspiring to see that also today, governments, and in particular tax administrations, are heavily interested in the fruits that data science can bring, see section 1.3 for the motivation of the Netherlands Tax and Customs Authority (NTCA) [79] for doing so.

Figure 1.2: Maria and Joseph on their way to Bethlehem. *Unknown Artist*

1.2 Tax Administrations in the Computer Age

A major task of tax administrations is to make it possible for the people to pay taxes and to check whether they do. To fulfill this task, many processes have to be organized, like receiving and processing tax returns. The processes themselves generate data (incoming tax returns for example) or need data (banking data for instance). It is thus logical that an organization that processes so much data can be helped enormously by computer power.

The arrival of computers, at the end of the twentieth century, has revolutionized data processing, also in the field of public administration and taxes. Computers allow for the storage of massive data sets and have replaced extensive paper archives. The new computing power has replaced the manual calculations of taxes and workflow management software has digitized processes within modern tax administrations.

The early applications of computers in tax administrations have been to store data and automate transaction processes. However, in the last decade tax administrations started to realize that data may provide valuable information about the efficiency of processes and interesting insights about the behavior and needs of taxpayers. Information that can be used for instance to cut inefficiencies or increase services.

Tax administrations are in fact natural places to exploit the benefits of data analysis. Tax administrations have access to a lot of relevant data. To give an impression, the Netherlands Tax and Customs Administration has access to, among others,

- Information about incomes,
- Financial information about companies like profits, costs, and turnover,

Figure 1.3: Trial of the Pix. By Matt Brown - https://www.flickr.com/photos/londonmatt/3269860504, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=52049921.

- Real estate information,
- Municipal data, like addresses, and (family) relations,
- Chamber of Commerce data,
- Counter-information from abroad,
- Some bank details.

See the information brochure [13] for a complete overview. Of course, this data can only be used for data analysis if there exists a legal basis.

Tax administration are also suitable places for data analysis, since a tax administration can make the required investments. Moreover, a tax authority may benefit from even a small increment in efficiency due to the vast sums of money that are processed. Besides, most tax administrations have a large, in-house ICT department, facilitating access to the required technical facilities. Tax administrations also fit with the required fact-driven culture, stemming in part from the nature of the business and the presence of many accountants.

1.3 Analytics at the Netherlands Tax and Customs Administration

The Netherlands Tax and Customs Administration (NTCA, or 'Belastingdienst' in Dutch) is the Dutch governmental organization responsible for collecting the main taxes in the Netherlands [79]. The organization falls under the responsibility of the Ministry of Finance. Its main tasks are to make it possible for taxpayers to pay taxes and to check whether this is done. Customs has been part of the NTCA since its founding in 1806 by finance minister Alexander Gogel, see Figure 1.4.

Figure 1.4: Portrait of Alexander Gogel, Minister of Finance, who founded the Netherlands Tax and Customs Administration in 1806. Source: Belasting & Douane Museum

Today, the organization collects about €200 billion annually and employs slightly less than 30,000 employees. It not only collects taxes but also hands out benefits. Benefits are income-dependent allowances from the government in, for example, healthcare costs, childcare costs and housing costs (rent allowance). The NTCA is present in most cities in the Netherlands, see Figure 1.5.

At least two major trends have contributed to the acceleration of analytics at the NTCA, that has started at about 2013. The first trend has its origin in the financial crisis of 2008. At that moment the Dutch government took a number of austerity measures, including major budget cuts for the various departments. These budget cuts were spread out over a five-year period, leading to ever severe budget cuts each year, increasing the need for efficient operations. The second major trend has to do

Figure 1.5: Locations of the NTCA in four major cities: Amsterdam (top left), Groningen (top right), Apeldoorn (bottom left), and Utrecht (bottom right).

with the age structure of the Tax and Customs Administration staff, which means that since 2016 around 2000 employees retire yearly. This urged the top management to invest in promising techniques that could increase effectiveness.

Another long term trend has been the steady increase of the workload for the NTCA without a steady increase in the number of employees, see Figure 1.6. The figure shows that the number of entrepreneurs have tripled in the 27 year period from 1986 to 2013, a consequence of economic growth but also of a policy of stimulating entrepreneurship by the government. Many of these entrepreneurs are self-employed and take more effort to check for the tax administration compared to employees. The figure also highlights the doubling of the number of income tax returns in this period.

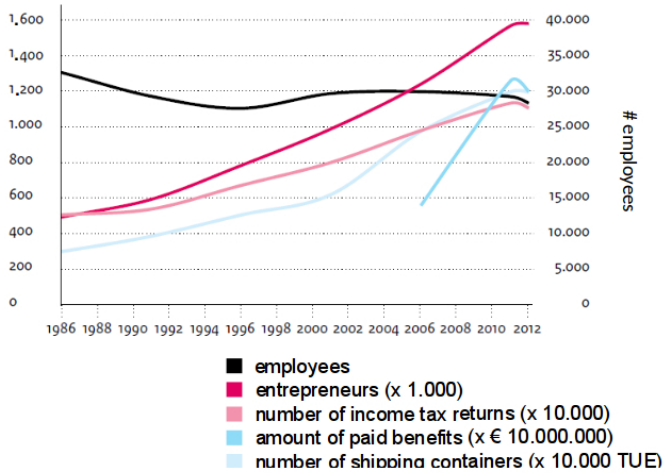

Figure 1.6: Trends related to the workload of the Netherlands Tax and Customs Administration over a 27-year period (1988 - 2013). The colored lines (blue and pink) depict trends that have a major influence on the workload. The scale belonging to the colored lines is depicted on the left vertical axis. The black line represents the number of Full Time Equivalent (FTE) employees. Its scale is depicted on the right vertical axis.

This rise has partly to do with a rise in the general population, but also with the fact that the lifestyle of citizens has become more dynamic and thus more citizens are asked to send in a tax return with detailed information. Also visible in the figure is the fact that since 2006 the NTCA was asked to take over the payment of several benefits from other departments. Finally the figure shows that the number of containers entering the Port of Rotterdam has more than tripled in this period, an indication of growing internationalism, but also of a lot more work for Customs. The black line in Figure 1.6 represents the number of employees in this time period, showing a constant number near 30.000. Considering these numbers, the interest and support of the NTCA for analytics is clear.

1.4 Analytics and Related Concepts

Analyzing data in the computer age comes down to extracting information from the raw data stored in computer systems. The systematic retrieval of data and the transformation of data into information and relevant knowledge is called *data mining*. Although various tax administrations have reached different levels of maturity in data

mining, it is safe to state that currently, no tax administration has fully exploited the information embedded in the data.

Extracting insights from data does not yet create tangible business value. For instance, insights should be 'actionable', i.e. management must be able to change current practices based on the insights. To maximally reap the (business) benefits of data mining, many activities have to be organized beyond the data mining itself, and strategic choices have to be made. Davenport and others [32, 33, 31] have studied data mining in organizations from a business science perspective. They call the application of techniques from statistics and machine learning on data in order to improve business processes *analytics*.

Recently, the vocabulary around data analysis has expanding further with terms as *deep learning*, *big data*, *machine learning*, *advanced analytics*, and *predictive modelling*. Figure 1.7 is an attempt to explain the term 'Analytics' in relation to 'Statistics', 'Machine Learning' and 'Data Science' as an understanding of these relations helps to avoid confusion while reading this thesis.

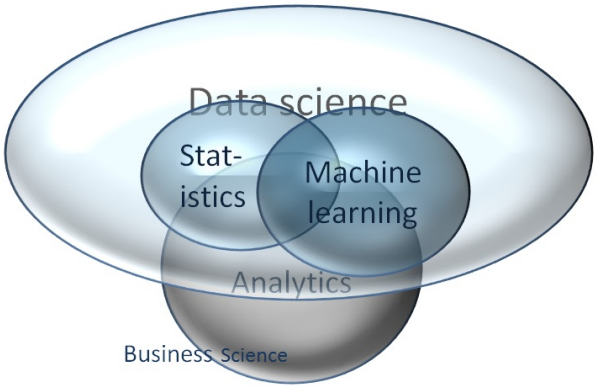

Figure 1.7: Venn-diagram visualization of the concepts *Analytics*, *Data Science*, and *Machine Learning*

The core of the diagram is formed by the components 'Statistics' and 'Machine learning'. These are fields of research that have produced a large collection of techniques that are helpful in analyzing data. Statistics is the oldest discipline of the two and has strong connections with mathematics, especially probability theory. Statistics has introduced many important basic concepts in data analysis. It originated in the pre-computer era, and presumably therefore its foundations and methods are more inclined to rigorous, analytical (i.e. closed form) solutions over heuristic algorithms that come to solutions by exploiting computing power. The connections with mathematics makes that statisticians are likely to think in probabilistic models underlying

a data set.

Machine Learning is a younger scientific discipline and has close connections with Computer Science and (Electrical) Engineering. Machine learning often puts emphasis on challenging practical problems that mimic human intelligence, like image recognition, natural language processing, automatic translation, speech recognition and learning strategies to play games. The connection with computer science makes machine learning in general more 'computation-intensive'. The influence of engineering makes the discipline practical, preferring in general working solutions over theoretical achievements. This said, the boundaries between the two disciplines in modern research are often hard to distinguish.

Around the core in Figure 1.7 the large component called 'Data Science' is shown. Data Science is a broad concept that roughly contains all activities that have to do with the transformation of data to knowledge. It is broader than statistics and machine learning as it does also include techniques to extract data and transform data. In that sense it resembles Data Mining. However it also covers applications of the findings of data analysis and as such has an overlap with many application domains, like geology, medicine, biology, social sciences, and engineering. In fact, in a funny, extreme sense, one could argue that all scientific disciplines are practising data science as long as they infer conclusions from data.

The last component in Figure 1.7, Analytics, was mentioned above. Analytics is understood to be the application of techniques from statistics and machine learning to add value to organizations. Analytics pays attention to embed knowledge from data in existing or new business processes. Analytics states that organizations can achieve a strategic advantage by excelling in data analysis.

1.5 Research Questions

In this section we position the topics of this thesis in the wider field of scientific research. This way it becomes clear where our results fit within the vast body of knowledge already present. To achieve this, we first shortly sketch the research disciplines of the administration of taxes as well as data science. Subsequently, we will formulate the research problem and the related research questions. Finally we will fit the research questions into the framework provided by the Organization of Economic Co-operation and Development (OECD).

Taxation is a large and flourishing research field as is evident from a long list of specialized journals like *Fiscal Studies*, *The National Tax Journal*, *Tax Law Review* and many, many more. Also the more practical side of taxation, namely the administration of taxes, receives considerable attention, see for instance the many articles on the

subject in the *Journal of Public Economics* and the *Journal of Tax Administration*. A scientific conference on the administration of taxes exists as well: the *International Conference on Tax Administration*.

Data Science is a flourishing research field as well. In the field of machine learning (a major sub-discipline of data science, see Figure 1.7), peer-reviewed conferences play a fundamental role. Some well-known, large conferences are the *International Conference on Machine Learning*, the *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* and the *Conference on Neural Information Processing Systems*. Some well-known journals in the field are the *Journal of Machine Learning Research*, *Machine Learning*, *Statistical Modelling*, and others.

At the intersection of the two academic disciplines of the 'administration of taxes' and 'data science', surprisingly little research is available. Currently, there is no scientific journal or conference specialized in this topic. A reason may be that tax data is typically not made available publicly, which makes publication hard as scientific results need to be reproducible. Although some journals publish articles on the application of data science to tax administration (notably *Expert Systems with Applications*, and *Electronic Journal on e-Government* as well as some conferences in machine learning), the application area of data science to tax administration is currently small.

This leads us to the research problem that underlies this thesis:

> **Research problem** *Expand our knowledge of the applications of data science to improve the administration of taxes in terms of effectiveness, efficiency, or improved compliance.*

The research problem is of interest as the potential of data science to the administration of taxes looks promising (see sections 1.2, 1.3, as well as the general news items on the potential of data science for data-rich domains), while many tax administrations are still at the start or intermediate level of implementing data science techniques [40] in their day-to-day business.

The research problem is too broad to answer in a single PhD-track. For this reason we limited the scope of this thesis to the research questions below that are all directly related to the research problem:

Research Questions

1. What data science / analytical techniques can be used in *taxpayer supervision* and what contributions may be expected from these techniques (Chapter 2)?
2. Can we improve on current audit selection models by incorporating categorical variables with many values (like 'industry sector') that are valued by experts doing manual audit selection (Chapter 3)?
3. Can we develop a new anomaly detection algorithm tailored to find the outliers that are of interest to a tax administration (Chapter 4)?
4. Unsupervised anomaly detection algorithms play an important role in (tax) fraud detection. What algorithms can be expected to work well under what conditions (Chapter 6)?
5. Can we develop a data science technique that assists in testing similar treatment of similar (tax) cases (Chapter 5)?

During the PhD track we have come across some smaller topics that are worth noting, but are not directly related to the research questions formulated above. These topics will be covered in Chapter 7.

A framework for the research at the intersection of data science and the administration of taxes, can be found in the OECD report *Advanced Analytics for Better Tax Administration; Putting Data to Work* [40]. The OECD is an established intergovernmental organization that facilitates cooperation between 36 industrialized countries. Its objective is to shape policies that foster prosperity, equality, opportunity and well-being for all. This objective is achieved, among others, by setting standards and exploring new developments.

According to Chapter 2 of the OECD-report *Advanced Analytics for Better Tax Administration*, analytics can be applied within the administration of taxes to the following tasks:

a audit case selection,
b filing and payment compliance,
c debt management,
d taxpayer services,
e policy evaluation,
f taxpayer segmentation.

Some tasks above have a clear link to effectiveness ('audit case selection', 'taxpayer segmentation', 'taxpayer segmentation'), while others are more focused on improving taxpayer compliance ('filing and payment compliance', 'taxpayer services') or efficiency ('debt management').

When considering these tasks in the light of the general distinction of analytical tasks between internal and external processes (more details can be found in Chapter

Task	Links to	Research Question
a. audit case selection	effectiveness	2
b. filing and payment compliance	compliance	5
c. debt management	efficiency	none, but small contribution in section 7.2
d. taxpayer services	compliance	none
e. policy evaluation	effectiveness	1
f. taxpayer segmentation	effectiveness	none, but small contribution in section 7.4

Table 1.1: Relation of research question to OECD tax administration tasks suitable for data science applications

2.1), made by Davenport and Harris [32], it might be said that items 'b' and 'e' belong mainly to internal processes, while items 'a','c', 'd', and 'f' are more directed to external processes of a tax administration.

The relation between the research questions and the six main tasks of tax administrations above, is shown in Table 1.1. The table reveals that most tasks have been touched in this thesis. It also becomes clear that research questions 3 and 4, that deal with fraud detection, do not fit in the table. One could argue that fraud detection is part of audit case selection, or one could add fraud detection as an additional task of tax administrations suitable for improvement by data science.

1.6 Outline of Thesis

As explained in the previous section, four main topics have been studied:

- The role of analytics in taxpayer supervision (Chapter 2),
- Categorical features with many levels in risk models (Chapter 3),
- Anomaly detection (Chapters 4 and 6), with possible applications for fraud detection or data quality,
- Process mining to test the similar treatment of similar cases (Chapter 5).

The first topic, the role of analytics in taxpayer supervision, is at the intersection of computer science and business administration. The chapter addresses the first research question. The interest in the subject stems from two opposite opinions with respect to data science and tax administrations that could be heard some years ago at different tax administrations. One opinion put analytics away as 'another hype'. The

other put analytics on a pedestal and expected great results. By looking more closely at analytics techniques and comparing these with the activities associated with taxpayer supervision, a clearer picture has been created of what one can realistically expect from analytics.

The second topic combines the recently developed technique of Factorization Machines with the classical classification algorithm of logistic regression and thereby answers the second research question. The combination allows logistic regression models to incorporate interactions of categorical features with many levels, without the need for an excessive amount of data. When developing risk models at the NTCA, these categorical variables appeared, but could not be handled with classical techniques.

The third topic contains two contributions to the field of anomaly detection. The first contribution, in Chapter 4, answers research question 3. In this chapter a particular class of anomalies, called *singular anomalies*, are put forward. This type of anomalies are typical encountered when dealing with tax evasion. These anomalies are observations characterized by a single or a few anomalous features, while the other features display regular values. An algorithm is presented for finding these anomalies. The second contribution, in Chapter 6, adds knowledge to the anomaly detection domain by testing three anomaly detection algorithms on the recently published benchmark of Goldstein and Uchida [43]. This chapter answers research question number 4.

The fourth topic (see Chapter 5) answers the last research question. The topic is in the field of process mining, a relatively young branch of data science that takes process logs as input and tries to describe and optimize (business) processes. We contributed a test procedure to verify, based on event logs, whether two groups of process workers treat similar cases similarly. Treating similar cases in a similar way is part of the permanent task of the NTCA.

Besides the four main topics mentioned above, we have touched some others as well. The contributions to these topics are limited, so none of them deserves a chapter of its own. Nevertheless we felt that we had to include these topics, mainly to show the reader the broadness of the field of applications of data science to tax administration. We have collected these topics in a final chapter, Chapter 7. These side-topics are:

- application of analytics within Human Resources,
- reinforcement learning applied to tax debt collection,
- explaining risk models to non-experts, and
- recommendations based on fuzzy sets.

1.7 Publications Related to Thesis

The content of this thesis is based on the following, previously published articles (in chronological order).

- C. Boon, F. D. Belschak, D. N. Den Hartog, and M. Pijnenburg. Perceived human resource management practices. *Journal of Personnel Psychology*, 2014
- M. Pijnenburg and W. Kowalczyk. Applying analytics for improved taxpayer supervision. In *Proceedings of 16th European Conference on e-Government ECEG 2016*, pages 145–153. Academic Conferences and publishing limited, 2016
- M. Pijnenburg, W. Kowalczyk, E. van der Hel-van Dijk, et al. A roadmap for analytics in taxpayer supervision. *Electronic Journal of e-Government*, 15:19–32, 2017
- M. Pijnenburg and W. Kowalczyk. Extending logistic regression models with factorization machines. In *International Symposium on Methodologies for Intelligent Systems*, pages 323–332. Springer, 2017
- M. Pijnenburg and W. Kowalczyk. Singular outliers: Finding common observations with an uncommon feature. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 492–503. Springer, 2018
- M. Pijnenburg and W. Kowalczyk. Are similar cases treated similarly? a comparison between process workers. In *International Conference on Business Information Systems*, pages 1–15. Springer, 2019
- M. Pijnenburg and W. Kowalczyk. Extending an anomaly detection benchmark with auto-encoders, isolation forests, and rbms. In *International Conference on Information and Software Technologies*. Springer, 2019. Best Paper Award

The side-topic of applying reinforcement learning to the collection of taxes has led to the master thesis 'Tax data and reinforcement learning', written by Martijn Post under the author's supervision. The side-topic of explaining risk models to non-experts has led to a non-scientific publication,

- M. Pijnenburg and K. Kuijpers. Explaining risk models to the business. *Tax Tribune*, 35:57–62, 2016.

A full list of publications of the author can be found at page 141 of this thesis.