Value-added tax fraud detection with scalable anomaly detection techniques

Reference:

Vanhoeyveld Jellis, Martens David, Peeters Bruno.- Value-added tax fraud detection w ith scalable anomaly detection techniques
Applied soft computing - ISSN 1568-4946 - 86(2020), 105895
Full text (Publisher's DOI): https://doi.org/10.1016/J.ASOC.2019.105895
To cite this reference: https://hdl.handle.net/10067/1636820151162165141

Value-added tax fraud detection with scalable anomaly detection techniques

Jellis Vanhoeyveld[a,1,2,*], David Martens[a,2], Bruno Peeters[b,1]

[a] *Faculty of Applied Economics, University of Antwerp, Prinsstraat 13, 2000 Antwerp, Belgium*
[b] *Faculty of Law, University of Antwerp, Venusstraat 23, 2000 Antwerp, Belgium*

Abstract

The tax fraud detection domain is characterized by very few labelled data (known fraud/legal cases) that are not representative for the population due to sample selection bias. We use unsupervised anomaly detection (AD) techniques, which are uncommon in tax fraud detection research, to deal with these domain issues. We analyse a unique dataset containing the VAT declarations and client listings of all Belgian VAT numbers pertaining to ten sectors. Our methodology consists in applying AD methods to firms belonging to the same sector and enables an efficient auditing strategy that can be adopted by tax authorities worldwide. The high lifts and hit rates observed in most sectors demonstrate the success of this approach. Sectoral differences exist due to varying market conditions and legal requirements across sectors and we show that the optimal AD method is sector dependent. We focus on three methodological problems that show issues in the related literature. (1) Can we design suitable input features? We develop new fraud indicators from specific fields of the VAT form and client listings and show the predictive value of the combination of these features. (2) Can we design fast algorithms to deal with the large data sizes that can occur in the tax domain? New methods are developed and we demonstrate their scalability both theoretically as well as empirically. (3) How should fraud detection performance be assessed? A new evaluation methodology is proposed that provides reliable performance indications and guarantees that fraud cases are effectively detected by the proposed methods.

Keywords: Unsupervised anomaly detection, Tax fraud detection, Scalable algorithms

1. Introduction

The current fight against tax fraud is confronted with a number of significant challenges [69]. Fraudsters adopt ever growing complex structures and operate in an organized fashion. The globalization, digitalization and differences in tax rates across countries have stimulated fraudsters to set up cross-border fraud schemes.

[*] Corresponding author

Email addresses: vanhoeyveld.jellis@gmail.com or jellis.vanhoeyveld@uantwerpen.be (Jellis Vanhoeyveld), david.martens@uantwerpen.be (David Martens), bruno.peeters@uantwerpen.be (Bruno Peeters)

[1] Member of the Antwerp Tax Academy https://www.uantwerpen.be/en/about-uantwerp/organisational-structure/centres-and-institutes/antwerp-tax-academy/

[2] Member of the Applied Data Mining research group http://applieddatamining.com/cms/

An important technique of tax fraud (or tax evasion) is the intentional misrepresentation of information in tax declarations in order to decrease the amount of tax liability [17]. Tax fraud can broadly be categorized as the evasion of direct and indirect taxes and has a direct impact on society [30]: the unfair redistribution of wealth; tax increases; cuts in the public services and the hampering of economic growth. In Belgium, tax fraud is a severe problem as an estimated €25 billion is lost annually hereby, which corresponds to 6,25% of its Gross Domestic Product (GDP) [59]. This figure rises to a staggering €1 trillion in the European Union (EU) [22]. The European Commission provides a detailed impact assessment regarding the effects of aggressive tax schemes in the EU [23].

The fight against tax fraud is a priority at the political level and therefore a series of measures have been proposed. If we focus on the European level, one could mention the European council directive 2011/16/EU[3] of 15 February 2011 as frequently amended, regarding the administrative cooperation between EU member states on matters of tax [26]. Specifically, the council regulation (EU) No 904/2010[4] of 7 October 2010 deals with cooperation and combating fraud in the area of value-added tax (VAT) [25]. Furthermore, there is an urgent desire to share the knowledge and expertise of tax experts among different member countries [21].

VAT is a form of indirect taxation that is worn by the end user [17]. We illustrate this by means of an example. Say that company C buys goods of supplier S for a base amount of €100 and a VAT amount of €21 (a 21% tax rate). C sells these goods to the end user E for €150 (excluding VAT) + €31,5 (VAT charges). In this example, S should pay €21 to the tax administration. If C is allowed to deduct all taxes[5] on incoming transactions,[6] then C should pay a total of €31,5 − €21 = €10,5 to the tax administration. Note that this amount corresponds exactly to the tax rate (21%) applied to the added value (€150 − €100) that C brings to the product (hence the name VAT). The tax administration receives a total amount of €31,5 that is completely worn by the end user E and is indirectly collected through S and C.

The VAT taxation system, like any tax regime, is prone to fraudulent abuse. There are several ways an entity can exploit the system, but we restrain ourselves to listing a few: exaggerating ones purchases in order to deduct more taxes on incoming transactions; under-declaring[7] outgoing transactions (sales) to limit the amount of tax due; feigning a foreign sale (not liable to inland VAT) by means of a fake invoice to

[3] The Directive was recently amended several times as a result of which its scope was significantly extended to also cover: automatic exchange of financial account information (Council Directive 2014/107/EU of 9 December 2014); exchange of information involving cross-border tax rulings and advanced pricing arrangements (Council Directive (EU) 2015/2376 of 8 December 2015); country-by-country reporting requirements for multinational enterprise groups operating in the EU (Council Directive (EU) 2016/881 of 25 May 2016) and the insurance that tax authorities have access to beneficial ownership information gathered in the context of anti-money laundering (Council Directive (EU) 2016/2258 of 6 December 2016). All consolidated versions can be found on: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32011L0016

[4] This document has been amended as well. For a consolidated version, see: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32010R0904

[5] This situation occurs when an entity is engaged in activities (outgoing transactions) that fall within the VAT-system and all its purchases (incoming transactions) are deployed to realize these activities. VAT on purchases for personal use or for activities that fall outside of the scope of VAT cannot be deduced.

[6] Throughout the text, incoming transactions denote purchases and outgoing transactions correspond to sales.

[7] Under-declaring means reporting a lower value than the actual value.

cover up a domestic transaction (liable to inland VAT) or simply not declaring anything at all.

VAT forms a considerable contribution with respect to the total of tax revenues. The Organisation for Economic Cooperation and Development (OECD) reports the 2014 VAT earnings in Belgium to be €27,5 billion [48] (15,27% of the total tax revenues). At the level of the EU, the 2014 VAT earnings yield a total of €1,03 trillion (18,43% of the total tax revenues) according to Eurostat [27]. If we focus on the VAT-gap, the center for social and economic research (CASE) reports the 2014 Belgian gap to be €2,98 billion (9,77% of expected returns) and the 2014 EU gap to be €160,22 billion (14,09% of expected returns).

The impact and scale of the (VAT) tax evasion problem has triggered tax authorities to lookout for efficient detection methodologies. Governments increasingly rely on data mining methods, which are able to automatically distinguish fraud patterns from legal ones, as key risk management tools that allow them to concentrate their limited resources on the perceived high risk targets without sacrificing any capacity to the compliant entities [21, 69, 24]. Efficient fraud detection is vital as this leads to a recovery of financial losses and enables an enhanced deterrence. We will apply such data mining techniques to detect VAT fraud.

The area of (VAT) tax fraud detection has a number of domain specific challenges that are summarized in Figure 1 and that are further explained in the next paragraphs.

Obtaining labelled data (instances with fraud or compliant indications) is a difficult task because this translates to conducting costly and time consuming audits (investigations) by tax experts. Such an investigation consists of checking the firm's books, scanning the company's computers for irregularities, verifying trade transactions, contacting the clients and suppliers of the audited company, etc. This means that the vast majority of tax data is unlabelled by nature. We refer to Table 1 to illustrate the relatively small sizes of the audited sets (the labelled data) in comparison to the sizes of the populations (all data).

Because conducting audits requires a lot of resources (in terms of cost and time), tax fraud detection environments are typically confronted with a labelled sample[8] that is not representative for the population [46], similar to the reject inference problem in credit scoring [70]: (1) audited data represent a very small fraction of the entire population. Percentages smaller than 0,5% are quite typical (e.g. see the works of Bonchi et al. [9], Gupta & Nagadevara [35], Castellón González & Velásquez [13] indicated in Table 1). Furthermore, (2) there is a sample selection bias [5, 51]: auditors typically don't want to waste time on tax compliant entities and rely on their past experience to present similar neighbouring fraud cases. This means the labelled sample is a biased representation of the population. Also, as a direct consequence, (3) the proportion of fraudsters in the audited set is not in accordance with the expected proportion on the overall population (fraud is assumed to be a rare phenomenon). In the study of Basta et al. [5], the proportion is even reversed: the audited set consists of 85,29% of instances that pertain to the fraud class.

The tax domain is dynamic in nature and this aspect is commonly referred to as concept drift [31, 51],

[8] The labelled sample corresponds to the set of entities that have formed the subject of a tax audit in the past (the audited set). Hence we know whether or not they are tax compliant.

where the fraud/legal class distributions change over time. Indeed, some fraudsters evolve their modus operandi to approximate legal forms and/or to counter detection systems [49]. Also, the legal class distribution is time dependent due to changing legal requirements (modifications in tax law) and macro-economic factors (e.g. changing market environments). Hence the tax declaration behaviour of companies changes over time and this in turn requires a regular update of the fraud detection models.

The notion of fraudulent or legitimate behaviour of a company may differ depending on the specific sector where the entity is active. Different tax regimes and conducts are observed in different sectors because of non-identical market conditions and varying juridical requirements. For instance, a company active in the construction sector has a different way (behaviour) of filling in a tax declaration form than a firm operating in the catering sector.

In the area of taxation, because fraud can be penalized several years post occurrence, it is common to aggregate the tax declarations of a single entity across a time interval of interest and retrieve the most likely fraudsters within the population. This means the tax fraud detection domain can be confronted with large datasets that are processed in a batch mode (off-line mode). In a case study of VAT fraud detection in a small country as Chile, Basta et al. [5] mentions a total of 582.161 active enterprises in the period 2005-2007. For larger countries, the data sizes can become even larger (e.g. the number of listed domestic companies in Canada, USA, India and China exceeds 3 million [64]. In India, this number is around 5,5 million).

Tax administrations are requested to improve their efficiency and accountability while reducing budgets and staff requirements [24]. This means there are severe resource (capacity) constraints on their behalf, which implies that only a small fraction of the entities eligible for auditing can be effectively audited.

The (VAT) tax fraud detection domain's characteristics, as highlighted in the previous paragraphs, differ from those of other fraud areas. To illustrate, we present a number of specificities of credit card fraud, which is a commonly investigated fraud area [49, 47, 71]. Credit card fraud operates in an on-line environment [66], which means that transactions emerge in a streaming fashion. This implies that models should be developed (trained) on past transactions to predict the class label of new transactions [52]. With respect to the availability of labelled data, we note that fraudulent transactions that have not been flagged by the detection system will be reported a few weeks post-occurrence by customers that discover fraud on their account. Also, undisputed transactions can be considered genuine after a certain period of time. Hence the labels for all transactions become available after a certain time lag, a problem known as verification latency [51, 40]. Furthermore, the output of credit card fraud detection models seems to require a hard labelling decision (a binary indication of fraud or legal). These fundamentally distinct characteristics in comparison to tax fraud translate to different requirements in the modelling effort and also imply that the obtained results/conclusions do not generalize beyond the fraud domain under investigation.

The main goal of this paper is to develop fraud detection models that take into account the aforementioned VAT domain's characteristics. For reasons indicated in Section 2, we propose the use of unsupervised anomaly detection (AD) methods to deal with many of these domain problems. Our contributions, stated in Section 3, are geared towards solving a number of methodological problems that are raised by the VAT domain's characteristics and the use of AD methods: (1) Can suitable input variables be designed for the specific domain of VAT fraud detection? (2) Can we develop fast AD methods that are able to deal with the large data sizes encountered in the VAT domain? (3) How should one assess fraud detection performance? The literature review presented in Section 2 reveals a number of shortcomings related to each of these issues. Figure 1 presents a high-level overview of this paper. We explain each of the figure's components in Sections 2 and 3. Our target audience includes practitioners or academics in data mining based fraud detection as well as tax experts (academics, government officials, companies offering tax services).

The specific problem statement that we address and that is encountered by tax administrations worldwide can be articulated as follows: given a set of periodic tax declarations, within a time interval of interest, made by entities pertaining to a certain population (e.g. all companies active in a sector of interest). The problem is which of these entities should be selected for auditing so that as many fraud cases as possible can be detected. Indeed, capacity constraints imply that only a limited number of entities can be investigated and a sensible selection thereof is mandatory to enable an efficient auditing strategy.

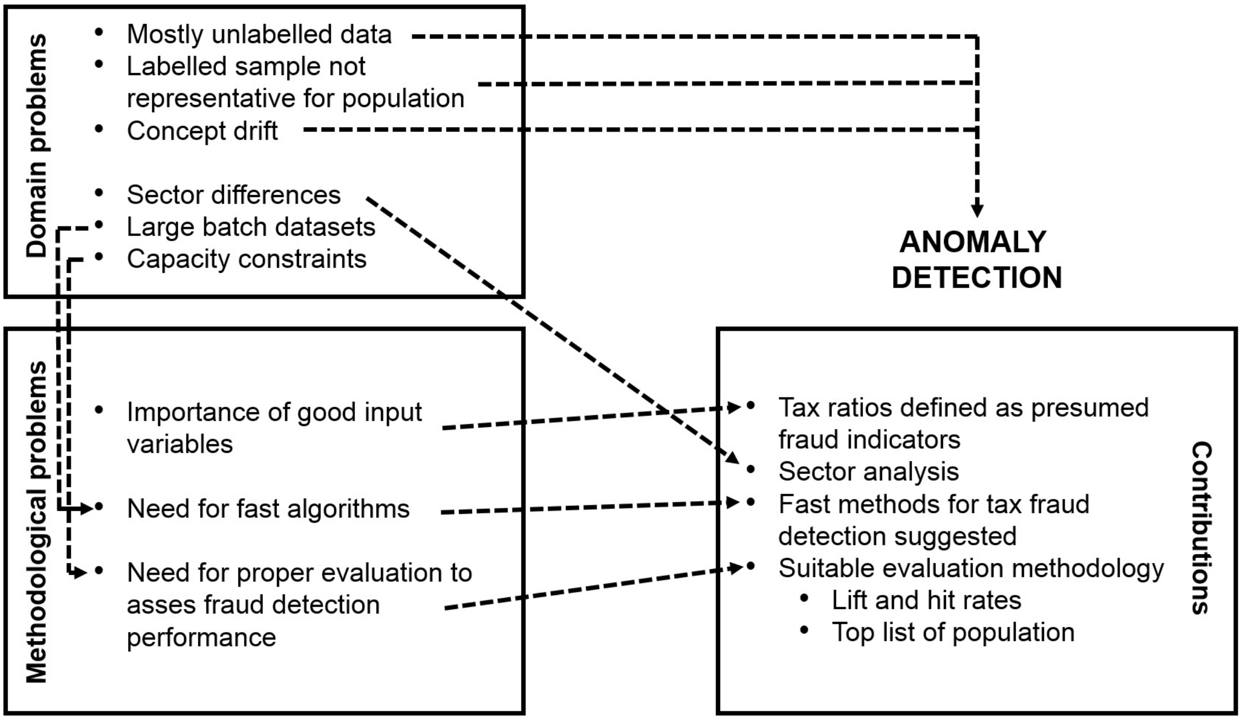
Domain problems
Mostly unlabelled data
Labelled sample not representative for population
Concept drift
Sector differences
Large batch datasets
Capacity constraints
ANOMALY DETECTION
Methodological problems
Importance of good input variables
Need for fast algorithms
Need for proper evaluation to asses fraud detection performance
Contributions
Tax ratios defined as presumed fraud indicators
Sector analysis
Fast methods for tax fraud detection suggested
Suitable evaluation methodology
Lift and hit rates
Top list of population

Figure 1: High-level overview of the paper.

2. Related work

2.1. Literature review on VAT fraud detection

The literature on data mining based VAT fraud detection can be categorized according to the type of AD techniques used: unsupervised or (semi-)supervised. This distinction is based on the extent to which labels (fraud or compliant) are available [15, 33]: supervised techniques construct their detection models based on a fully labelled dataset. They correspond to the traditional classification based techniques that learn to discriminate between the two classes. Regarding the semi-supervised approaches, we adopt the definition that is common in the AD literature: they develop a decision boundary around the instances of a single class and learn to recognize observations pertaining to this class. Any instance that deviates from the model is assigned to the other class. Unsupervised AD techniques, which constitute the focus of this paper, are the most flexible as they do not require labelled data to construct their models and can directly be applied to the entire population. In the remainder of this paper, when we allude to AD, we implicitly refer to its unsupervised variant. We will introduce the concept of AD and its related literature in Section 2.2.

Table 1 presents a detailed literature review. An indication regarding the type of detection mechanism is provided. Population or audit sizes are aggregated in case multiple years are analysed separately by the authors. It is clear that AD techniques have only been applied to a single published paper. Castellón González & Velásquez [13] give an overview of the data mining methods adopted by tax administrations all over the world.[9] None of them seem to make use of AD algorithms. This trend seems to be confirmed in the broader areas of tax fraud detection [58] and financial fraud detection.[10] Ngai et al. [47] reviewed the financial fraud detection literature and noted AD to be the least common, accounting for only 2% of all articles under survey. A more recent 2016 review [71] does not even mention the use of AD techniques.

The popular supervised classification approaches have proven their usefulness in several fraud detection applications. However, the specific VAT domain characteristics introduced in Section 1 pose problems related to the application thereof as we explain below. As mostly unlabelled data are observed, it is difficult to construct supervised models [58]. Also, the labelled sample is not representative for the population and suffers from sample selection bias, which means that the resulting classifier constructed on such data is also biased. Furthermore, because auditing is very time consuming, the labelled information may already be outdated due to concept drift. The dynamic nature of the tax environment may render auditing solely based on supervised scoring mechanisms inappropriate on the long run because classifiers are limited to flagging known fraud types (at classification time) [8, 20]. This means that new modus operandi of fraudsters may not be detected. Furthermore, cross-sector differences motivate the development of a classification model

[9] USA, Canada, Australia, UK, Bulgaria, Brazil, Peru and Chile.

[10] Financial fraud encompasses tax fraud and can be defined as the intentional use of illegal methods/practices to obtain financial gains [71] and includes such areas as bank fraud (credit card fraud, money laundering and mortgage fraud), insurance fraud, corporate fraud (securities and commodities fraud, financial statement fraud) and mass marketing fraud.

Table 1: Literature overview on data mining based VAT fraud detection. An indication of the type of detection mechanism is provided: supervised (Sup), semi-supervised (Semi) or unsupervised (Unsup).

Ref./Pop./Sample	Techn.	Description
Bonchi et al. [9] (1999, Italy) Population: 80.643 instances Audited set: 4.103 entities	Sup	Cost-sensitive decision trees to detect tax evasion. (Presumably, multiple tax fraud types are considered).
Gupta & Nagadevara [35] (2007, India - Delhi) Population: 180.000 instances Audited set: 1.608 entities	Sup	A variety of classification trees, logistic regression and discriminant function analysis were used to tackle VAT fraud.
Basta et al. [5] (2009, Italy) Population: 34 million VAT declarations Audited set: 45.442 entities	Sup	An iteratively refined rule-based system to detect VAT fraud and incorporate multiple objectives.
Wu et al. [74] (2012, Taiwan) Population: ? Audited set: 3.477 fraudsters	Semi	Learning association rules from business entities identified as likely involved in VAT evasion activities.
Castellón González & Velásquez [13] (2013, Chile) Population: 582.161 instances Audited set: 1.692 entities	Sup	Self-organizing map (SOM) for clustering/visualization of the universe of tax payers. Subsequently, neural networks, decision trees and Bayesian networks are applied.
Matos et al. [44] (2015, Brazil - Ceará) Population: 142.000 instances Audited set: 23.810 entities	Semi	Dimensionality reduction techniques (PCA/SVD) were applied on the non-compliant entities to reduce the 14 fraud indicators to a single dimension that serves as a fraud score.
Assylbekov et al. [4] (2016, Kazakhstan) Population: 116.494 instances Audited set: approx. 1.000 entities	Unsup	Statistical AD based on a multivariate Gaussian distribution. Its parameters are estimated from the largest cluster after a two-level SOM based clustering process on the population.
Mittal et al. [46] (2018, India - Delhi) Population: 315.191 firms Audited set: 538 fraud cases	Sup	Random forests are applied to detect bogus firms (bill traders or shell companies). Model construction and evaluation based on the entire population by relying on a special kind of cross-validation procedure.
Mehta et al. [45] (2019, India - Telangana) Population: 1.233 instances Audited set: not of relevance	Unsup	Clustering dealers represented by 4 features. Descriptive analysis of feature values within the obtained clusters. No assessment of fraud detection performance is conducted.

per sector. This would further reduce the size of the labelled sample and this poses major difficulties (e.g. in Section 6 Table 4, we observe that less than 10 fraud cases are present in 7 out of 10 sectors under consideration). Finally, the imbalanced learning issue [37, 10, 67] is hard to deal with as well.

AD techniques, on the other hand, are more suitable in a VAT fraud detection environment. Such methods can develop their model based on the entire (unbiased and unlabelled) population instead of the small sample of (biased and labelled) audited instances. They enable the detection of new types of fraud since, by assumption, fraud cases exhibit a behaviour that deviates from legal types of conduct [20]. Also, concept drift is easily addressed by training an AD model based on all entities' tax declarations in a certain time interval of interest (it is easy to update AD models as no labelled data are required).

In the following paragraphs, we focus on the methodological problems that are addressed in this study (see Figure 1): (1) the importance of good input variables. (2) The need for fast algorithms and, finally, (3) the need for a proper evaluation methodology to asses fraud detection performance. We highlight several shortcomings related to these issues in the VAT fraud detection literature, though many of them are also encountered in the broader area of tax fraud (or other financial fraud domains).

Provost & Fawcett [53] emphasize the requirement of a thorough data and business understanding in the predictive modelling effort. It is of the utmost importance that VAT domain knowledge is integrated in the feature development stage so that useful variables can be defined and shared with the research community. Studies in the VAT domain (see Table 1) fall in two categories: the first set of studies (see [9, 5, 44]) do not disclose the specific variables adopted in their work, yet mention general statements such as: *'information from tax declarations is integrated with data from other sources [9]'*. The second category of studies (see [35, 74, 13, 4, 46, 45]) partially disclose their defined variables, some of which take the form of tax ratios. However, their proposed tax ratios (e.g. tax/turnover, purchases/sales, income/assets) are defined from a high-level perspective and do not include specific fields of the VAT form. The use of ratios is common in related fraud areas as well (e.g. financial ratios to detect financial statement fraud [56, 38]).

The supervised classification studies in the VAT fraud detection domain do not focus on the need for fast algorithms because of the limited sizes of the labelled sample (see Section 1). In the tax area of customs, scalability considerations are more important because the number of labelled cases is far higher (audits require less time) and the large number of declarations need to be processed within a matter of seconds in an on-line environment. However, customs fraud detection studies commonly ignore the need for fast classification algorithms and rely on relatively slow techniques such as decision trees and neural networks [41, 36, 61]. Similarly, slow supervised algorithms such as neural networks and non-linear support vector machines (SVMs) are encountered in the broader financial fraud detection literature [47, 71].

AD methods should scale to the large data sizes that can occur in today's tax fraud detection problems because they should train and evaluate their models on the entire population. The VAT fraud detection study of Assylbekov et al. [4] assumes a multivariate Gaussian density function to assign the anomaly scores by estimating its parameters from the largest cluster obtained from a self-organizing map (SOM) clustering process. However, their proposal is computationally prohibitive for large datasets [57, 16]. In the tax area of customs, the AD based approach of Rad et al. [55] is slow because it relies on nearest neighbour computations (as explained in Section 2.2). The cluster based AD method of de Roux et al. [58] designed to target urban delineation tax fraud also suffers from scalability issues (spectral clustering and kernel density estimations are slow methods). In Section 2.2, a general literature review on the topic of scalable AD is conducted.

There are two concerns related to the evaluation of fraud detection models in the VAT fraud detection literature. (1) All studies indicated in Table 1, except for the supervised detection methodology of Mittal et al. [46], evaluate their detection models solely based on the audited set. However, as this labelled sample is not representative for the population (as explained in Section 1), one should have serious doubts whether performances reported on the audited set generalize to the entire population. Note that the issue of performance assessment that is limited to the collection of historically obtained labelled data is also encountered in the financial fraud detection literature. We refer to Kumar & Nagadevara [41], Shao et al. [61], Ravisankar et al. [56] and Pozzolo et al. [52, 51] for examples thereof. (2) The evaluation metrics that are commonly employed (accuracy, sensitivity, specificity) in (tax) fraud detection studies are derived from the confusion matrix. The latter is constructed based on a certain threshold (cut-off value) applied to the output scores of the model. Usually though, the choice of threshold is not discussed in these studies (e.g. see the works of Bonchi et al. [9], Gupta & Nagadevara [35] and Castellón González & Velásquez [13]) and a default (built-in) cut-off value is adopted, which could be irrelevant with respect to the available resources of a tax administration. Vanhoeyveld & Martens [67] and Pozzolo et al. [51] indicate that it is important to integrate capacity constraints in the evaluation process.

2.2. Related work on anomaly detection

AD methods detect anomalies (outliers) which show characteristics that differ markedly from the remaining population [2]. They have demonstrated their success across various fraud domains, such as credit card fraud [7], healthcare fraud [63], customs fraud [55] and urban delineation tax fraud [58]. All AD techniques rely on assumptions to distinguish anomalies from normal data (e.g. outliers have a larger distance to their nearest neighbours compared to normal examples). The effectiveness of any AD algorithm therefore depends on whether these assumptions are met for the dataset under consideration. AD methods can assign an anomaly score to each instance which is directly proportional to its probability of being an outlier. The latter is convenient in the tax domain because investigators will only examine a top list containing the most suspicious cases (the ones with the highest anomaly scores) in correspondence with their available capacity.

Outlier detection methods generally assume two basic conditions are fulfilled in order to work well [20, 15]: (1) the number of anomalies is far less compared to the number of normal instances. In a context of tax evasion, the first condition is satisfied [5]. (2) Anomalies, by definition, show a behaviour (in terms of their feature representation) that sets them apart from the normal instances. This second condition seems to be reasonably satisfied, though fraudsters may try to disguise their behaviour and attempt to approximate legal types of conduct. On the other hand, Castellón González & Velásquez [13] noted tax fraud cases to occur among extreme values of variables. We also refer to the related discussion in Section 4.2, where we explain that a fraudster may try to 'game' with the values reported in its VAT declarations, but, in doing so, may 'overshoot' his corrections. Hence the use of AD techniques seems plausible in a VAT environment.

Over the years, a variety of AD methods have been proposed that can be categorized into classification based, nearest neighbour based, clustering based and statistical techniques, where each category is explained in the survey of Chandola et al. [15]. Goldstein & Uchida [33] note the nearest neighbour and clustering based techniques to be the most popular categories. They find the nearest neighbour based algorithms to outperform the clustering based methods. This provides a motivation for our choice of nearest neighbour based AD techniques. In a recent (2016) study, Campos et al. [12] compared 12 nearest neighbour based algorithms and reached the following conclusion: "*After about* 15 *years of research, in general, the seminal methods knn*[11] *and Local Outlier Factor (LOF) remain the state of the art.*" In our study, we therefore included a variant of knn (see Section 5.1) and the LOF technique (see Section 5.2).

An important distinction between AD algorithms concerns the aspect of locality, where there is a difference between global (e.g. knn) and local methods (e.g. LOF) [75, 60, 12]. Regarding the former, a measure of 'outlyingness' (e.g. in knn, this is the average distance of a point to its k nearest neighbours) is defined and subsequently calculated for each instance in the dataset. The computed outlier scores represent global scores for each observation. Local methods similarly propose a measure of outlyingness, yet a *ratio* is employed that compares the outlyingness of an object to the same measure applied to its direct neighbourhood (e.g. a cluster of similar points or, with LOF, its k nearest neighbours).

AD techniques should scale to the large data sizes that can occur in the tax area. The nearest neighbour based AD methods that we adopt in this study are amongst the top performers in AD benchmark studies [33]. Unfortunately, they have a large complexity[12] of $O(N^2 \cdot d)$, with N the number of instances and d the dimensionality of the dataset, which excludes them from being used for large datasets [73]. Luckily, a couple of data compression techniques do exist that can improve the efficiency of any AD method. They can be divided into a reduction of the data size at the feature level or the instance level. The latter are more appropriate in a context of tax fraud because these problems are usually characterized by a large number of instances of low dimensionality. The instance based approaches are briefly reviewed next.

Data reduction techniques focusing on the instance level are called sampling (or subsampling) methods. Vanhoeyveld & Martens [68] recently proposed a new sampling approach, that we use in this study, which

[11] The knn anomaly score of instance x of the dataset D is the average distance of x to its k nearest neighbours in D.

[12] Nearest neighbour based techniques can be accelerated with Kd-trees. However, these techniques are not recommended for large dimensions and depend heavily upon the specific data structures encountered. MATLAB by default uses this approach only if $d < 10$ which is why we refrain from using Kd-trees.

addresses the shortcomings of the two remaining alternatives of random sampling [73] and density-biased sampling [39]. This approach gives rise to a new class of scalable AD methods called the Fixed-Width Anomaly Detection (FWAD) algorithms and is presented in more details in Section 5.1. The success thereof is demonstrated across various domains such as text/speech/image recognition, the medical domain, intrusion detection and spam detection. Vanhoeyveld & Martens [68] concluded that, in general, FWAD algorithms are significantly faster than their underlying AD method if the latter has a superlinear computational complexity in the number of instances N (e.g. the nearest neighbour based AD techniques). More benefits are obtained if the dataset has a large number of instances of low dimensionality. FWAD is outperformed by its underlying AD method in terms of predictive performance in general, because compression inherently implies an information loss, though this depends on the AD technique, the dataset characteristics and the amount of compression. However, there do exist datasets for which FWAD outperforms AD.

The specification of hyperparameter values for AD algorithms (e.g. the number of neighbours k for knn and LOF) in an unsupervised fashion is an important issue. The choice thereof has a large impact on the detection performance [65, 12]. Unfortunately, supervised hyperparameter tuning approaches are ruled out due to the absence of labelled data. Alternatively, one could opt to choose the parameter combination that optimizes an internal evaluation measure[13] tailored to the AD task. However, the existing works in this direction of Clémençon & Thomas [16] and Marques et al. [42] are too computationally involved to be applied to the large datasets that can be encountered in the tax domain. We should note that a number of unsupervised hyperparameter tuning heuristics have been developed, we refer to Song et al. [62] and Ghafoori et al. [32] for examples, yet they can only be used for a specific AD method (they do not generalize beyond the intended AD algorithm) and do not scale well with large data sizes.

Recently, Vanhoeyveld & Martens [68] developed a new algorithm to tune the hyperparameters of any AD algorithm in an unsupervised fashion resulting in a Pseudo-Optimal (PO) parameter choice that we adopt in our work. Given that many labelled datasets for the AD task have been made publicly available, the idea is to exploit this data repository to leverage suitable parameter settings that can subsequently be used for the particular unlabelled dataset at hand. The PO hyperparameter choice corresponds to the parameter setting that has the lowest average rank across the external labelled datasets. The authors compare a PO parameter choice to an average[14] performance across a range of hyperparameter settings for a variety of AD algorithms and datasets. For all AD methods (including the ones we use in this study), the PO choice statistically outperforms the average choice at the $\alpha = 0{,}10$ significance level according to the Wilcoxon signed-rank test [72, 18] with the Area Under Curve (AUC)[15] [28] performance measure.

[13] Internal evaluation measures are performance metrics that do not rely on labelled data and are solely based on the dataset characteristics (the feature values of the instances) and the anomaly scores assigned by an AD method.

[14] Due to the issues surrounding unsupervised hyperparameter tuning, a common approach adopted in recent AD benchmark studies is to report the average (or best) performance across a range of hyperparameter settings [33, 12, 19]. Alternatively, arbitrary (unmotivated) hyperparameter choices are also common.

[15] The AUC is a supervised evaluation metric and corresponds to the probability that an anomaly is ranked higher than a

3. Contributions

The literature review in Table 1 shows that AD methods have been rarely used for VAT fraud detection. With this work, we contribute in filling this void. In a former case study related to AD based VAT fraud detection conducted in Kazakhstan [4], a statistical outlier detection technique was able to outperform the predictive model of the tax administration in terms of cumulative revenues on the labelled sample. However, there are a number of issues, that we circumvent in our study, related to their proposal: (1) statistical AD methods are not amongst the top performers in AD benchmark studies [33]. Furthermore, the assumption that VAT declarations follow a multivariate Gaussian distribution has not been motivated nor validated. (2) The proposed algorithm is not applicable for larger datasets (as indicated in Section 2.1). (3) Performances are assessed based solely on the set of audited instances.

Section 1 indicates that non-identical market conditions and other tax law rules give rise to varying behaviours of companies across sectors. These sectoral differences motivate the development and assessment of a fraud detection model per sector. To the best of our knowledge, we are the first to conduct such an individual sector analysis and to frame the VAT fraud detection problem as a contextual[16] AD problem. The latter is implied when an instance can be considered as normal in a certain context, yet the same observation (in terms of its (behavioural) attributes) should be perceived as anomalous in a different context [15, 33]. In our study, the context of an instance coincides with its sector. Note that de Roux et al. [58] indicate that it is important to only compare tax declarations that are similar to each other. The contributions geared towards solving the methodological problems formulated in Section 2.1 are outlined next. We also refer to Figure 1.

As discussed in Section 2.1, the current VAT fraud detection studies do not share their adopted features or define high-level attributes that do not include specific fields of the VAT form. In Section 4.2, we propose a set of presumed fraud indicators for the specific domain of VAT. They take the form of tax ratios and can, with some modifications,[17] be used in future studies or adopted by tax administrations worldwide. We believe that we are the first to disclose such an exhaustive list of tax ratios based on specific fields of the VAT form and client listings. In Section 7, we show that the combination of tax ratios is predictive for VAT

normal example. The AUC is the most widely adopted/preferred evaluation metric in the area of AD [33, 12].

[16] The context defines the environment/world where the instance resides and is a vital part of the problem formulation. Chandola et al. [15] indicates that each instance is described by contextual attributes (that fix the context) and behavioural features (that describe the behaviour of an entity in the specified context).

[17] The main idea of designing tax ratios based on individual fields of a VAT declaration form is broadly valid. However, the specific development of these variables depends on a country by country basis. This is because non-identical VAT regimes and declaration forms are observed in different countries (even within the EU) [54]. If the analysis is conducted in a different country, one should assess the suitability of each tax ratio that we have proposed based on whether the required information is present in that country's VAT declarations (this is why we have included a description of each tax ratio in Table 3). Hence some of our variables may still be of use (e.g. Var_{18} is a basic ratio of sales to purchases and should be broadly valid), whereas others may not. Furthermore, new tax ratios could be developed according to the specificities of each country. Also note that countries with a detailed VAT form (e.g. Belgium) generally consider it an important element for data mining based risk analysis [54]. For some countries, more elaborate analyses may be possible (e.g. more logical checks) while, for others, the situation will be the opposite [54].

fraud detection.

In a VAT context, batch sizes can become large and this poses difficulties for several AD methods such as knn (see Section 2.2). In Section 5, we therefore develop the baseline (BL) and FWknn methods. We demonstrate their scalability in Sections 5.1 and 7.5. The BL approach is very intuitive and exploits the design of the proposed tax ratios. The FWknn approach is a particular instance of the class of FWAD methods, as introduced in Section 2.2, with knn as underlying AD technique. Note that the FWknn method was proposed by Vanhoeyveld & Martens [68], but it has not yet been validated in the area of tax evasion.

The common approach of evaluating VAT fraud detection models based on the labelled sample of historically conducted audits is questionable (as discussed in Section 2.1). We assess performances by looking at the top list (containing the instances with the highest anomaly scores) that is produced by applying our detection model to the entire population. This list should differ markedly from the top list that is limited to the audited set and should provide a more accurate representation regarding the nature of the instances that are flagged [4]. In Section 7, Table 5, we show that many instances appearing in the top list of the population differ from the ones pertaining to the audited set. With this methodology, we have the guarantee that (known or unknown) fraud cases appearing in the top list of the population, as produced by our detection methods, would be effectively detected if the tax administration verifies this list.[18] To the best of our knowledge, we are the first AD based study that comes with such a guarantee (and the second[19] VAT fraud detection study).

Many studies in the VAT domain ignore the fact that tax administrations have limited resources and can only inspect a small fraction of entities eligible for auditing (as discussed in Sections 1 and 2.1). In Sections 7.2 and 7.3, the lift and hit rate metrics are proposed that take such capacity constraints into account. Note that these measures themselves are not new, but the novelty lies in the adoption thereof in the area of VAT fraud detection.

To summarize, AD methods are applied to all instances pertaining to the same sector (a sector analysis) and where each company is represented by a set of tax ratios (fraud indicators). In Section 7, we demonstrate that this methodology is successful in detecting VAT fraud and therefore contributes towards an efficient auditing strategy. Also, tax administrations worldwide have access to similar data repositories and our methodology could therefore be transferred with some modifications to their problems as well.

[18] There is no guarantee that a detection method flags any fraud cases when examining a top list limited to the set of audited instances, because, in practice, the detection method would be applied to the entire population and may therefore result in a completely different top list.

[19] Mittal et al. [46] also report performances on the top list of the population. However, their work adopts a supervised fraud detection method which requires a special type of cross-validation procedure to generate predictions for the entire population. Our work is similar to their evaluation procedure, but is tailored to the unsupervised setting.

4. Data

4.1. Raw data

The Belgian tax administration provided us with fully anonymized periodic[20] VAT declarations and client listings[21] for tax year 2014 for a subset of organisations with a Belgian VAT number[22] pertaining to 10 anonymized sectors. Each sector corresponds to a 5-digit NACE activity code and was selected such that (a) it contains an adequate number of organisations and (b) it has a relatively high or low prevalence of (known) fraud. Furthermore, the Belgian tax administration did not disclose any of the aforementioned information regarding mixed taxpayers, ‘small’ companies and VAT units (as discussed in Section 6).

The VAT declarations provide the tax administration with all the required details to determine the tax liability. This information can be divided into four main categories: outgoing transactions (sales); incoming transactions (purchases); taxes due and deductible taxes. The VAT declaration form [29] contains a total of 34 grids that need to be completed by the declarant. Table 2 provides a brief description of these grid codes and they will be used in the construction of the tax ratios in Section 4.2. An elaborate explanation regarding all grid codes can be found in the VAT manual of Govers & Deschacht [34].

Besides the aforementioned VAT returns, Belgian companies also have to present a yearly client listing. This is a list of all[23] the Belgian VAT numbers to whom the company under consideration has delivered goods or performed services for a total amount of more than €250 (excluding VAT). The list contains, for each client, the total revenue (excluding VAT) and the VAT charged throughout an entire year.

Due to confidentiality and the highly sensitive nature of the data, only sector aggregates of fraud signals, investigations and cases were communicated rather than labels on individual Belgian companies (see the related discussion in Section 6). Therefore, unsupervised techniques need to be applied to the data at hand.

4.2. Tax ratios

In AD, one wishes to find those entities that display a behaviour that is significantly different from the normal behaviour within the population under consideration. The conduct of a company is captured through a set of manually defined variables (presumed fraud indicators) that take the form of tax ratios. Table 3 presents an overview of the 18 tax ratios that are used in this study, motivated by domain knowledge. Note that all variables from the periodic VAT returns are aggregated over an entire year, since the client listings have a one year periodicity. Each tax ratio is designed from a perspective that a relatively high or low value with respect to the rest of the population could arouse suspicion. Castellón González & Velásquez [13] already noted fraud cases to occur among the extreme values of variables. Due to confidentiality reasons,

[20] Depending on the company characteristics, this can be monthly or quarterly.

[21] The client listings data of the suppliers of the companies under consideration were also disclosed.

[22] In the remainder of this article, for simplicity reasons, we’ll call these Belgian companies.

[23] Belgian clients that perform only VAT exempted actions are not incorporated in the listings.

Table 2: Description of grid codes in the Belgian VAT declaration form (free translation of authors). Category II = outgoing transactions; category III = incoming transactions; category IV = taxes due and category V = deductible taxes.

Category	Grid	Description
II	[00]	Actions subject to a special arrangement
II	[01]	Actions for which VAT is due for the declarant at a rate of 6%
II	[02]	Actions for which VAT is due for the declarant at a rate of 12%
II	[03]	Actions for which VAT is due for the declarant at a rate of 21%
II	[44]	Services for which the foreign VAT is due by the co-contractor
II	[45]	Actions for which the VAT is due by the co-contractor
II	[46]	Exempt intra-community supply of goods conducted in Belgium and triangular ABC-sales
II	[47]	Other VAT exempt actions and other actions conducted abroad
II	[48]	Amount of the issued credit notes and the negative corrections with respect to actions declared in grids [44] and [46]
II	[49]	Amount of the issued credit notes and the negative corrections with respect to the other actions declared in category II
III	[81]	Amount of the incoming transactions taking into account the received credit notes and other corrections with respect to commodities, raw and auxiliary materials.
III	[82]	Amount of the incoming transactions taking into account the received credit notes and other corrections with respect to services and miscellaneous goods
III	[83]	Amount of the incoming transactions taking into account the received credit notes and other corrections with respect to business assets.
III	[84]	Amount of the received credit notes and negative corrections with respect to actions indicated in grids [86] and [88]
III	[85]	Amount of the received credit notes and negative corrections with respect to the other actions declared in category III
III	[86]	Intra-community acquisitions conducted in Belgium and triangular ABC-sales
III	[87]	Other incoming transactions for which the VAT is due by the declarant
III	[88]	Intra-community services with reverse charge mechanism
IV	[54]	VAT on actions declared in grids [01], [02] and [03]
IV	[55]	VAT on actions declared in grids [86] and [88]
IV	[56]	VAT on actions declared in grid [87], excluding imports with reverse charge mechanism
IV	[57]	VAT on imports with reverse charge mechanism
IV	[61]	VAT adjustments in favour of the State
IV	[63]	VAT refund due to the receipt of credit notes
IV	[65]	Not to be completed
V	[59]	Deductible VAT
V	[62]	VAT adjustments in favour of the declarant
V	[64]	VAT to be recovered due to issued credit notes
V	[66]	Not to be completed

we are not able to disclose the specific fraud types that are potentially captured with each individual ratio. In general, the combination of fraud indicators is presumed to enable the detection of (but not limited to) the following fraud types: the use of fake invoices (potentially with a foreign client); feigning sales to private individuals; under-declaration of revenues and the reporting of fictitious incoming transactions.

One may wonder whether disclosing the tax ratios is potentially harmful for tax administrations. Fraudsters adjust the information they disclose to the government and consequently ‘game’ with the value of a ratio. However, changing the numerator or denominator of a single ratio would create a snowball effect with respect to the values of the other ratios. Hence this would require multiple adjustments that affect several fraud indicators. The main argument, however, relates to the fact that a fraudster would need to know the common value (the behaviour) of a tax ratio within the population and this information is unavailable from the perspective of a fraudster.

Table 3: Tax ratios (18 in total) obtained from the VAT declarations and client listings.

Name	Tax ratio	Description
Var_1	$\frac{[01]}{[01]+[02]+[03]}$	Proportion of the outgoing transactions at a rate of 6%.
Var_2	$\frac{[02]}{[01]+[02]+[03]}$	Proportion of the outgoing transactions at a rate of 12%.
Var_3	$\frac{[44]+[45]+[46]+[47]}{[01]+[02]+[03]+[44]+[45]+[46]+[47]}$	Proportion of sales not subject to domestic (Belgian) VAT.
Var_4	$\frac{\sum_{OUT,Listing}(revenue{>}0)I(VAT)}{[01]+[02]+[03]}$	Ratio of sales in client listing to sales in VAT declarations that are subject to domestic (Belgian) VAT. The numerator contains the total revenues that are subject to VAT and excludes credit notes.
Var_5	$\frac{[48]+[49]}{[01]+[02]+[03]+[44]+[45]+[46]+[47]}$	Ratio of issued credit notes and negative corrections to the total revenues from sales.
Var_6	$\frac{[86]+[88]}{[81]+[82]+[83]}$	Ratio of Intra-community acquisition of goods and services to the total amount of incoming transactions.
Var_7	$\frac{[84]}{[84]+[86]+[88]}$	Ratio of amount of received credit notes and negative corrections (wrt grids [86] & [88]) to the total Intra-community acquisition of goods and services.
Var_8	$\frac{[85]}{[81]+[82]+[83]+[84]+[85]}$	Ratio of received credit notes and negative corrections (wrt other actions than grids [86] & [88]) to the total amount of incoming transactions.
Var_9	$\frac{[81]+[82]+[83]-[86]-[87]-[88]}{\sum_{IN,Listing} revenue(VAT{\neq}0)}$	Ratio of domestic (Belgian) purchases to the total purchases (subject to VAT) as indicated in the client listing.

Table 3: Continued.

Name	Tax ratio	Description
Var_{10}	$\frac{[54]}{0{,}06\times[01]+0{,}12\times[02]+0{,}21\times[03]}$	Ratio of declared VAT due to sales indicated in grids [01], [02] & [03] to the 'correct' VAT from those sales.
Var_{11}	$\frac{[55]}{[86]+[88]+[84]}$	Ratio of VAT due as a result of intra-community purchases (grids [86] & [88]) to the total amount from those purchases
Var_{12}	$\frac{[56]+[57]}{[87]}$	Ratio of VAT due as a result of actions indicated in grid [87] to the amount declared in this grid.
Var_{13}	$\frac{\sum_{OUT,Listing} VAT(VAT>0)}{[54]}$	Ratio of VAT collected from all clients as indicated in the client listing (excluding credit notes) to the VAT declared through domestic (Belgian) sales.
Var_{14}	$\frac{[61]+[63]}{[54]+[55]+[56]+[57]+[61]+[63]}$	Ratio of VAT corrections in favour of the government and VAT on received credit notes to the total amount of VAT due.
Var_{15}	$\frac{[59]}{[81]+[82]+[83]+[84]+[85]}$	Ratio of deducted VAT (excluding corrections) to the total amount of incoming transactions.
Var_{16}	$\frac{[62]+[64]}{[59]+[62]+[64]}$	Ratio of VAT corrections in favour of declarant and VAT on issued credit notes to the total amount of deducted VAT.
Var_{17}	$\frac{\sum_{OUT,Listing} revenue.I(VAT)}{\sum_{OUT,Listing} revenue}$	Fraction of sales in client listing that are subject to VAT.
Var_{18}	$\frac{[01]+[02]+[03]+[44]+[45]+[46]+[47]-[48]-[49]}{[81]+[82]}$	Ratio of sales to purchases (excluding purchases of business assets)

4.3. Preprocessing

The fraud indicators designed in the previous section require some additional preprocessing. First of all, for a tax ratio under consideration, its mean μ, median and standard deviation *std* for all the Belgian companies pertaining to the same population (sector) are calculated (excluding companies that show infinity or undefined values). Secondly, all undefined values (numerator and denominator both zero) are replaced with the median value of the population. Thirdly, a standard statistical normalization is applied to all variables to ensure that all attributes are expressed in the same scale:

$$Var_{prep} = \frac{Var - \mu(Var)}{std(Var)}. \tag{1}$$

Finally, if $Var_{prep} > 3$ or $Var_{prep} < -3$, the preprocessed variable gets assigned a value of 3 or -3 respectively. We have included this operation because this delivered better results, as indicated by the tax administration, compared to previous experiments that exclude this truncation. Without this operation, extreme values on a single tax ratio would dominate the end result. Hence it is better to flag companies that have more variables with slightly less extreme values than companies with few variables having very unusual values. The particular values of 3 and -3 were agreed upon with the tax administration, though we cannot guarantee their optimality as we do not focus on investigating which type of preprocessing[24] is most optimal for the current application.

5. Method description

Due to the fact that we didn't have access to any labelled data, we were limited in the number of top lists (with most suspicious cases, see Section 6) that could be send to the tax administration for verification. As a direct consequence, a careful consideration was made regarding the choice of fraud detection methods. We motivated our choice for nearest neighbour based AD methods in Section 2.2.

5.1. Fixed-Width Anomaly Detection (FWAD)

FWAD algorithms rely on a Fixed-Width (FW) clustering process as a first data compression step. In FW clustering, the data partitions are allowed to overlap (instances can belong to multiple clusters). Each cluster represents a hypersphere of radius (width) W around its centre. All instances falling within a (Euclidean) distance of W to this centre are associated with the cluster. The FW clustering approach relies on the following steps [50, 20, 14]: in a first step, the dataset is traversed iteratively in order to determine the cluster centres. If, during a certain iteration, the instance under consideration has a distance larger than W to all existing cluster centres, the data point forms the centre of a new FW cluster. In any case, the algorithm continues with the next instance until the entire dataset is processed. In a second step, each instance of the dataset is associated with those FW clusters that have a distance smaller than W to the instance. Because the cluster centres coincide with a small subset of data points from the original dataset, FW clustering can be regarded as a sampling approach. Its complexity is $\mathrm{O}(N \cdot |FWC| \cdot d)$. If the radius W is assigned a sufficiently large value, the number of FW clusters $|FWC|$ will be several orders of magnitude smaller than the number of instances N, making the procedure scalable for large datasets.

We present a high-level description of the class of FWAD techniques in this paragraph. For a more elaborated technical discussion, containing a pseudo-code implementation thereof, we refer to Vanhoeyveld & Martens [68]. Figure 2 illustrates the basic steps of a FWAD algorithm. In a first step, the FW clustering technique is used to obtain a small subsample of the dataset (the FW cluster centres). This process keeps track of the number of data points (*Count)* assigned to the FW clusters. This *Count* term is an unnormalized approximation of the local density because each cluster has the same radius W (and hence the same volume). The latter approximation is useful as anomalies occur in low-density regions of the (true yet unknown)

[24] Campos et al. [12] recommends the normalization of datasets (preprocessing) as this outperforms unnormalized datasets (without preprocessing). They indicate that the issue of normalization is rarely addressed in the AD literature and hence it is unclear which type of preprocessing is most recommended.

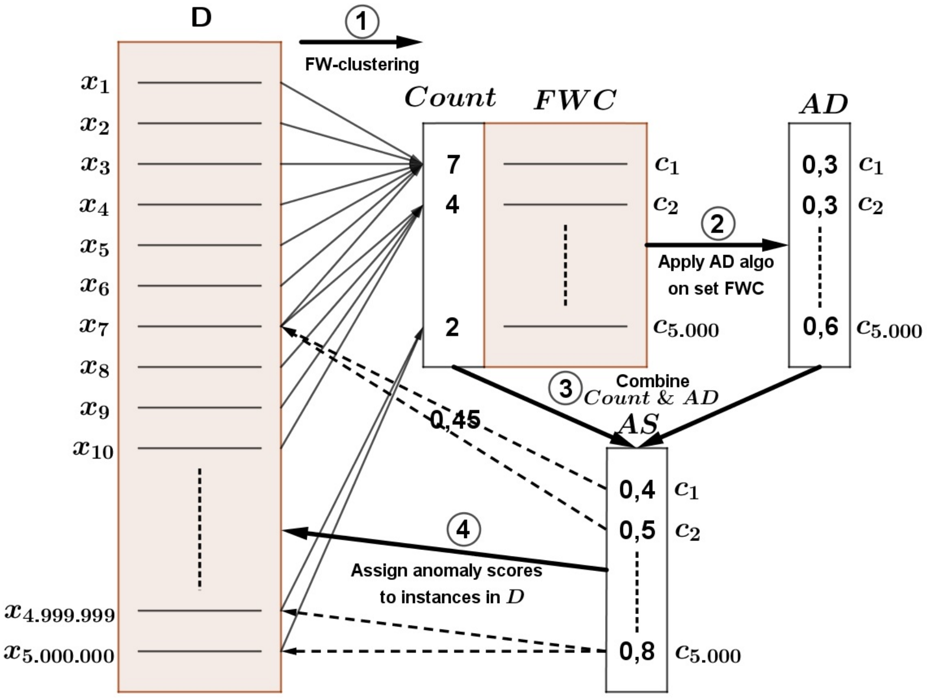

Figure 2: High-level overview of FWAD's basic components.

probability density function $f(x)$ that generated the data [16]. In a second step, a standard AD algorithm is applied to the compressed dataset of FW clusters of size $|FWC|$ (which is much smaller than N and the application of the AD method should therefore be substantially faster than for the original dataset). In a third step, FWAD aggregates the *Count* term (step 1) and the AD algorithm's score (step 2) into a single anomaly score for each FW cluster centre. Finally, in a fourth step, the clustering nature of FW clustering is exploited to assign outlier scores to every instance of the original dataset by averaging the anomaly scores of its associated FW clusters.

The knn AD algorithm is computationally prohibitive for large datasets due to its $O(N^2 \cdot d)$ complexity (as discussed in Section 2.2). We therefore integrate the knn AD method in the FWAD formulation and call this the FWknn technique. The latter has a computational complexity of $O(N \cdot |FWC| \cdot d)$ (FW clustering) + $O(|FWC|^2 \cdot d)$ (knn on set of FW clusters).

The compression level of the FWAD method is an important hyperparameter that controls the predictive performance versus time requirements trade-off [68] and it is governed by the choice of width percentage W_p (which implicitly defines the cluster width W). In general, increasing the compression level W_p translates into significant speed-ups at the cost of a potential decrease in predictive power. In Section 6, we will indicate and motivate our choice of compression level W_p.

5.2. LOF

Local Outlier Factor (LOF) [11] was the first proposed local AD method and compares the local density of the nearest neighbours of a point to the local density of the point itself. It assumes that an anomaly has a relatively larger distance to its neighbours compared to the same distance computed for its neighbours. The calculation of the LOF score involves the following steps:

- Find the k nearest neighbours of instance x, these form the set $N_k(x)$.
- Approximate the local density for a record x using Euclidean distance metric[25]:

$$LD(x) = 1 \bigg/ \frac{\sum_{o \in N_k(x)} d(x,o)}{|N_k(x)|} \tag{2}$$

- Compute the LOF score as a ratio of local densities:

$$LOF(x) = \frac{1}{|N_k(x)|} \sum_{o \in N_k(x)} \frac{LD(o)}{LD(x)}. \tag{3}$$

The algorithm depends on a single parameter k that controls the size of the neighbourhood. LOF has a complexity[26] of $O(N^2 \cdot d)$.

5.3. BL

Besides the FWknn and LOF methods, a baseline (BL) approach is proposed, where the anomaly score of an instance corresponds to the sum of the absolute values of the preprocessed fraud indicators. Because each individual tax ratio was designed from a perspective that a relatively high or low value with respect to the rest of the population is suspicious, the BL method is a very intuitive approach to combine the 'suspiciousness' scores[27] of the individual variables. We assume that when such relatively high or low values are reported for more variables, this should be regarded as more suspicious. The BL technique will flag those companies that show the largest absolute values for as many preprocessed tax ratios as possible. Also, it is very attractive in terms of computational requirements. The BL method has a linear complexity $O(N \cdot d)$. Furthermore, it does not require the specification of a hyperparameter.

[25] The original formulation employs a slightly different measure of distance, called the reachability distance. Except in some very rare situations in highly dense clusters, this is the Euclidean distance [33]. In Campos et al. [12], no significant statistical difference is observed using either definition.

[26] We have included the local LOF method to illustrate differences with the global FWknn and BL approaches. One could design a FWLOF method. However, because the *Count* term in FW clustering is a global outlier score, FWLOF would be a hybrid between global and local methods. We wanted to assess performances based entirely on a local approach.

[27] The suspiciousness score of a variable corresponds to the absolute value of the implied preprocessed tax ratio.

6. Methodology

The evaluation methodology presented in Figure 3 consists of the following five steps of which the first two steps were performed internally at the tax administration: extract all companies pertaining to a given sector (activity code or NACE); apply filtering to discard companies that show certain pre-defined characteristics; perform pre-processing on the instance representation; use an anomaly detector to assign a fraud score (anomaly score) to each data point and finally, evaluate the results. The next paragraphs highlight each of the subsequent steps in more detail.

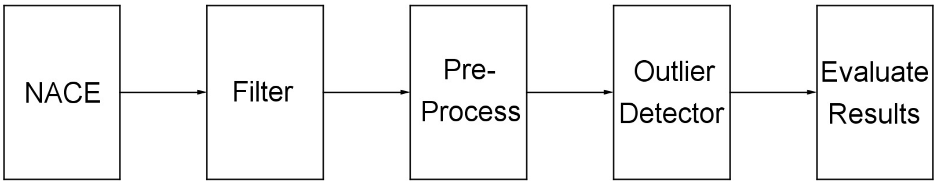

Figure 3: Overview of the evaluation methodology.

The Belgian tax administration provided information regarding a subset of organisations pertaining to 10 anonymized sectors (see Section 4.1). Table 4 includes summary information regarding each activity code. A detailed description of the columns $|A|$, $|S|$, $|I|$ and $|F|$ is provided later in this section. At this stage, we emphasize that an AD method will be applied to all companies pertaining to the same sector under consideration (an individual sector analysis).

Table 4: Activity code information (after filtering) regarding sector size $|A|$, signalisations $|S|$, investigations $|I|$ and fraud cases $|F|$.

Sector	Prevalence	$\|A\|$	$\|S\|$	$\|I\|$	$\|F\|$
S-A	Low	4.955	13	6	1
S-B	Low	13.939	54	33	3
S-C	Low	9.552	27	16	5
S-D	Low	11.066	43	25	6
S-E	Low	11.007	41	16	6
S-F	High	1.738	32	15	7
S-G	High	1.466	30	16	8
S-H	High	8.046	100	34	17
S-I	High	10.862	133	29	17
S-J	High	5.494	154	86	24

In the second stage, a filtering process is adopted by the tax administration to further reduce the number of companies within a certain sector. Mixed taxpayers[28] (grid [00] $\neq 0$) are discarded since AD might mark these companies as anomalous purely due to their intrinsic nature which differs from standard taxpayer

[28] Mixed taxpayers perform activities for which he is not liable to charge VAT and activities for which he is liable to charge VAT. In essence, VAT on purchases directly related to the activities that are not liable to VAT cannot be deduced.

behaviour. As such, they require a separate analysis. Also, the focus is on fraud cases of a sufficiently large monetary value. Therefore, companies reporting a tax amount due of less than €1.000 *and* a total amount of sales smaller than €7.500 *and* a total amount of purchases less than €5.000 are discarded. Note that all three conditions have to be satisfied simultaneously in order to be excluded. Another reason for this filtering relates to the fact that the fraud databases contain cases that are flagged in the years 2014 - 2017. Hence, we hereby ensure that there are at least some traces of activity in the VAT declarations of 2014. Furthermore, all VAT units[29] (and its members) are discarded since they have a different way of filling in the declaration form and client listings as compared to other Belgian companies. Note that they represent a small group that is limited to around 1% of all Belgian VAT numbers.

In a third stage, all instances from a certain activity code that remain after filtering undergo a preprocessing step. The behaviour of each Belgian company is characterized by a vector x containing the tax ratios $Var_i,\ i = 1, \ldots, 18$ described in Section 4.2. The preprocessing stage transforms this initial vector x into the feature vector x_{prep} according to the procedure outlined in Section 4.3.

Next, a fraud detection method (FWknn, LOF or BL, see Section 5) is applied to all instances in the sector under investigation. Note that all preprocessed tax ratios are considered simultaneously. Regarding the specification of hyperparameter values, we adopt the proposal of Vanhoeyveld & Martens [68]. They obtain the following parameter settings by applying a PO hyperparameter tuning procedure (as introduced in Section 2.2) to a collection of 14 labelled datasets tailored to the AD task:[30] $S_p = 10$, $W_p = 1$, $a = 0{,}2$, $k_p = 10$ for FWknn (we refer to Vanhoeyveld & Martens [68] to explain the meaning of these parameters) and $k = 20$ for LOF.

In the final stage, the AD algorithms are assessed with appropriate evaluation metrics. The description of these measures is presented in Sections 7.2 and 7.3. The evaluation mechanism entails the following process: for each of the ten sectors and for each of the three algorithms under investigation, the top 400 list (the 400 instances with the highest anomaly scores) is send to the tax administration for verification and is evaluated per multiple of 50 organisations. Each evaluation reports the aggregates of signalisations S, investigations I and fraud cases F. There is a natural ordering for these types in the sense that the set S consists of weak signals that were not substantial enough for follow-up investigation together with the strong signals (the set I); the set I contains (possibly ongoing) investigations with or without extra taxation and the set F consists of cases where additional taxation or fines were due. In mathematical notation, $F \subset I \subset S$.

[29] Under very strict conditions, several Belgian companies can form a VAT union. In this case, a new VAT number is established for the unit which is used for communication with the tax administration and contains the aggregated declaration of its individual members. Members of the unit retain their personal VAT number that is used in the yearly client listings data.

[30] A PO$_{AUC}$ parameter setting (that emphasizes predictive performance) is adopted and favoured over a PO$_{Time}$ setting (that focuses on reducing computational timings). This is because the analysis is conducted in a relatively small country (Belgium) and therefore the sizes of the 10 sectors under consideration are relatively small (see Table 4). Hence timings are not much of an issue in this Belgian case study. A PO$_{AUC}$ tuning procedure automatically results in the adoption of the lowest compression level ($W_p = 1$) within the proposed range of compression levels ($W_p = [1, 5, 10, 15, 20, 25, 30]$ [68]). Indeed, increasing the compression level reduces computational timings, yet it also decreases predictive performances.

Figure 4 shows the relevant groups occurring in our study. Besides the S, I and F groups explained in the previous paragraph, the figure also highlights the A group (represents all entities with the same activity 495 code); the group $F^{\star}$ (denotes unobserved fraud cases) and the T group (represents the entities appearing in a possible top $|T|$ list of an anomaly detector). We refer to Table 4 for a summary on the sizes of A, S, I and F for each of the ten sectors. Note that, due to capacity constraints faced by tax administrations, the sizes $|S|$, $|I|$ and $|F|$ are very small compared to $|A|$.

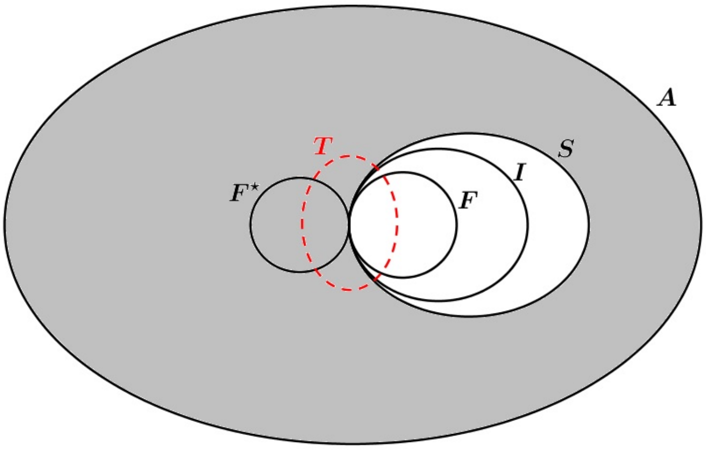

Figure 4: Representation of the relevant sets of a sector: all companies A, signalisations S, investigations I, fraud cases F, unobserved fraud cases $F^{\star}$ and a possible top list T of an AD technique. The relative sizes of the bubbles are not representative.

7. Results and Discussion

500 7.1. Feedback format

As indicated in the previous section, the top $|T|$ Belgian companies are matched with the fraud databases of the tax administration for each sector and fraud detection method under investigation. Their feedback reports the number of signalled cases in the top list: $|S_T = S \cap T|$; the number of investigated companies in the top list: $|I_T = I \cap T|$ and the number of fraud cases occurring in the top list: $|F_T = F \cap T|$. Table 5 505 presents the result of this procedure at the $|T| = 400$ stratum. With respect to the $|T|$ value, the following choice was agreed upon with the tax administration to reflect capacity constraints on their behalf:

$$|T| = [50, 100, 150, 200, 250, 300, 350, 400] \,. \tag{4}$$

In an environment of unlimited resources, tax administrations could choose the optimal $|T|$ value, denoted as $|T^{\star}|$, to be the setting that minimizes the total cost $C(T)$ on the population (we translate the work of Provost & Fawcett [53] to our setting): $C(T) = C_{audit}.|T| + C_{fraud}.(|F| + |F^{\star}| - |F_T| - |F^{\star}_T|)$, with 510 $F^{\star}_T = F^{\star} \cap T$. The first term reflects that cases being flagged by our method (occurring in T) should undergo an audit. The second term includes costs related to fraud cases that do not appear in T and would remain undetected. Because $F^\star$ and $F^\star_T$ are unknown, one can leave them out of the equation: $\tilde{C}(T) = C_{audit}.|T| + C_{fraud}.(|F| - |F_T|)$. C_{audit} denotes the expected cost of an audit and C_{fraud} denotes the expected monetary loss of a fraud case. Determining these costs requires a separate analysis by domain experts and is a difficult task [53] due to the VAT domain's characteristics outlined in Section 1. However, tax administrations face resource constraints which means that $|T^\star|$ cannot be achieved in practice and $|T|$ coincides with the available capacity [53]. Hence $|T|$ is not a design parameter but is fixed beforehand.

Table 5: Feedback results with $|T| = 400$ for each sector and fraud detection method under consideration. Results show the number of signalisations $S_T = S \cap T$, investigations $I_T = I \cap T$ and fraud cases $F_T = F \cap T$ occurring in the top T list.

Sector	BL			FWknn			LOF		
	$\lvert S_T\rvert$	$\lvert I_T\rvert$	$\lvert F_T\rvert$	$\lvert S_T\rvert$	$\lvert I_T\rvert$	$\lvert F_T\rvert$	$\lvert S_T\rvert$	$\lvert I_T\rvert$	$\lvert F_T\rvert$
S-A	2	1	1	3	1	1	3	1	1
S-B	7	6	1	6	5	1	5	4	1
S-C	2	1	1	2	1	1	1	0	0
S-D	8	5	1	6	4	0	2	1	1
S-E	3	1	0	4	1	0	1	1	1
S-F	16	7	3	12	6	3	8	4	3
S-G	18	10	5	14	10	4	12	10	5
S-H	15	8	4	17	9	4	8	5	2
S-I	7	3	2	10	3	2	2	1	0
S-J	12	7	1	13	8	0	11	5	2

7.2. Lift analysis

Lift curves [53, 6] are well suited to assess the prevalence of fraud cases among entities ranked highly by a scoring mechanism. They can be evaluated according to the available capacity. The lift compares the occurrence of fraud in the top $|T|$ list ($|F_T|/|T|$) with the overall (base) occurrence of fraud ($|F|/|A|$):

$$lift(T) = \frac{|F_T|}{|F|} \bigg/ \frac{|T|}{|A|} = \frac{|F_T|}{|T|} \bigg/ \frac{|F|}{|A|} . \tag{5}$$

If the method is any good, it is able to improve upon the base rate (i.e. $lift > 1$). Note that a random model[31] has a lift of 1. As an example, the sector S-B has $|F| = 3$ and $|A| = 13.939$ (see Table 4), corresponding to a base fraud rate of 0,0215%. A randomly picked company from this sector will have a probability to be fraudulent of 0,0215%. The FWknn method with $|T| = 400$ finds $|F_T| = 1$ fraud case in its top list (see Table 5). Hence the fraud rate improves to 0,25%, which is 11,6 (the lift) times larger than a random model would find.

Figure 5 comprehensively visualizes the $lift(T)$ values for each of the sectors, fraud detection methods and capacities under consideration. The horizontal line represents the lift of a random model. Values above

[31] A random model would make an arbitrary ranking of the instances corresponding to a strategy of random inspections.

this line indicate that the proposed method is able to find more known fraudsters (F) than a random model. In case the corresponding marker is missing, the lift is equal to zero meaning that there are no fraud cases in the top $|T|$ list. As the goal of the proposed methodology is to detect VAT fraud, one can expect/hope to retrieve a number of known fraudsters in the top list. The high lift values in most sectors reveal this is certainly the case. We can see that, depending on the sector, between 5 and 100 times more fraud cases can be detected with our methodology compared to a strategy of randomly auditing companies.

Many companies appearing in our top $|T|$ lists have not yet been signalled (see Table 5) and hence belong to the 'gray' area of Figure 4. For those entities, there is no information regarding their fraud/compliant nature. As the proposed methods are able to reveal known fraud cases, we can hope/expect to find undiscovered fraud cases in the top list pertaining to the group $F^{\star}$. In that case, the lift values would increase. Hence, it is important to note that the results displayed in Figure 5 should be regarded as lower bounds.

Regarding the effect of $|T|$, we analyse the LOF method for sector S-A (see Figure 5), as these results are mostly encountered for the other combinations as well. Initially, for very low values of $|T|$ (i.e. $|T| = 50$), there are no known fraud cases in the top $|T|$ list (i.e. $|F_T| = 0$) and the lift is 0. Usually, $|S_T| = 0$ as well and this means the top $|T|$ list consists of new cases that have not yet been signalled by the tax administration. When $|T|$ is increased (i.e. $|T| = 100$), known fraud cases start to appear in the top list resulting in non-zero lift values. Increasing the $|T|$ value even further results in either extra fraud cases being found that improve upon the lift value or no additional fraud cases are discovered and this decreases the lift value (i.e $|T| > 100$). Eventually, the lift values converge to 1 as $|T|$ becomes even larger.[32]

7.3. Hit rate analysis

Another interesting metric is to look at the ratio of the number of fraud cases to the total number of flagged cases (signalled or investigated). This ratio is called the hit rate (or strike rate or precision). The European Commission [24] noted that audits contribute 3,17% of the total VAT collected (on average for 22 EU member states). Hence the majority of revenues arise from a fear of being audited. From that perspective, ensuring a high hit rate is essential. Note that this boils down to minimizing the number of false positives (FP) policy, which is a popular strategy [9, 35, 5].

More formally, in our setting, we define the hit rate (HR) for signalled or investigated cases as follows:

$$
\begin{aligned}
HR_S(T) &= |F_T|/|S_T| && (6) \\
HR_I(T) &= |F_T|/|I_T| \,. && (7)
\end{aligned}
$$

Because we deploy a novel method, which has not yet been used by the Belgian tax administration, we can expect the entities appearing in our top lists to differ from the ones that have already been flagged. The

[32] if $|T| \to |A|$, then $|F_T| \to |F|$ and therefore $lift \to 1$ (see Equation 5).

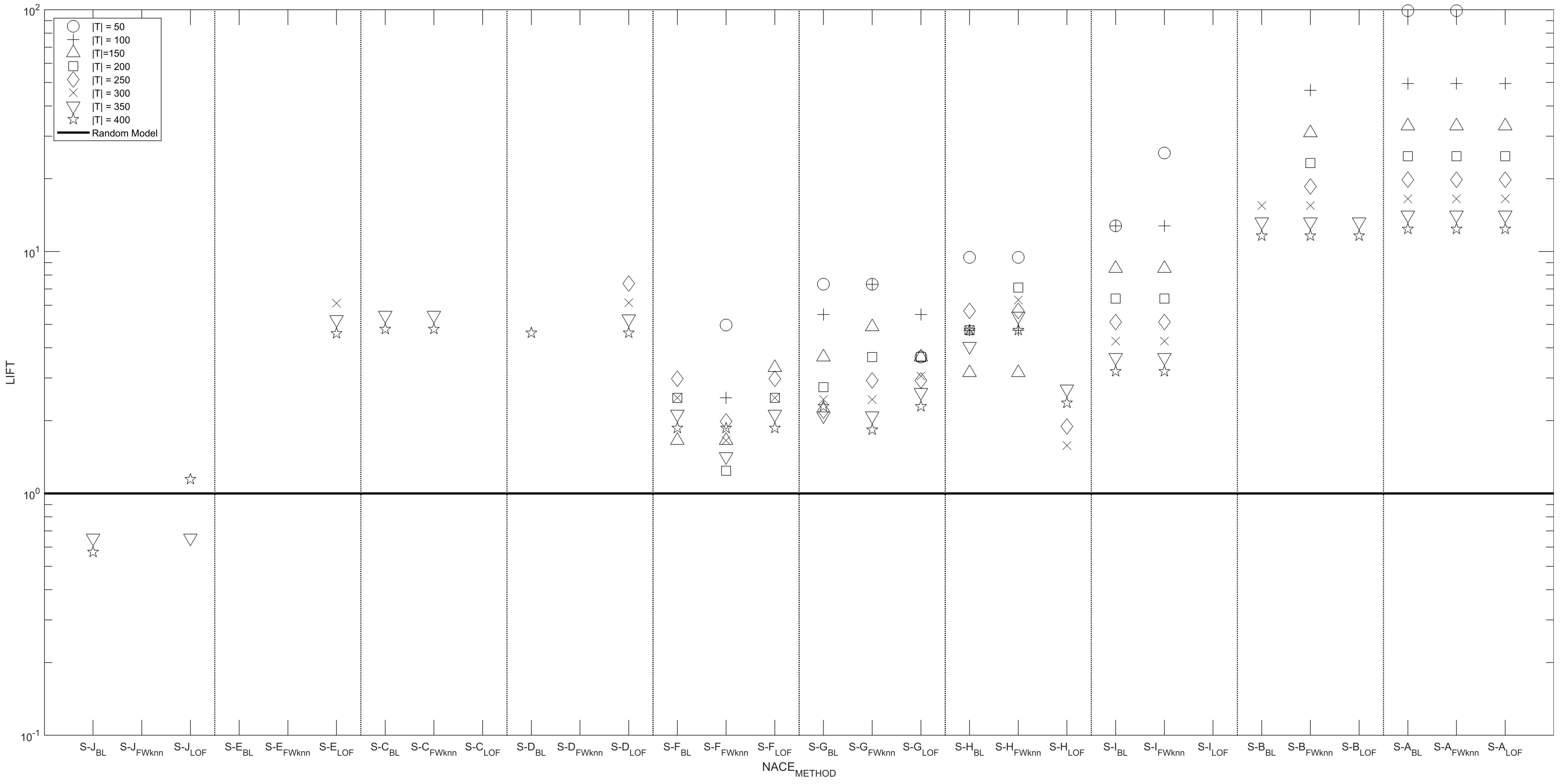

Figure 5: Lift at different $|T|$ values. Vertical lines separate different sectors.

low $|S_T|$ and $|I_T|$ values of Table 5 confirm this. Hence, for the majority of companies in our top $|T|$ lists, we don't know whether they are compliant or not. This is the main reason why the terms $|S_T|$ and $|I_T|$ appear in the denominator (as opposed to say $|T|$).

Figures 6 and 7 show the hit rates for each sector, fraud detection method and capacity value under consideration for the signalled and investigated case respectively. The overall (base) hit rate $HR_{Overall}$ corresponding to $|F|/|S|$ (signalled) or $|F|/|I|$ (investigated) is also shown for each activity code. In the case of signalled companies (see Figure 6), for at least one $|T|$ value, the proposed methodology is able to outperform the base HR in 22 out of 30 cases (10 sectors and 3 AD methods). Regarding investigated companies (see Figure 7), the results are less clear, with 20 wins out of 30 cases. The latter performance drop occurs because investigated companies have already undergone an a-priori analysis by a tax expert.

Two important remarks should be considered: (1) the HR in our top $|T|$ list should be compared with $HR_{Overall}$ of the tax administration for the same $|T|$ value. This would allow for a more fair comparison, however, we didn't have access to this kind of information. (2) the HRs in our top $|T|$ lists relate only to companies that have already been flagged by the tax administration (at the intersection of T with S or I). Whether these results generalize to the entire list can only be assessed in case additional audits are effectively conducted on these currently unsignalled companies. However, auditing is a very time consuming approach and this is the main reason results are lacking in this direction. Note that each of the studies in the literature review of Section 2.1 limit themselves to reporting performances on the sample of historic audits (no new cases resulting from their methodology are investigated).

7.4. Important observations and method comparisons

The previous analyses of Sections 7.2 and 7.3 clearly demonstrate the success of AD methods in a VAT fraud detection setting. This observation confirms the finding of Assylbekov et al. [4]. Tax authorities worldwide can adopt our proposal and can hereby overcome the VAT domain specific problems outlined in Section 1. As explained in Section 2.1, supervised classification methods have difficulties with some of these issues. Our methodology therefore contributes towards an efficient auditing strategy that should translate into an improved retrieval of tax losses and an enhanced deterrence.

Even though we have provided a number of reasons for the existence of sectoral differences in Section 1, we have not yet analysed the impact thereof on the obtained results. Figures 5, 6 and 7 clearly indicate that lifts and HRs can differ markedly across sectors. Also, the optimal method is sector dependent (e.g. in sector S-D, the LOF method is superior, whereas in sector S-I the BL or FWknn methods are recommended). This confirms that VAT fraud detection should be framed as a contextual AD problem (see Section 3) and this motivates an individual sector analysis. As all previous data mining based studies in the VAT domain (see Table 1) develop/evaluate their models based on instances pertaining to several sectors, this implies that methods were recommended that are suboptimal for certain activity codes under consideration.

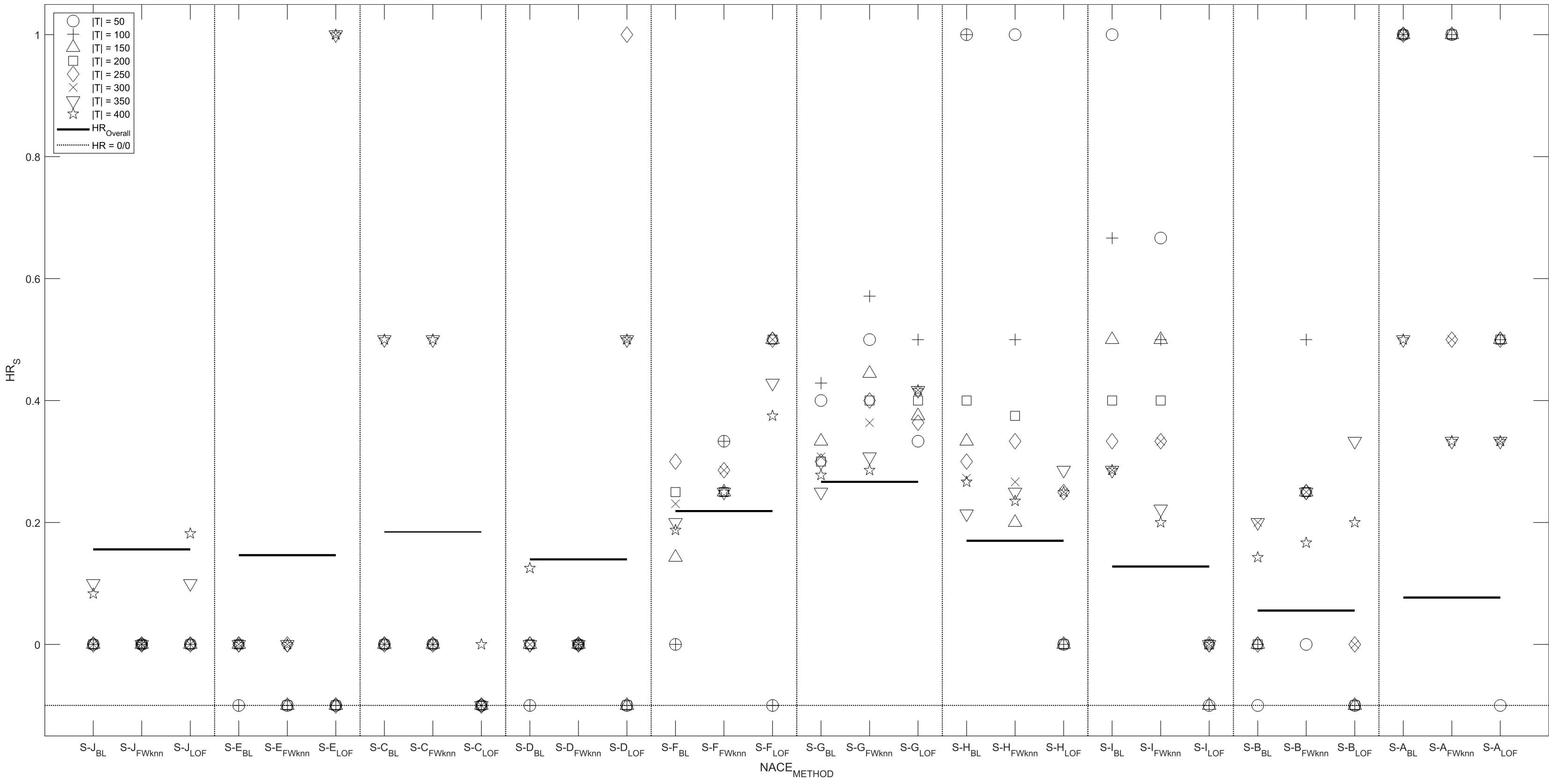

Figure 6: HR_S at different $|T|$ values. $HR_{Overall}$ represents the overall hit rate $|F|/|S|$. The horizontal dashed line points to undefined hit rates with numerator and denominator both zero. Vertical lines separate different sectors.

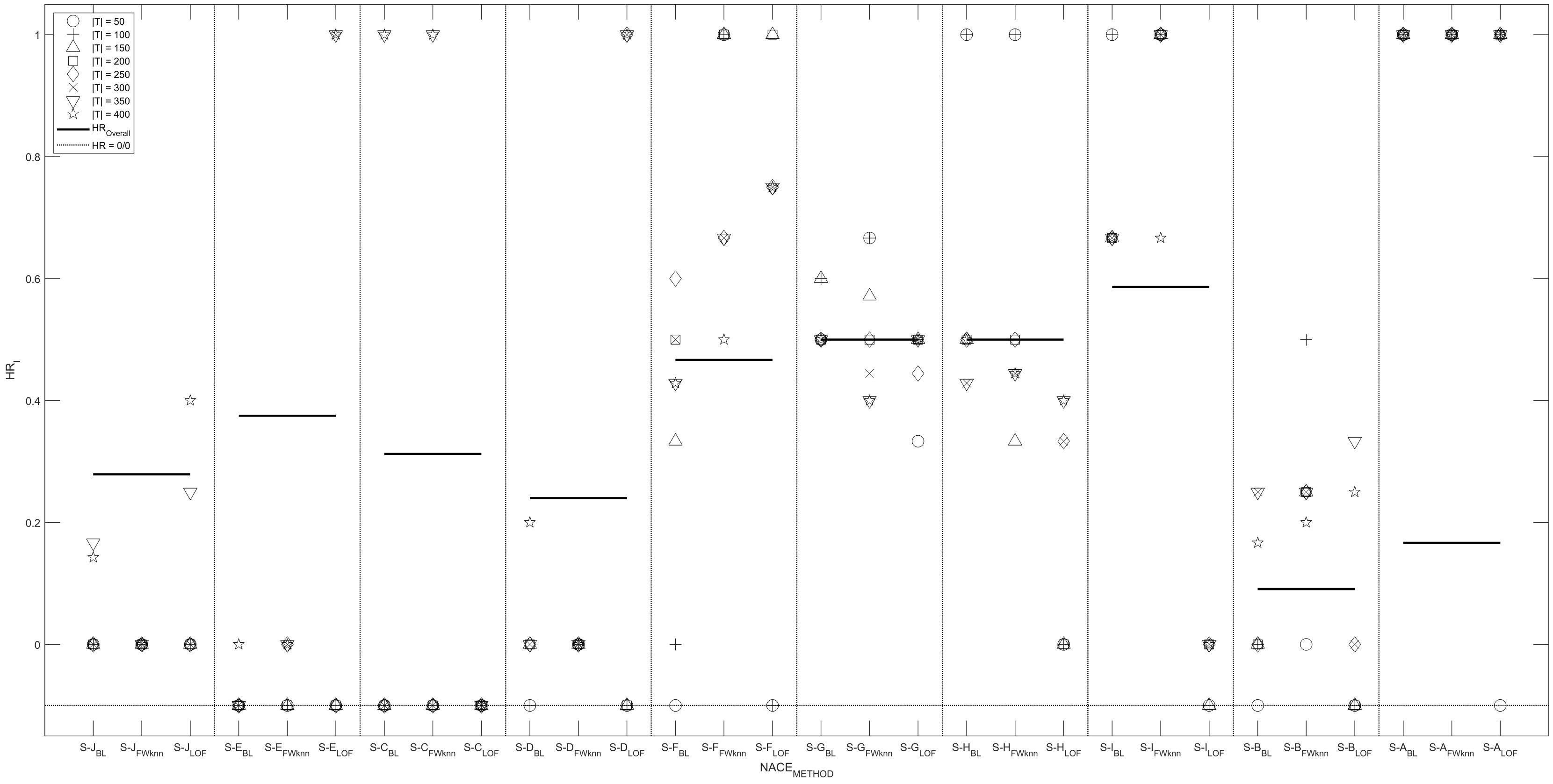

Figure 7: HR_I at different $|T|$ values. $HR_{Overall}$ represents the overall hit rate $|F|/|I|$. The horizontal dashed line points to undefined hit rates with numerator and denominator both zero. Vertical lines separate different sectors.

The lift and HR analyses (see Sections 7.2 and 7.3) demonstrate that the BL method is successful in detecting VAT fraud. This implies that a simple combination of the proposed tax ratios is valuable from a predictive point of view. Note that we have not shown the predictive value of each tax ratio individually. In fact, previous experiments suggested the use of a single fraud indicator to be inferior to the combination of tax ratios, because a single variable only provides a limited view on a company's tax declaration behaviour. Furthermore, for a particular firm that is being flagged, the most important preprocessed tax ratios (those with the largest contribution to the BL outlier score) are the ones having the largest absolute values. These variables can form the starting point of a tax audit. We have disclosed the 5 most important variables to the Belgian tax administration for each company appearing in the top list. We found that a different subset of tax ratios may be of relevance for different companies (a company can be flagged owing to a different type of suspiciousness). We consider such instance dependent explanations to be complementary to the global explanations[33] that are commonly reported in related studies [9, 35, 5, 13].

The feedback results of Table 5 clearly indicate that many companies appearing in our top list have not yet been signalled or audited by the Belgian tax administration. Hence the top list of the population differs markedly from the top list limited to the set of audited instances. The latter approach, which is often adopted in other VAT fraud detection studies (as discussed in Sections 2.1 and 3), provides an unrealistic indication of the predictive performance of a detection method because, in practice, tax administrations would examine a top list obtained by applying the method to the population. Furthermore, our evaluation takes into account capacity constraints faced by tax administrations and this aspect has been commonly ignored in previous studies as well (as indicated in Section 2.1).

When comparing the BL and FWknn approaches, we noted many identical companies appearing in the top lists. This translates to similar results with respect to the lift and HRs. Hence anomalies detected by the FWknn approach seem to conform[34] to instances with large absolute values on the majority of preprocessed tax ratios. This observation confirms the finding of Castellón González & Velásquez [13]. They also note fraud cases to occur for extreme values of variables. We recommend the use of the BL method over the FWknn technique because it is more intuitive, more comprehensible and has a much lower computational cost. Note that these aspects have also been ignored in other AD studies in the VAT domain [4].

As introduced in Section 2.2, the difference between global and local AD methods is important [75, 60, 12]. The distinction is based on different assumptions regarding the nature of outliers that are flagged. Which type of method is more appropriate depends on the dataset (sector) characteristics. Indeed, the local LOF method outperforms the global BL and FWknn approaches in terms of lift and HR in the sectors S-J, S-E

[33] Decision trees are popular global explanation techniques, which deliver a global ranking of variables. These are the most important attributes in general (when considering the entire collection of instances), yet those are not instance dependent.

[34] In general, an AD method retrieves instances residing in low-density areas of the (true yet unknown) probability density function. This does not necessarily translate to the values occurring at the extremes of a variable's range (i.e. if a variable has a range $[a, b]$, then an anomaly does not necessarily occur near a or b).

and S-D, yet for the sectors S-C, S-H, S-I and S-B, the LOF method is inferior. Similar performances are observed in the sectors S-F, S-G and S-A. Hence the complementarity of the local LOF and global FWknn/BL techniques seems to be confirmed. We recommend that the practitioner adopts the BL approach initially and if this approach delivers unsatisfactory results, one can switch to a local method such as LOF. The latter has a large computational complexity and this may motivate the use of a FWLOF method for large datasets (in Section 5.2, we have indicated why we have chosen not to include this version).

7.5. Timing-wise comparison

Table 6 presents computational timings of the three proposed detection methods when applying them to the total collection of around 950.000 Belgian companies. Each firm is again represented by the 18 preprocessed tax ratios presented in Section 4.2. Hyperparameter settings for the FWknn and LOF methods are described in Section 6. Regarding FWknn, we also examine the effect of increasing the compression level to $W_p = 30$. We have conducted 5 trials and reported average timings for the FWknn and BL method. Because LOF is very slow (and deterministic), we only report the result of a single run.

It is clear that the BL approach is very fast and can compute outlier scores for the entire population in less than 1 second. The FWknn method with $W_p = 1$ is still fast and requires 2 hours and 13 minutes. As indicated in Section 5.1, computational timings can be reduced by increasing the compression level (at the possible expense of a drop in predictive performance). Indeed, with $W_p = 30$ and the same settings for the other hyperparameters, FWknn requires around 17 minutes. The maximal time requirement is determined by the application at hand and the practitioner can impose a compression level that matches this requirement [68]. LOF is the slowest method, due to its quadratic complexity in terms of the number of instances, and requires 20 hours and 50 minutes (around 10 times slower than FWknn with $W_p = 1$). In Section 1, we mentioned that VAT databases can reach sizes of 5,5 million firms. A simple extrapolation would imply a runtime of around 28,5 days for LOF and this seems clearly infeasible.

Table 6: Computational timings of the BL, FWknn and LOF methods on a dataset of around 950.000 Belgian companies represented by 18 variables. See Section 6 for a description of the adopted hyperparameter values. FWknn uses $W_p = 1$ and FWknn ($W_p = 30$) uses $W_p = 30$. Best performance is indicated in boldface.

	BL	FWknn	FWknn ($W_p = 30$)	LOF
Time(s)	**0,069**	7985,6	1045,0	75.014,4

8. Conclusions and future research directions

The VAT domain characteristics of having very few labelled data (cases with fraud/legal indications) that are not representative for the population and the phenomenon of concept drift motivates our choice for AD techniques which have been rarely investigated. Our methodology consists of applying these methods to all companies pertaining to the same sector and that are represented by a set of tax ratios that are constructed based on domain knowledge. The success thereof is demonstrated by the high lift and hit rates achieved within most sectors under consideration. Our methodology should therefore enable an efficient auditing strategy by decreasing fraud losses and enhancing deterrence. Furthermore, it does not depend on prior time consuming audits and can be adopted by tax administrations worldwide.

To the best of our knowledge, we are the first VAT fraud detection study that examines the effect of sectoral differences that occur because of varying market conditions and tax law rules that apply in different sectors. We frame the problem as a contextual AD problem, where the context corresponds to the sector of the company under consideration. We have shown that the predictive performance and the optimal method are sector dependent. Hence the choice of fraud detection method should be determined by the characteristics of the sector and this aspect has been ignored in previous studies.

We have developed new fraud indicators in the form of tax ratios that are based on specific fields of the VAT return form (and client listings data). Other studies in the VAT domain lack in this direction. The proposed baseline method returns a fraud score by summing the absolute values of the (preprocessed) tax ratios. The BL method was shown to be successful in detecting VAT fraud and this implies that a simple combination of the proposed tax ratios is valuable from a predictive point of view. The new BL algorithm also provides an instance level explanation of why a particular firm is flagged for fraud and such an indication can form the starting point of a tax audit. We believe that we are the first study to provide such instance level explanations in the tax fraud detection domain.

Large batch datasets can be encountered in the tax domain which necessitates the use of fast algorithms. However, this aspect is commonly ignored in the tax fraud detection community. We have developed the BL method and used the FWAD sampling algorithm as fraud detection techniques. We have theoretically and empirically shown that these algorithms are scalable. Furthermore, the FWAD method enables the control of the predictive performance versus time requirement trade-off by modifying the compression level.

Most studies in the tax domain report performances based entirely on the labelled sample. As this set is not representative for the population, these results do not generalize beyond the sample. We have assessed performances by examining the appearance of (known) fraud cases in a top list (containing the most suspicious cases) constructed for the entire population. This evaluation methodology matches the practical application of fraud detection methods and provides a better indication regarding the nature of instances that are flagged. Furthermore, to the best of our knowledge, we are the first AD based study that provides a guarantee that fraud cases are effectively detected. Also note that we have taken capacity constraints faced by tax authorities into account and this issue has been ignored in many previous studies. In short, we have presented a new evaluation methodology to reliably assess the fraud detection performance of AD methods.

Next, we present a number of future research directions. Extensions of the current case study include the integration of other data sources (e.g. other tax types such as income tax, the released financial statements, commodity usage via electricity and telephone bills), preferably spanning a period of several years (to include growth features in the model). It is important to assess the predictive value of each data source individually and when combining them [53]. Also, the client listings data are currently used to obtain aggregated information (total sales, total purchases). However, this approach loses the fine-grained information available at identifier level. Therefore, it would be interesting to discover anomalies in the supplier/client network through graph based AD techniques [3] and assess the complementarity with our results.

Tax experts can benefit from indications of why a model has flagged a particular company for fraud because such explanations can form the starting point of a tax audit [30]. Even though the proposed BL method provides such indications, many AD techniques haven't got this explanatory power [5]. Additional research is therefore required on the design of a comprehensible AD system [43]. Aggarwal [1] presents a survey of techniques in this direction, the subspace AD methods that look for outliers in the subspace of the most important features for a certain instance, though he notes these approaches to be computationally intensive and they may not scale with the large data sizes that can be encountered in the tax domain.

An inspection strategy based solely on supervised classification models is inappropriate as these kind of methods are limited to the detection of known fraud patterns [8]. Indeed, in a dynamic environment, this model can become obsolete after a certain period of time. However, supervised techniques typically outperform unsupervised methods on the *known* set of fraudsters. Ideally, an inspection strategy should be developed that integrates classification with AD (and possibly other techniques such as random targeting and non data-driven approaches). In other words, how much 'weight' should be given to each detection method to result in an optimal auditing strategy in the short and long term. However, the development of such a framework is still an open issue.

Acknowledgements

We would like to thank the Belgian tax administration (Federal Public Service Finance, the Special Tax Inspectorate) for the provision of the dataset and their involvement throughout the project. Note that the models and tax ratios described in this paper are not necessarily the ones used by the tax administration.

This work was supported by a BOF-DOCPRO research project of the University of Antwerp, Belgium.

References

[1] Aggarwal, C. C. (2017). High-dimensional outlier detection: The subspace method. In *Outlier Analysis* (pp. 149–184). Cham: Springer International Publishing. doi:10.1007/978-3-319-47578-3_5.

[2] Agyemang, M., Barker, K., & Alhajj, R. (2006). A comprehensive survey of numeric and symbolic outlier mining techniques. *Intell. Data Anal.*, *10*, 521–538. URL: http://dl.acm.org/citation.cfm?id=1609942.1609946.

[3] Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, *29*, 626–688. URL: https://doi.org/10.1007/s10618-014-0365-y. doi:10.1007/s10618-014-0365-y.

[4] Assylbekov, Z., Melnykov, I., Bekishev, R., Baltabayeva, A., Bissengaliyeva, D., & Mamlin, E. (2016). Detecting value-added tax evasion by business entities of kazakhstan. In I. Czarnowski, A. M. Caballero, R. J. Howlett, & L. C. Jain (Eds.), *Intelligent Decision Technologies 2016: Proceedings of the 8th KES International Conference on Intelligent Decision Technologies (KES-IDT 2016) – Part I* (pp. 37–49). Cham: Springer International Publishing. URL: https://doi.org/10.1007/978-3-319-39630-9_4. doi:10.1007/978-3-319-39630-9_4.

[5] Basta, S., Fassetti, F., Guarascio, M., Manco, G., Giannotti, F., Pedreschi, D., Spinsanti, L., Papi, G., & Pisani, S. (2009). High quality true-positive prediction for fiscal fraud detection. In *2009 IEEE International Conference on Data Mining Workshops* (pp. 7–12). doi:10.1109/ICDMW.2009.59.

[6] Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, *3*, 27–38.

[7] Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. In *Proc. Credit Scoring and Credit Control VII* (pp. 235–255).

[8] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, *17*, 235–255. URL: https://doi.org/10.1214/ss/1042727940. doi:10.1214/ss/1042727940.

[9] Bonchi, F., Giannotti, F., Mainetto, G., & Pedreschi, D. (1999). Using data mining techniques in fiscal fraud detection. In M. Mohania, & A. M. Tjoa (Eds.), *DataWarehousing and Knowledge Discovery: First International Conference, DaWaK'99 Florence, Italy, August 30 – September 1, 1999 Proceedings* (pp. 369–376). Berlin, Heidelberg: Springer Berlin Heidelberg. URL: https://doi.org/10.1007/3-540-48298-9_39. doi:10.1007/3-540-48298-9_39.

[10] Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, *49*, 31:1–31:50. URL: http://doi.acm.org/10.1145/2907070. doi:10.1145/2907070.

[11] Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). Lof: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* SIGMOD '00 (pp. 93–104). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/342009.335388. doi:10.1145/342009.335388.

[12] Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenková, B., Schubert, E., Assent, I., & Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, *30*, 891–927. URL: https://doi.org/10.1007/s10618-015-0444-8. doi:10.1007/s10618-015-0444-8.

[13] Castellón González, P., & Velásquez, J. D. (2013). Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Syst. Appl.*, *40*, 1427–1436. URL: http://dx.doi.org/10.1016/j.eswa.2012.08.051. doi:10.1016/j.eswa.2012.08.051.

[14] Chan, P. K., Mahoney, M. V., & Arshad, M. H. (2003). A machine learning approach to anomaly detection. Technical report CS-2003-06, Department of Computer Science, Florida Institute of Technology Melbourne.

[15] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.*, *41*, 15:1–15:58. URL: http://doi.acm.org/10.1145/1541880.1541882. doi:10.1145/1541880.1541882.

[16] Clémençon, S., & Thomas, A. (2018). Mass volume curves and anomaly ranking. *Electron. J. Statist.*, *12*, 2806–2872. URL: https://doi.org/10.1214/18-EJS1474. doi:10.1214/18-EJS1474.

[17] Couturier, J., Peeters, B., & Van De Velde, E. (2017). *Belgisch belastingrecht in hoofdlijnen* volume 22. Antwerpen-Apeldoorn: Maklu.

[18] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

[19] Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms. *Pattern Recogn.*, *74*, 406–421. URL: https://doi.org/10.1016/j.patcog.2017.09.037. doi:10.1016/j.patcog.2017.09.037.

[20] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection. In D. Barbará, & S. Jajodia (Eds.), *Applications of Data Mining in Computer Security* (pp. 77–101). Boston, MA: Springer US. URL: https://doi.org/10.1007/978-1-4615-0953-0_4. doi:10.1007/978-1-4615-0953-0_4.

[21] European Commission (2012). An action plan to strengthen the fight against tax fraud and tax evasion. https://ec.europa.eu/taxation_customs/sites/taxation/files/resources/documents/taxation/tax_fraud_evasion/com_2012_722_en.pdf. Accessed on 28.08.2018.

[22] European Commission (2017). Fight against tax fraud and tax evasion: a huge problem. https://ec.europa.eu/taxation_customs/fight-against-tax-fraud-tax-evasion/a-huge-problem_en. Accessed on 05.10.2017.

[23] European Commission (2017). Impact assessment. accompanying the document proposal for a council directive. amending directive 2011/16/eu as regards mandatory automatic exchange of information in the field of taxation in relation to reportable cross-border arrangements. https://ec.europa.eu/taxation_customs/sites/taxation/files/impact-assessment-2017.pdf. Accessed 05.10.2017.

[24] European Commission (2017). Report from the commission to the council and the european parliament. Eighth report under article 12 of regulation (EEC, Euratom) nr 1553/89 on VAT collection and control procedures. https://ec.europa.eu/taxation_customs/sites/taxation/files/2017_report_vat_collection_control_procedures_en.pdf. Accessed on 29.12.2018.

[25] European Council (2010). Council regulation (EU) no 904/2010 of 7 October 2010 on administrative cooperation and combating fraud in the field of value added tax. https://eur-lex.europa.eu/eli/reg/2010/904/oj. Accessed on 28.05.2015.

[26] European Council (2011). Council directive 2011/16/EU of 15 February 2011 on administrative cooperation in the field of taxation and repealing directive 77/799/eec. https://eur-lex.europa.eu/eli/dir/2011/16/oj. Accessed on 28.05.2015.

[27] Eurostat (2016). Tax revenue statistics. http://ec.europa.eu/eurostat/statistics-explained/index.php/Tax_revenue_statistics. Accessed on 05.10.2017.

[28] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861 – 874. doi:https://doi.org/10.1016/j.patrec.2005.10.010.

[29] Federal Public Service Finance (2016). Periodic VAT declaration form (document 625). https://finances.belgium.be/sites/default/files/downloads/165-625-formulaire-2016.pdf. Accessed on 05.10.2017.

[30] Junqué de Fortuny, E., Stankova, M., Moeyersoms, J., Minnaert, B., Provost, F., & Martens, D. (2014). Corporate residence fraud detection. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '14 (pp. 1650–1659). New York, NY, USA: ACM. doi:10.1145/2623330.2623333.

[31] Gama, J. a., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Comput. Surv.*, *46*, 44:1–44:37. URL: http://doi.acm.org/10.1145/2523813. doi:10.1145/2523813.

[32] Ghafoori, Z., Rajasegarar, S., Erfani, S. M., Karunasekera, S., & Leckie, C. A. (2016). Unsupervised parameter estimation for one-class support vector machines. In J. Bailey, L. Khan, T. Washio, G. Dobbie, J. Z. Huang, & R. Wang (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 183–195). Cham: Springer International Publishing.

[33] Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLOS ONE*, *11*, 1–31. URL: https://doi.org/10.1371/journal.pone.0152173. doi:10.1371/journal.pone.0152173.

[34] Govers, M., & Deschacht, H. (2017). *Btw-praktijkboek 2017*. Mechelen, Belgium: Wolters Kluwer.

[35] Gupta, M., & Nagadevara, V. (2007). Autit selection strategy for improvig tax compliance - application of data mining techniques. In *Foundations of Risk-Based Audits. Proceedings of the eleventh International Conference on e-Governance, Hyderabad, India, December* (pp. 28–30).

[36] Han, C.-R., & Ireland, R. (2014). Performance measurement of the kcs customs selectivity system. *Risk Management*, *16*, 25–43. doi:10.1057/rm.2014.2.

[37] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1263–1284. doi:10.1109/TKDE.2008.239.

[38] Karlos, S., Fazakis, N., Kotsiantis, S., & Sgarbas, K. (2016). Semi-supervised forecasting of fraudulent financial statements. In *Proceedings of the 20th Pan-Hellenic Conference on Informatics* PCI '16 (pp. 34:1–34:6). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/3003733.3003740. doi:10.1145/3003733.3003740.

[39] Kollios, G., Gunopulos, D., Koudas, N., & Berchtold, S. (2003). Efficient biased sampling for approximate clustering and outlier detection in large data sets. *IEEE Transactions on Knowledge and Data Engineering*, *15*, 1170–1187. doi:10.1109/TKDE.2003.1232271.

[40] Krempl, G., & Hofer, V. (2011). Classification in presence of drift and latency. In *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 596–603). doi:10.1109/ICDMW.2011.47.

[41] Kumar, A., & Nagadevara, V. (2006). Development of hybrid classification methodology for mining skewed data sets - a case study of indian customs data. In *IEEE International Conference on Computer Systems and Applications, 2006.* (pp. 584–591). doi:10.1109/AICCSA.2006.205149.

[42] Marques, H. O., Campello, R. J. G. B., Zimek, A., & Sander, J. (2015). On the internal evaluation of unsupervised outlier detection. In *Proceedings of the 27th International Conference on Scientific and Statistical Database Management* SSDBM '15 (pp. 7:1–7:12). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2791347.2791352. doi:10.1145/2791347.2791352.

[43] Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. *MIS Quarterly*, *38*, 73–100.

[44] Matos, T., de Macedo, J. A. F., & Monteiro, J. M. (2014). An empirical method for discovering tax fraudsters: A real case study of brazilian fiscal evasion. In *Proceedings of the 19th International Database Engineering Nr 38, Applications Symposium* IDEAS '15 (pp. 41–48). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2790755.2790759. doi:10.1145/2790755.2790759.

[45] Mehta, P., Mathews, J., Kasi Visweswara Rao, S. V., Kumar, K. S., Suryamukhi, K., & Babu, C. S. (2019). Identifying malicious dealers in goods and services tax. In *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)* (pp. 312–316). doi:10.1109/ICBDA.2019.8713211.

[46] Mittal, S., Reich, O., & Mahajan, A. (2018). Who is bogus?: Using one-sided labels to identify fraudulent firms from tax returns. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies* COMPASS '18 (pp. 24:1–24:11). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/3209811.3209824. doi:10.1145/3209811.3209824.

[47] Ngai, E., Hu, Y., Wong, Y., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*, 559 – 569. URL: http://www.sciencedirect.com/science/article/pii/S0167923610001302. doi:https://doi.org/10.1016/j.dss.2010.08.006. On quantitative methods for detection of financial fraud.

[48] OECD (2016). *Revenue Statistics 2016*. OECD Publishing. URL: http://www.oecd-ilibrary.org/taxation/revenue-statistics-2016_rev_stats-2016-en-fr. doi:http://dx.doi.org/10.1787/rev_stats-2016-en-fr.

[49] Phua, C., Lee, V. C. S., Smith-Miles, K., & Gayler, R. W. (2010). A comprehensive survey of data mining-based fraud detection research. *CoRR*, *abs/1009.6119*. URL: http://arxiv.org/abs/1009.6119.

[50] Portnoy, L., Eskin, E., & Stolfo, S. (2001). Intrusion detection with unlabeled data using clustering. In *In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001* (pp. 5–8).

[51] Pozzolo, A. D., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, *29*, 3784–3797. doi:10.1109/TNNLS.2017.2736643.

[52] Pozzolo, A. D., Caelen, O., Borgne, Y.-A. L., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card

fraud detection from a practitioner perspective. *Expert Systems with Applications*, *41*, 4915 – 4928. URL: http://www.sciencedirect.com/science/article/pii/S095741741400089X. doi:https://doi.org/10.1016/j.eswa.2014.02.026.

[53] Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

[54] PWC (2013). Study on the feasibility and impact of a common EU standard VAT return. Specific contract No. 9, TAXUD/2011/DE/329. Final report (27 february 2013). https://ec.europa.eu/taxation_customs/business/vat/vat-reports-published_en. Accessed on 20.08.2019.

[55] Rad, H. A., Arash, S., Rahbar, F., Rahmani, R., Heshmati, Z., & Fard, M. M. (2015). A novel unsupervised classification method for customs fraud detection. *Indian Journal of Science and Technology*, *8*. URL: http://www.indjst.org/index.php/indjst/article/view/87306.

[56] Ravisankar, P., Ravi, V., Raghava Rao, G., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decis. Support Syst.*, *50*, 491–500. doi:10.1016/j.dss.2010.11.006.

[57] Roussinov, D. G., & Chen, H. (1998). A scalable self-organizing map algorithm for textual classification: A neural network approach to thesaurus generation. *Communication Cognition and Artificial Intelligence, Springer*, *15*, 81–112.

[58] de Roux, D., Perez, B., Moreno, A., Villamil, M. d. P., & Figueroa, C. (2018). Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery &#38; Data Mining* KDD '18 (pp. 215–222). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/3219819.3219878. doi:10.1145/3219819.3219878.

[59] Schneider, F. (2015). Size and development of the shadow economy of 31 european and 5 other OECD countries from 2003 to 2014: Different developments. *Journal of Self-Governance & Management Economics*, *3*, 7 – 29. URL: http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,url,uid&db=bth&AN=110610310&site=eds-live.

[60] Schubert, E., Zimek, A., & Kriegel, H.-P. (2014). Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, *28*, 190–237. URL: https://doi.org/10.1007/s10618-012-0300-z. doi:10.1007/s10618-012-0300-z.

[61] Shao, H., Zhao, H., & Chang, G.-R. (2002). Applying data mining to detect fraud behavior in customs declaration. In *Proceedings. International Conference on Machine Learning and Cybernetics* (pp. 1241–1244 vol.3). volume 3. doi:10.1109/ICMLC.2002.1167400.

[62] Song, J., Takakura, H., Okabe, Y., & Nakao, K. (2013). Toward a more practical unsupervised anomaly detection system. *Inf. Sci.*, *231*, 4–14. URL: https://doi.org/10.1016/j.ins.2011.08.011. doi:10.1016/j.ins.2011.08.011.

[63] Tang, M., Mendis, B. S. U., Murray, D. W., Hu, Y., & Sutinen, A. (2011). Unsupervised fraud detection in medicare australia. In *Proceedings of the Ninth Australasian Data Mining Conference - Volume 121* AusDM '11 (pp. 103–110). Darlinghurst, Australia, Australia: Australian Computer Society, Inc. URL: http://dl.acm.org/citation.cfm?id=2483628.2483641.

[64] TheWorldBank (2018). Listed domestic companies, total. World Federation of Exchanges database. https://data.worldbank.org/indicator/CM.MKT.LDOM.NO?end=2017&start=1975&view=map. Accessed on 15.08.2018.

[65] Thomas, A., Clémençon, S., Feuillard, V., & Alexandre, G. (2016). learning hyperparameters for unsupervised anomaly detection. URL: https://sites.google.com/site/icmlworkshoponanomalydetection/accepted-papers presented at ICML2016 Anomaly Detection Workshop, New York, NY, USA.

[66] Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, *75*, 38 – 48. URL: http://www.sciencedirect.com/science/article/pii/S0167923615000846. doi:https://doi.org/10.1016/j.dss.2015.04.013.

[67] Vanhoeyveld, J., & Martens, D. (2018). Imbalanced classification in sparse and large behaviour datasets. *Data Mining and*

Knowledge Discovery, *32*, 25–82. URL: https://doi.org/10.1007/s10618-017-0517-y. doi:10.1007/s10618-017-0517-y.

[68] Vanhoeyveld, J., & Martens, D. (2018). Towards a scalable anomaly detection with pseudo-optimal hyperparameters. https://hdl.handle.net/10067/1546510151162165141. Research paper, University of Antwerp, Faculty of Business and Economics.

[69] Vanhoeyveld, J., Martens, D., & Peeters, B. (2016). Datamining voor fraudedetectie. In A. Verhage, A. Jorissen, R. Prins, & J. Jaspers (Eds.), *Criminele organisaties en organisatiecriminaliteit* (pp. 167–212). Antwerp,Belgium: Maklu-Uitgevers nv. URL: http://www.maklu-online.eu/nl/tijdschrift/cahiers-politiestudies/jaargang-2016/39-criminele-organisaties-en-organisatiecriminalit/.

[70] Verstraeten, G., & Van den Poel, D. (2005). The impact of sample bias on consumer credit scoring performance and profitability. *Journal of the Operational Research Society*, *56*, 981–992. URL: https://doi.org/10.1057/palgrave.jors.2601920. doi:10.1057/palgrave.jors.2601920.

[71] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers 'I&' Security*, *57*, 47 – 66. URL: http://www.sciencedirect.com/science/article/pii/S0167404815001261. doi:https://doi.org/10.1016/j.cose.2015.09.005.

[72] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*, 80–83. URL: http://www.jstor.org/stable/3001968.

[73] Wu, M., & Jermaine, C. (2006). Outlier detection by sampling with accuracy guarantees. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '06 (pp. 767–772). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1150402.1150501. doi:10.1145/1150402.1150501.

[74] Wu, R.-S., Ou, C., Lin, H.-y., Chang, S.-I., & Yen, D. C. (2012). Using data mining technique to enhance tax evasion detection performance. *Expert Syst. Appl.*, *39*, 8769–8777. URL: http://dx.doi.org/10.1016/j.eswa.2012.01.204. doi:10.1016/j.eswa.2012.01.204.

[75] Zimek, A., Gaudet, M., Campello, R. J., & Sander, J. (2013). Subsampling for efficient and effective unsupervised outlier detection ensembles. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD '13 (pp. 428–436). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2487575.2487676. doi:10.1145/2487575.2487676.