# Bayesian Networks on Income Tax Audit Selection - A Case Study of Brazilian Tax Administration

**Leon Sólon da Silva** [*]
Secretariat of Federal Revenue of Brazil
Universidade de Brasília
leon.silva@rfb.gov.br

**Henrique de C. Rigitano** [†]
Secretariat of Federal Revenue of Brazil
henrique.rigitano@rfb.gov.br

**Rommel N. Carvalho**
Brazil's Office of the Comptroller General [‡]
Universidade de Brasília [§]
rommel.carvalho@cgu.gov.br

**João Carlos F. Souza** [¶]
Universidade de Brasília
jocafs@unb.br

## Abstract

Tax administrations in most countries have more corporate and personal information than any other government office. Data mining techniques can be used in many different problems due to the large amount of tax returns received every year. In the present work we show an essay of the Brazilian Tax Administration on using Bayesian networks to predict taxpayers behavior based on historical analysis of income tax compliance. More specifically, we tried to improve a previous risk based audit selection which detects a large amount of taxpayers as high risk. However, in its current form it identifies much more cases than the tax auditors can handle. Our first results are promising, considerably improving tax audit performance.

## 1 INTRODUCTION

Tax administrations have more information on people and companies than any other government office. Tax returns, bank transactions, and invoices arrive as hundreds of millions of records every year. The Secretariat of Federal Revenue of Brazil (RFB) is the Brazilian Tax Administration and Brazilian Customs as well. This combination is a major leverage and also a challenge.

Basically, there are two types of taxes: sales taxes and income taxes. Sales taxes includes value-added taxes and they are based on the value of the product being sold. Income tax is based on how much a person or a company earns. In most countries, sales taxes amount are considerably larger than income taxes (OECD, 2013). In Brazil, corporate and personal income taxes are about 50% of the country's revenue (RFB, 2016). Although corporate tax has much greater impact on final numbers, personal income tax audits affects a considerably large share of the Brazilian citizens. There are 27 million individual taxpayers in Brazil, about 13% of the population (RFB, 2016).

In order to facilitate and prioritize tax audits on personal income tax, RFB created the concept of a "fiscal lattice". One can understand the fiscal lattice as a first audit selection based on historical risk analysis of tax compliance by taxpayers. This lattice is a complex process in which many tax auditors specialized in personal income tax frauds create risk based rules for audit selection. The main difference between a regular audit and fiscal lattice audit is that the former has a much simpler process of analysis in order to determine whether to punish a taxpayer or not.

Since the number of taxpayers has increased, and the ratio between tax auditors and citizens has been reducing (RFB, 2016), the number of income taxpayers caught on fiscal lattice has increased as well. From 2010 to 2014, the taxpayers selected for this kind of audit highly increased (RFB, 2016). This changing scenario is pushing the tax administration to a limit of the tax auditors capacity of analysis. RFB's major office, has about 10,000 tax auditors and a huge backlog of fiscal lattice audits to analyze.

Data mining techniques can help better selecting taxpayers for audit and the present work offers one solution to improve the selection of this kind of audits. In Section 2.1 we discuss how Bayesian networks can be used as a classification algorithm in order to create predictive models.

The document is organized as follows: Section 2 describes some background information about Bayesian

[*] Anexo Ministério da Defesa, 5o andar Brasília, DF, Brazil
[†] Av. Rogerio Weber, 1752 - Centro, Porto Velho, RO, Brazil
[‡] SAS, Quadra 01, Bloco A, Edificio Darcy Ribeiro Brasilia, DF, Brazil
[§] Campus Darcy Ribeiro Brasilia, DF, Brazil
[¶] Campus Darcy Ribeiro Brasilia, DF, Brazil

networks; Section 3 details the solution for the tax audit selection problem, from its methodology to our first results; Section 4 presents the conclusion and future work.

## 2 BACKGROUND

In this section we bring some tax administration concepts, formulate the problem assessed by the present work, and discuss Bayesian networks for prediction.

### 2.1 BAYESIAN NETWORKS FOR PREDICTIVE MODELS

As stated by (Korb and Nicholson, 2010) Bayesian networks (BNs) are graphical models for reasoning under uncertainty, where the nodes represent variables (discrete or continuous) and arcs represent direct connections between them. These direct connections are often causal connections. In addition, BNs model the quantitative strength of the connections between variables, allowing probabilistic beliefs about them to be updated automatically as new information becomes available.

Bayesian networks are useful to learn from data and discover causalities between variables and it can be used as a classifier algorithm. It is being used for prediction in many different problems, from genetics (Jansen et al., 2003) and prognostics of breast cancer (Gevaert et al., 2006), to identification of split purchases (Carvalho et al., 2014). In the present work, we use Bayesian networks as a solution for predicting a taxpayer to be compliant or non-compliant in terms of tax obligations. In more detail, our approach presents an improvement of tax audit selection using Bayesian networks to build predictive models. In the next section we present the details for the solution to our problem, as well as the first results.

The next subsections describe two different types of Bayesian networks, Naïve Bayes and Tree-Augmented Naïve Bayes.

#### 2.1.1 Naïve Bayes

Naïve Bayes is the most simple version of Bayesian network. It uses strong connections between the nodes and it considers all explanatory variables (nodes) as independent. Despite its simpleness it has many applications with good results and great run performance as stated in (Zhang, 2004).

#### 2.1.2 Tree-Augmented Naïve Bayes

Tree-Augmented Naïve Bayes (TAN), as explained in (Zheng and Webb, 2011), relaxes the assumption of complete independence of the explanatory variables by enforcing a tree structure. In this case, each explanatory variable only depends on the class and one other variable. This relaxation allows the representation of more complex models, leading to possible performance improvements, as shown in (Carvalho et al., 2014).

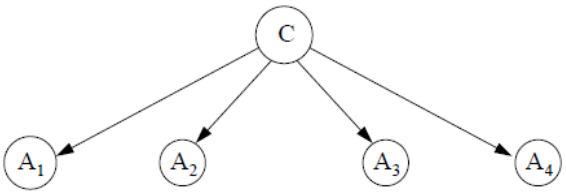


Figure 1: Example of Naïve Bayes Network (Zhang, 2004)

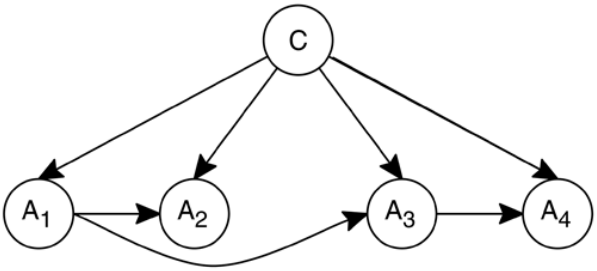


Figure 2: Example of Tree-Augmented Naïve Bayes Network (Jiang et al., 2009)

### 2.2 RELATED WORK

As stated in (Silva et al., 2015) many tax administrations have been using data mining techniques to create predictive models for tax compliance risk. Despite being a topic of great interest, tax administrations have many concerns in publishing internal projects. Since taxpayer information is classified and should be protected by tax officers, many of them do not share the details of tax compliance risk projects.

A source of such information, case studies, methodolo-

gies, and best practices are intergovernmental organizations. For tax administrations and customs the World Customs Organization (WCO) and the Organization for Economic Cooperation and Development (OECD) are important sources. In a recent survey that gathered many countries, OECD presented a comparative chart that shows the use of data mining to detect tax fraud (OECD, 2013).

Tax Administrations internal publications also present many studies that can be applied by other countries and many of them have developed methodologies based on statistical analysis and data mining to create tax compliance risk systems. Most countries use data mining for taxpayers classification considering its risks of non-compliance.

Some studies, however, reveal different data analysis approach being held in tax administration. The US Internal Revenue Service (IRS) uses data mining for different purposes, according to (Castellón González and Velásquez, 2013), among which are tax compliance risk based taxpayer classification, tax fraud detection, tax refund fraud, criminal activities, and money laundering (Watkins et al., 2003).

Another related reference is Jani Martikainens master thesis (Martikainen et al., 2012). He presents results of studies conducted by the Australian Taxation Office (ATO) concerning the usage of models to detect high-risk tax refund claims. Also according to the author, the ATO avoided the payment of refunds of about US$ 665,000,000.00 between 2010 and 2011 based on data mining tools. ATO uses refund models based on social networking discovery algorithms that detect connections between individuals, companies, partnerships, or tax returns. The models are updated and refined to enhance detection and increase the recognition of new fraud (Martikainen et al., 2012).

More related to the present work Gupta *et al.* in (Gupta and Nagadevara, 2007) describes in details different approaches on using data mining techniques to improve tax audit selection. The main difference is that in (Gupta and Nagadevara, 2007) the main taxes are value-added taxes in contrast with income taxes, object of the present research. Also in (Kirkos et al., 2007) data mining is used to detect frauds on financial statements, which can be easily customized to tax returns and tax evasion/fraud.

# 3 SOLUTION AND FIRST RESULTS

In this section we describe the methodology used in the present work and detail each step of the data analysis from the information and data gathering to the construction of predictive models for improvement of tax audit selection.

## 3.1 METHODOLOGY

The methodology of the present work follows the well-known CRISP-DM (CRISP-DM). The Cross Industry Standard Process for Data Mining is a technology-independent methodology and reference model to implement data mining process in every business. It describes each phase every data mining work should pass. Each phase is equally relevant for the success of the data analysis process and should not be underestimated. The process has six phases and it is possible to perform the same step more than once. The phases of CRISP-DM are (Wirth and Hipp, 2000):

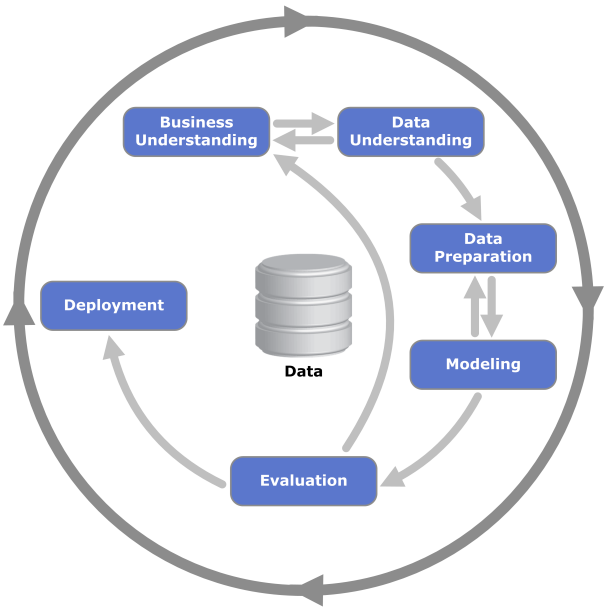


Figure 3: CRISP-DM Reference Model (Wirth and Hipp, 2000)

Business Understanding

Every data analysis process is designed to answer business questions to achieve business goals. In the business understanding phase of CRISP-DM these questions are asked and possible solutions are also proposed. Possible quantitative and qualitative business process' improvements are also detailed, in order to justify the use of data mining techniques to solve business problems.

According to (Chapman et al., 2000), this initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives.

Data Understanding

Once the business questions are clear, it is time to understand the required information to perform the changes needed in the business process and achieve the goals identified in the previous phase. In data understanding, all sources of information needed to perform the analysis are determined. The first insights and main patterns are also identified in the first contact with the data available from the possible sources. Each business question needs to be mapped to every data source (systems, databases, webpages, etc.) in order to address every goal and identify possible gaps and lack of information.

In (Wirth and Hipp, 2000) it is stated that there is a close link between business understanding and data understanding. The formulation of the data mining problem and the project plan require at least some understanding of the available data.

Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools.

Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques require specific data formats. There is a close link between data preparation and modeling. Often, one realizes data problems while modeling or one gets ideas for constructing new data.

Evaluation

At this stage in the project you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment

Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.

### 3.2 BUSINESS UNDERSTANDING

Our main goal is to improve individuals tax audit selection. We try to achieve a better audit process performance by better using the tax auditors knowledge and time available to perform these audits. As in any tax administration, there are far more taxpayers returns and information to analyze than tax officers, and to achieve the revenue goals and tax fairness it is major that the selection of audit is as risk based as possible.

In Brazil, personal taxpayers pay their income taxes every month. Since the tax is calculated on a year based, by April of the next year, taxpayers are obliged to send their income tax return in order to adjust their debt (or credit). Every year, tens of million of returns are sent to RFB, much more than it could handle if there were no risk based selection.

RFB created the concept of “fiscal lattice” to select personal income tax returns based on tax compliance risk. In this technique personal income tax fraud experts analyze the historical of all taxpayers and their previous knowledge in order to come up with parameters to select the tax returns for audit. Once caught on “fiscal lattice”, only a tax officer could release the tax return, preventing fraudsters from receiving a possible credit. There are three main purposes in using this technique:

- to better select taxpayers based on tax compliance risk;
- to facilitate the verification of tax auditors, since each parameter has a well defined analysis and treatment activities;
- to ease the auto-correction of tax returns by taxpayers, since many of them were caught due to filling

errors.

Besides all Brazilian tax administration efforts to select the individuals tax audits, the number of audits selected by fiscal lattice has increased from 569,000 in 2011 to 937,000 in 2014[1] in contrast with the number of tax officers, that decreased from 12,273 in 2010 to 10,419 in 2015 (RFB, 2016).

More specifically, we intend to use data mining techniques to discharge as many taxpayers as possible of fiscal lattice, with the minimum compliance risk to tax administration. With thousands of audits already finalized by experienced tax auditors, it is possible to assess this problem with machine learning tools and achieve best results in letting go those taxpayers that offer less risk of tax compliance.

In our first approach on trying data mining techniques to address the problem, we selected a certain RFB's unit that has been suffering from the large number of fiscal lattice audits. The "Delegacia Especial de Pessoa Fsica" (DERPF) or "Individual Taxpayers Special Office" is an individual taxpayer specialized unit located at Sao Paulo City, the Brazilian biggest city, in the most economically active federation unit (State of Sao Paulo). This unit has come to its limit of fiscal audits since its creation in 2014, and has the largest number of this kind of audits in the whole country. It was a natural choice for our first experiments.

### 3.3 DATA UNDERSTANDING AND PREPARATION

To answer the business question on how to improve the selection of individual taxpayers caught in fiscal lattice, we evaluated the sources of the information needed to perform the data mining analysis. Our sample was taken from audits performed by DERPF from years 2014 to 2016.

Basically, all individuals taxpayer information was taken from internal systems, from online systems to datamarts and datawarehouses. Most of taxpayer information caught in fiscal lattice is available from tax returns, but some information is taken from invoices and financial operations. The exact properties retrieved by the data extraction as well as the fraud/non-compliance rate are classified information.

The final taxpayer table has 25,322 taxpayer's returns analyzed by tax auditors and classified as compliant or non-compliant. Each line has, besides the dependent variable (compliant) other 20 characteristics of taxpayers and information retrieved from returns and other systems. From these, 13,547 are women and 10,730 are men. Other explanatory variables are information of tax return and unfortunately cannot be specified because it could present classified information, since the result of the analysis could lead taxpayers to learn fraud patterns and use that information to avoid being caught.

For preparation, all independent variables were analyzed in order to remove the incomplete rows and to discretize continuous ones to comply with the Bayesian network algorithms constraints. The numeric variables where classified within bands in terms of average multipliers (one average, half average, three times average, etc.). After data preparation the final number of individual taxpayers returns was 24,277.

All data preparation took place using R language[2] and its packages.

### 3.4 MODELING AND EVALUATION

We used $bnlearn$ R package[3] in order to run the Bayesian network algorithms. Specifically the functions $naive.bayes$ and $tree.bayes$ where chosen to create the predictive models. The first is the well-known Naïve Bayes algorithm, which does not take parameters for customizing the models and the former is an implementation of the Tree-Augmented Naïve Bayes (TAN) algorithm. The TAN algorithm takes white list (force the inclusion of arcs in Bayesian network), black list (force the exclusion of arcs in Bayesian network), and mi[4] parameters.

To create the predictive models we took the compliant variable as dependent and the other 35 (thirty five) information as independent variables. The sample of 24,277 where divided into training (80%) and test (20%). No validation sample was needed since we used 10-fold cross-validation technique with $bnlearn$'s function $bn.cv()$.

As stated in $bn.cv()$ documentation (CRAN, 2016) k-fold is a technique where the data is split in k subsets of equal size. For each subset in turn, bn is fitted (and possibly learned as well) on the other k - 1 subsets and the loss function is then computed using that subset. Loss estimates for each of the k subsets are then combined to give an overall loss for data.

[1] In 2015 this number decreased to 670,000 due to efforts in better selecting individuals tax returns for audits

[2] `https://www.r-project.org/`.

[3] `http://www.bnlearn.com/`.

[4] The estimator used for the mutual information coefficients for the Chow-Liu algorithm in TAN. Possible values are mi (discrete mutual information) and mi-g (Gaussian mutual information). We use discrete since all explanatory variables have been discretized

Since the proportion of compliant/non-compliant taxpayers is classified information, we present the results of the predictive models in terms of improvements from the actual process of discharging taxpayers from fiscal lattice. Since our dependent variable is compliant/non-compliant, we are interested in evaluating the models by specificity more than sensitivity, since it is more dangerous to let a non-compliant taxpayer go away without being audited than to select one that is compliant to be audited.

Each Brazilian tax administration local unit is autonomous and may choose whatever criteria it finds best to dismiss taxpayers from fiscal lattice. So, to a matter of possible comparison with our proposal, we consider a linear cut (random selection) of taxpayers until it reaches a units capacity. If, for example, an office has the capacity to audit 2,000 taxpayers per month, and there are 3,000, we consider the actual process to randomly choose the 1,000 to be dismissed. The overall taxpayers wrongly dismissed, is the same as the proportion between non-compliant taxpayers from overall caught on fiscal lattice. Our goal is to better predict if a taxpayer caught on fiscal lattice is compliant or not. If we come to a specificity considerably better than random selection, we achieve our goal to let go as few non-compliant taxpayers as possible.

As we learn from Table 1, using Naïve Bayes is already a good tool to select those taxpayers which can and cannot be dismissed from being audited. Tree-Augmented Naïve Bayes had no major advantages, despite the customization of parameters (root chose automatically or user defined).

Table 1: Predictive Models by Algorithm/Parameters

| Algorithm | Performance Rate |
|---|---|
| Naive Bayes | 41 % |
| TAN (auto root) | 34 % |
| TAN (selected root) | 35 % |

Therefore, the predictive models in this first results showed optimistic results, resulting in a increase of more then 30% in tax audit selection in comparison to randomly discharging taxpayers. It is major to recollect that the taxpayers caught in fiscal lattice have already been through a risk based process of selection and any increase in this criteria is a leverage in using Bayesian networks to build models of tax compliance.

# 4 CONCLUSION AND FUTURE WORK

Brazil has been through a major crisis and the responsibility of the RFB as a tax administration has also increased in order to guarantee the revenue for public policies. A better selection of tax audits save resources and increase the performance of the collecting tax process. Our approach on creating predictive models to improve the risk based selection of the so called "fiscal lattice" proved to be a promising one based on the first results.

We intend to use different approaches and Bayesian networks algorithms in order to create compliance risk scores and leave the decision of taxpayers being compliant or not to the tax officers and possibly increase the specificity. The approach in the present work delegates this decision to the prediction algorithm.

Furthermore we will try and build Bayesian networks with larger samples and more tax units and include more information about the taxpayer, since in this work we basically used income tax returns and registry information. Financial transactions and invoice data could be interesting explanatory variables and will be used in future applications.

**Acknowledgements**

The authors would like to thank RFB, specially DERPF, for providing the resources necessary to work in this research, as well as for allowing its publication.

## References

Rommel N Carvalho, Leonardo Sales, Henrique A Da Rocha, and Gilson Libório Mendes. Using bayesian networks to identify and prevent split purchases in brazil. In *BMA@ UAI*, pages 70–78, 2014.

Pamela Castellón González and Juan D Velásquez. Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications*, 40(5):1427–1436, 2013.

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. 2000.

CRAN. Cran project. package bnlearn. https://cran.r-project.org/web/packages/bnlearn/index.html, 2016. Accessed: 2016-05-08.

Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray

data with bayesian networks. *Bioinformatics*, 22(14): e184–e190, 2006.

Manish Gupta and Vishnuprasad Nagadevara. Audit selection strategy for improving 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*, pages 28–30, 2007.

Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F Greenblatt, and Mark Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, 2003.

Liangxiao Jiang, Harry Zhang, and Zhihua Cai. A novel bayes model: Hidden naive bayes. *Knowledge and Data Engineering, IEEE Transactions on*, 21(10): 1361–1371, 2009.

Efstathios Kirkos, Charalambos Spathis, and Yannis Manolopoulos. Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, 32(4):995–1003, 2007.

Kevin B Korb and Ann E Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.

Jani Martikainen et al. Data mining in tax administration-using analytics to enhance tax compliance. *Department of Information and Service Economy. Aalto University*, 2012.

OECD. Tax administration 2013 - comparative information on oecd and other advanced and emerging economies. Technical Report 2308-7331, Organisation for Economic Co-operation and Development, Paris, 2013. URL `http://www.oecd-ilibrary.org/content/serial/23077727`.

RFB. Secretariat of federal revenue of brazil (rfb) website. http://www.receita.fazenda.gov.br, 2016. Accessed: 2016-05-08.

Leon Sólon da Silva, Rommel Novaes Carvalho, and João Carlos Felix Souza. Predictive models on tax refund claims-essays of data mining in brazilian tax administration. In *Electronic Government and the Information Systems Perspective*, pages 220–228. Springer, 2015.

R CORY Watkins, K Michael Reynolds, Ron Demara, Michael Georgiopoulos, Avelino Gonzalez, and Ron Eaglin. Tracking dirty proceeds: exploring data mining technologies as tools to investigate money laundering. *Police Practice and Research*, 4(2):163–178, 2003.

Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39. Citeseer, 2000.

Harry Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004.

Fei Zheng and Geoffrey I Webb. Tree augmented naive bayes. In *Encyclopedia of Machine Learning*, pages 990–991. Springer, 2011.