National College of Ireland

Analysis of Brazilian deputies expenses claims from 2013 to 2016

MSc Research Project Data Analytics

William Fabiano Bueno
x15046231

School of Computing
National College of Ireland

Supervisor: Dr. Simon Caton

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	William Fabiano Bueno
Student ID:	x15046231
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Dr. Simon Caton
Submission Due Date:	11/12/2017
Project Title:	Analysis of Brazilian deputies expenses claims from 2013 to 2016
Word Count:	5488

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	10th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

1 Introduction **3**
1.1 Statement of the Research Problem . . . 4

2 Related Work **4**
2.1 Outliers Identification by machine learning . . . 4
2.1.1 Supervised Anomaly Detection . . . 5
2.1.2 Semi-supervised Anomaly Detection . . . 5
2.1.3 Unsupervised Anomaly Detection . . . 5
2.1.4 k-means clustering . . . 5
2.2 Forecast of time series data . . . 5
2.2.1 Forecasting with Holt-Winters Exponential Smoothing Methods . . . 6
2.2.2 Forecasting with Artificial Neural Networks . . . 6

3 Methodology **6**
3.1 Data Acquisition . . . 7
3.2 Data Preparation . . . 7
3.3 Evaluation strategy and expected performance . . . 7

4 Implementation **7**
4.1 k-means . . . 7
4.2 Time Series . . . 8
4.2.1 Holt-Winters . . . 8
4.2.2 Artificial Neural Network . . . 8

5 Evaluation **8**
5.1 k-means . . . 8
5.2 Time Series . . . 11
5.2.1 Holt-Winters . . . 11
5.2.2 Artificial Neural Network . . . 12
5.3 Discussion . . . 12

6 Conclusion and Future Work **13**

Analysis of Brazilian deputies expenses claims from 2013 to 2016

William Fabiano Bueno
x15046231
MSc in Data Analytics

11th December 2017

Abstract

After the Operation Car Wash, deflagrated by the federal police force in Brazil, politicians are on the spot of attention as consequence of the several corruption scandals becoming public. Applying machine learning techniques, this work analysed the Brazilian deputies claim expenses for the exercise of the parliamentary activity from a party perspective and identified unusual activity in regarding the claims from 2013 to 2016. It was also applied two different time series techniques, Holt Winters and an Artificial Neural Network, to forecast the claim expenses and propose a reduction in the deputies allowance.

1 Introduction

Financial fraud attracts attention and concern from both private and public sectors due to numerous cases of corporate fraud, money laundering, credit card fraud among other crimes. Financial fraud is a serious problem when considering the internal policies of corporations (Ngai et al.; 2011), accentuating the necessity of transparency and the annual balance report. The United Kingdoms Fraud Authority suggested that in 2016, the amount of losses due to fraud only in the United Kingdom were estimated in 193 billion, from which the private sector was responsible for 144 billion and the public sector was responsible for 37.5 billion [1].

Regarding strategies to attempt to minimise financial fraud, governments and corporations worldwide face challenges to overcome internal barriers, such as lack of alignment with the business strategy and not integrated legacy systems (Wong and Venkatraman; 2015). Despite the availability of new software in the market, several entities still control their employees expenses with improper tools, transforming the identification and explanations of abnormalities into a difficult assignment (Lecue and Wu; 2017). Furthermore, the task of identifying misconduct increases by the fact that often, financial managers and executives do not want the internal and external auditors to identify irregularities in their department (Fanning and Cogger; 1998). In this scenario, auditors are required either in private companies as well as in the public sector to monitor expenses, to minimize unnecessary expenditure and to assure that the financial laws/rules are being followed. Both private and public sectors, need effective auditing tactics and systems to reduce losses and fiscal fraud among their employees.

Audit in the public sector aims to assure the adequate and timely application of the several public resources, searching for the common good of society (JuND; 2006). In other words, the importance of public auditing rests on the reassurance that the public entities can accomplish their proposed obligations with transparency and at the same time, to achieve their duties ethically, economically and professionally (Goodson et al.; 2012). Like in the private sector, government auditing can be internal or external, with the later officially performed by a federal auditor. By offering impartial objective assessments of how the public resources have been managed, the objective of the federal auditor is to detect inconsistencies and misconduct, to safeguard the public interest. The auditing process, contributes to create conditions for public services to be satisfactorily provided and to reflect the expectation that public organizations and their employees will perform their functions efficiently, ethically and in accordance to the laws and regulations (Santos; 2015).

[1] https://www.gov.uk/government/organisations/national-fraud-authority

1.1 Statement of the Research Problem

Lately, more than 1000 politicians in Brazil have been accused of corruption and fraud in the context of a police investigation named Car Wash Operation. It is estimated by the Brazilian Federal Public Ministry that R$ 200 billion, approximately 52 billion (European central Bank, exchange rate 1 euro = 3.8564 reais on 27/Nov/2017) [2], are diverted yearly due to bribes that involve not only the government but also the private sector (Ribeiro et al.; 2016). Moreover, the country is facing an economic recession that lead to the approval of a bill that freezes the public expenditures for the next 20 years.

Brazilian democracy is formed by the executive power (1 president), legislative power (two chambers 81 senators and 513 federal deputies) and the judicative power (supreme court with 11 ministers) [3]. The lower chamber, which is the subject of this study, is constituted by 513 deputies representing 27 states or federation units. The number of deputies per state is determined by the size of the population in each state, where states with higher number of voters, elect more deputies when comparing to states less populated (table X available in the configuration manual). According to the Brazilian law, each deputy is granted with a monthly budget to spend on the exercise of parliamentary activity and related, such as:

- rent allowance of R$3,800.00 per month;
- health insurance no limit;
- printing of documents to show the parliamentary - R$20,000.00 per year;
- allocated funds to hire staff - R$936,000.00 yearly;

Allowance for the exercise of the parliamentary activity monthly, which there is a variance in the amount, depending on the state the deputy represents, which varies from R$ 30,788.66 to 45,612.53, available in the configuration manual [4].

In the context of the poor and fragile audit process that the Brazilian deputies allowance faces and since data analytics is a science specialized in detecting patterns and pointing data anomalies, this work aims to analyse the Brazilian deputies claim expenses made in the period of time from 2013 to 2016, using machine learning techniques, to (1) identify unusual activities and (2) predict the claims for year 2017, verifying the feasibility of a reduction in the total budget destined for the exercise of the parliamentary activity.

The research question proposed in this work is: in which extent is it possible to forecast the Brazilian deputies claim expenses for 2017 and identify anomalies in the fiscal years from 2013 to 2016?

The overview of this work comprises:

This work is organized as follows. In Section 2, the related work is explored, highlighting existing methods for outlier detection and forecasting expenses. The dataset, tools and methodology are described in Section 3. In Section 4 the implementation of the experiments is detailed, explaining the parameters used on the identification of outliers by k-means and on the development of an exponential smoothing method and Artificial Neural Network (ANN) models. Results are illustrated in Section 5, followed by a discussion on the comparison between the time series prediction methods in Section 6, exploring also the possibility of a reduction in the deputies allowance, presenting conclusions and remarks for future work.

2 Related Work

2.1 Outliers Identification by machine learning

In statistical language, unusual activity is known as outliers, exceptions or anomalies, and it is characterized as patterns found in data that do not present the expected behaviour. Anomalies can potentially bias or skew an analysis and its results. With the possibility to apply anomaly detection in a wide range of applications such as insurance, cyber-security intrusion detection, credit card, health care (Chandola et al.; 2009), Kumar (2005) discuss the possibility to identify a potential hacked computer and loss of

[2] https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/eurofxref-graph-brl.en.html

[3] http://www2.camara.leg.br/

[4] http://www2.camara.leg.br/camaranoticias/noticias/POLITICA/474313-CONHECA-O-VALOR-DO-SALARIO-DE-UM-DEPUTADO-E-DEMAIS-VERBAS-PARLAMENTARES.html

classified information to unauthorized personnel due to an abnormal traffic pattern in a computer network. Anomaly detection may also be used to verify the incidence of malignant tumours that could be showed by an anomalous magnetic resonance image (Spence et al.; 2001).

Machine learning methods with the purpose for anomaly detection, have been used mostly to identify outliers and remove them from the dataset, what is known as to clean the dataset (Goldstein and Uchida; 2016). Currently the outliers identification is used for different purposes such as fraud detection, where the outliers analysis could be used to point system misuse (Phua et al.; 2010), intrusion detection used to monitor servers and networks where outliers could be identified as intrusion attempts (Garcia-Teodoro et al.; 2009) or even for medical applications where the outliers detection can be used to analyse for example, medical images such as computed tomography to spot tumours or abnormal cells (Lin et al.; 2005).

The main goal of anomaly detection is to build up an anomaly detector model, usually dividing the dataset to create two different process: training and testing. The different anomaly detection setups are classified based on the labels available in the dataset. Supervised Anomaly Detection, Semi-supervised Anomaly Detection and Unsupervised Anomaly Detection are different approaches available for identifying anomalies, each of them offering several different algorithms (Ding et al.; 2014), (Goldstein and Uchida; 2016).

2.1.1 Supervised Anomaly Detection

In a supervised setup, all the data including the anomalies are identified and labelled upfront and the dataset must be divided into training and testing. With the dataset divided, an ordinary classifier can be trained and applied. Examples of data mining methods that uses this modality of training are Support Vector Machine (SVM) (Ji and Xing; 2017) and ANNs. In study to compare better machine learning methods for fraud detection, SVM outperformed other classifiers such as logistic regression, C5.0 DT and back-propagation neural network (Song et al.; 2014).

2.1.2 Semi-supervised Anomaly Detection

Semi-supervised setup also requires training and testing datasets but the training dataset must be cleaned and not present anomalies. The concept is that a model of the normal class is learned with the training dataset and it detects the anomalies deviating from the learned model (Goldstein and Uchida; 2016). Examples of algorithm that handles semi-supervised setup are SVM and association rules. Kim et al (2003) implemented an anomaly detection model based on association rules algorithm Apriori, designed to scale to a huge volume of data for fraud detection.

2.1.3 Unsupervised Anomaly Detection

This setup is the most flexible among the three, where no labels are required. It is not necessary to divide the dataset into training and testing. The algorithm calculates the distance to give an estimation on what is normal and what is unusual in the dataset. As an example of an unsupervised method k-means clustering, was applied in an insurance company data set to detect outliers by analysing the clustered based outliers (Thiprungsri and Vasarhelyi; 2011).

2.1.4 k-means clustering

k-means clustering is an unsupervised machine learning method that divide the data into clusters where the entities allocated in the same group (cluster) are more similar when compared to each other than when compared to those from other groups and this technique works by measuring squared Euclidean distance (Albashrawi and Lowell; 2016).

Since k-means can generate clusters with relatively uniform sizes and is less time consuming, Wu et. al (2007) used this clustering technique to identify rare classes in credit card fraud.

Virdhagriswaran and Dakin (2006), used k-means technique focused in account fraud where the objective was to identify public companies present in the United States Securities and Exchange Commission, that were potentially committing accounting fraud. .

2.2 Forecast of time series data

Using statistical and data mining models, it is possible to validate the possibility of forecasting claim expenses. By using past information to learn about variables, time series forecasting is known as a

popular technique for predictive analysis. It is mandatory that the data is ordered to apply time series technique and, even though most currently data is ordered by time, it is also a possibility to apply time series on data ordered by other factors, such as distance or height. Time series can be applied in several areas such as credit scoring, financial analysis, auditing and accounting, decision support and many others (Kotu and Deshpande; 2014), (Tkáč and Verner; 2016).

Considering non-series data, normally the aim is to comprehend the correlation between variables from one observation and the influence they may produce in the parameter that is object of the study (Jayasree and Balan; 2013). Nonetheless, with time-series data, it is important to comprehend the present value of the wanted variable with its preceding or forthcoming values, known as autocorrelation. Autocorrelation can indicate that previous data of a variable can correlate with the future data of the same variable (Kotu and Deshpande; 2014), (Mills; 2011).

Two different data series methods are available: model-driven and data-driven. In the method known as univariate or data-driven, since the target variable is also the predictor, the difference between target and predictor is not relevant. In other words, this method has only one variable. Examples of data-driven methods are Simple average and Naive Forecast. Model-driven or multivariate techniques require dependent and independent variables, where time must be the independent variable.

2.2.1 Forecasting with Holt-Winters Exponential Smoothing Methods

The exponential smoothing methods were created by Charles C. Holt (2004) and Robert Goodell Brown (1960) unaware of each other. Commonly used to remove noise from the data, the exponential smoothing methods deals with smoothing parameters that are determined according to the past data. Older observations get a smaller weight in forecasting then the most recent ones which get relatively more weight (Tratar and Strmčnik; 2016). There are two parameters that are assigned by the user, the first one is the weight that decreases as the observations are getting older and the second one is the initial value. For the initial value, the user may choose first observation as the value (ÇAPAR; 2015).

After applying time series methods to forecast the heat load in Slovenia, Tratar L. and Strmcnik E. (2016) identified that Multiple Linear Regressions offered better forecast results for short-term and Holt Winters offered better results for long-term. Kotsialos A. et al. (2005) applied Holt Winters and ANN methods for sales forecasting and their conclusion was that even though ANN-based forecast performed slightly better, it does not justify the application of a complex method for a slightly increased accuracy.

2.2.2 Forecasting with Artificial Neural Networks

ANN is a machine learning technique that imitates the natural capacity of the brain to gain and accumulate knowledge. It was described by Tkac, M. and Verner, R. (2016) as computational structures designed to emulate the accumulation of knowledge in the biological central nervous system. The ability of an ANN to learn can be pointed as the main differential when comparing it to other different machine learning methods. Due to its adaptability, efficiency, robustness and wide range of different ANN methods, this machine learning technique turn out to be wildly popular for credit scoring, classification, financial analysis and others (Tkáč and Verner; 2016).

Differently than other computational techniques, ANNs are capable to resolve nonlinear problems. By producing an ANN model to predict economic crisis, Chirita et al. (2012) affirmed that the ability to classify non-linearly data, makes ANN an exceptional option for solving economic world problems. Even though acquiring training data to train the model remains a challenge, as well as, design factors impacting the accuracy of the model, ANNs are still a reliable option for forecasting financial time series.

According to Hamid Shaikh A. and Habib A. (2014), there are two major limitations when considering ANNs. The first one is that the ANNs work as a black box, in other words, the steps which involves the connection weights cannot be explained in an equation and secondly, ANNs have a generic tendency to overfit the model.

3 Methodology

To address the research question of this work, it was applied the Cross Industry Standard Process for Data Mining (CRISP-DM) as general methodology (Paula et al.; 2016), which consists of six stages, described as follows:

- Business Understanding Acquiring knowledge of claim expenses to achieve the objectives;

- Data Understanding Getting the data from the Brazilian government web page;
- Data Preparation Cleaning the data, selecting the attributes, creating extra attributes, to prepare the dataset for the analysis;
- Modelling Selecting the methods to achieve the proposed objectives;
- Evaluation Comparing the model results from the previous step with the objectives;
- Deployment Applying the solutions to increase opportunities.

3.1 Data Acquisition

The data is available on the Brazilian official web page [5], from where four datasets were collected. The datasets comprise the yearly claimed expenses made by the deputies of the lower chamber from 2013 to 2016.

3.2 Data Preparation

Since Portuguese is a language with several types of accentuation in its grammar, Excel was used to handle the special characters issue as well as to remove blank instances, remove unnecessary attributes and creation of new attributes (detailed information available in the configuration manual). Three tables were created with the metadata for the id codes used to populate the attributes created for the time series method and k-means technique. To perform the k-means, it was necessary to hot encode these three new attributes due to the fact of k-means cannot handle categorical attributes. The experiments were carried out in R Studio, which is a tool capable to process large volumes of data in a faster manner and the packages necessary to perform the proposed analysis are mentioned in the configuration manual. Considering the large amount of observations, they were monthly aggregated for analysis.

3.3 Evaluation strategy and expected performance

To evaluate the accuracy of the forecasts, data from 2013 to 2015 was used as training data and data regarding 2016 was used as test data, allowing the proposed methods to be verified with actual data. Considering that it was used the same dataset for both proposed forecast methods, the Root Mean Square Error (RMSE), which represents the standard deviation of the differences when considering the observed values and the predicted values (Patrick et al.; 2000), was considered as one of the accuracy metrics. Mean Absolute Percentage Error (MAPE) which is commonly used in forecast methods is another evaluation measure used due to its intuitive and easy interpretation (Sagaert et al.; 2018). Bratu M. (2012) used Mean Absolute Error (MAE) to verify which of the efficiency of the time series models applied to forecast macroeconomics in Romania Should consider using just the 2 measures first mentioned. These were the used accuracy measures methods to evaluate this work.

4 Implementation

To analyze the Brazilian deputies claim expenses, it was used k-means to identify outliers and to forecast years 2016 and 2017 was used Holt-Winters and an ANN. All three methods were implemented with R.

4.1 *k*-means

k-means is a clustering technique that works by grouping observations with similarities, reducing the distance from observations to its respective cluster center by summing the squares (Ding et al.; 2014). k-means needs the user to provide the number of clusters desired for the algorithm, which for the present work was 5. The data was aggregated by claimed expense amount and month, in a way where the final dataset is visualised by parties, state and general description. For example, the dataset presents only one instance for Postal Services per month for each party, for each state and the same criteria for the remaining eleven description codes as illustrated by table 1. Similarly, it is necessary to set up the nstart, which determines at which observation to start the run, being 20 the number chosen for this work. The

[5] http://www2.camara.leg.br/transparencia/cota-para-exercicio-da-atividade-parlamentar/dados-abertos-cota-parlamentar

Table 1: Dataset attributes for analysis

State_code	Party_Code	General_Disc_Codes	Invoice_Month	Invoice_amount

outliers selected for this analysis were the further five observations from the cluster center of each year, from 2013 to 2016, totalizing 25 outliers.

4.2 Time Series

Figure 1 is the time series model for the Brazilian deputies claim expenses from January/2013 to December/2015, created in R to support the forecast methods described in the two subsequent parts of this section. This time series model works as training data for the forecast models. By observing the model, it can be stated that the data is seasonal.

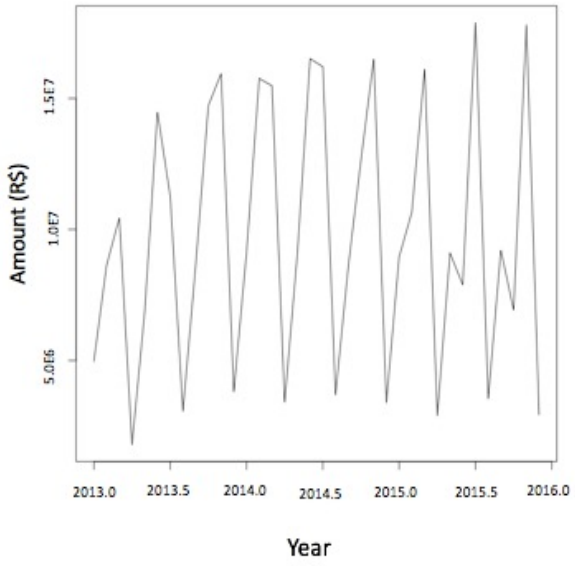

Figure 1. Time Series model

4.2.1 Holt-Winters

Widely applied for forecasting analysis, the exponential smoothing methods, are suitable for either seasonal or data with no trend. To run the proposed method, with the assistance of the forecast package in R, it was used the time series model, mentioned in the previous part, to fit the Holt Winters model in a forecast function to project a prediction for 24 months from January/2016 to December/2017.

4.2.2 Artificial Neural Network

Similarly to the Holt-Winters implementation method, by using R, the ANN was also feed with the time series model and set up to provide a prediction model from January/2016 to December/2017.

5 Evaluation

5.1 *k*-means

The clusters created by k-means are not going to be part of the discussion, since the objective is to get the outliers only. The k-means outputs analysis was divided in 5 parts, showing, firstly, the results for monthly aggregated datasets from 2013 to 2016 and, subsequently, the results year by year. Each output shows the total expenditure (Invoice Amount) in terms of three different categories to broaden the analysis: 1) the type of service corresponding to the invoice (General Description), 2) the party that generated the invoice (Party Code) and 3) the state that generated the invoice (State Code). The

squares in the graphs represent the clusters centres. The round dots represent the instances, the colours represent the different clusters and the crosses are marking the five outliers.

Considering the outputs for the aggregated period of 2013 2016 (Figure 2), it is noticeable that the majority of the outliers is spotted manly in one category (Figure 2-a), corresponding to divulgation of the parliamentary activity (code 4). Only one of the outliers correspond to a different category, aircraft rental (code 1). Figure 2-b demonstrates that the outliers are spread in three different parties, PODE (code 9) and PP (code 10) with one observation each and PTB (code 23) with three outliers, corresponding to four states (Figure 2-c), Para (code 14), Goias (code 9), Amazonas (code 4) and Sao Paulo (code 25).

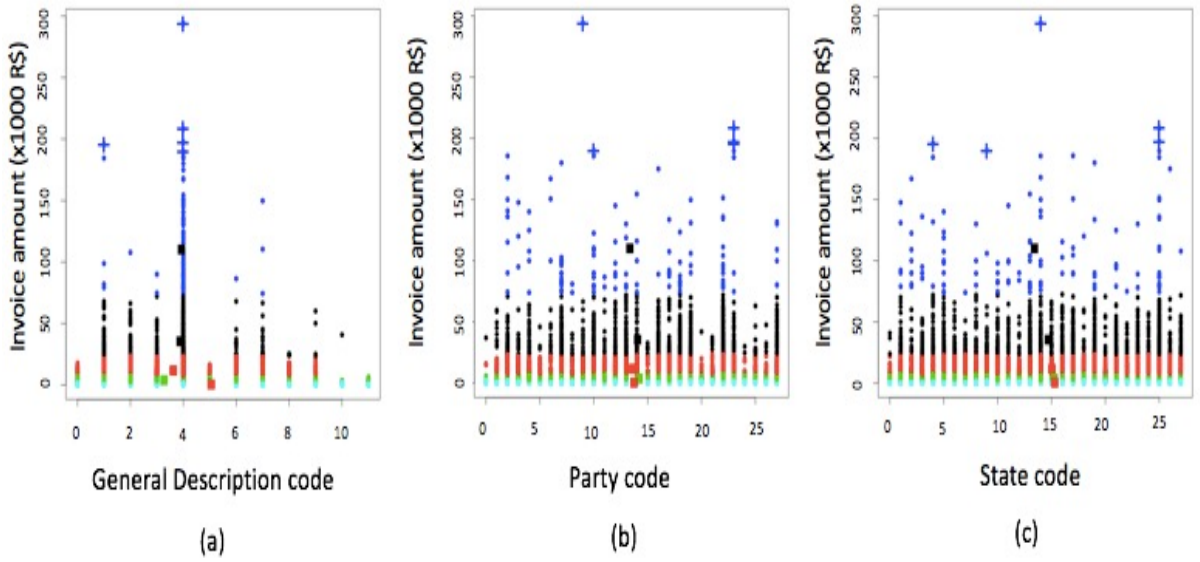

Figure 2. Total expenditure clustering output for the aggregated period of 2013-2016 in terms of (a) General description, (b) Party code and (c) State code.

Drilling down the dataset and analysing the outputs for 2013 (Figure 3), Figure 3-a points four outliers in the category of divulgation of the parliamentary activity (code 4) and one in the category of airline tickets (code 2). Regarding the parties, Figure 3-b points two outliers corresponding to PT (code 22), and one outlier each for AVANTE (code 2), PSD (code 18) and PTB (code 23), corresponding to three states (Figure 3-c), So Paulo (code 25), Para (code 14) and Minas Gerais (code 13).

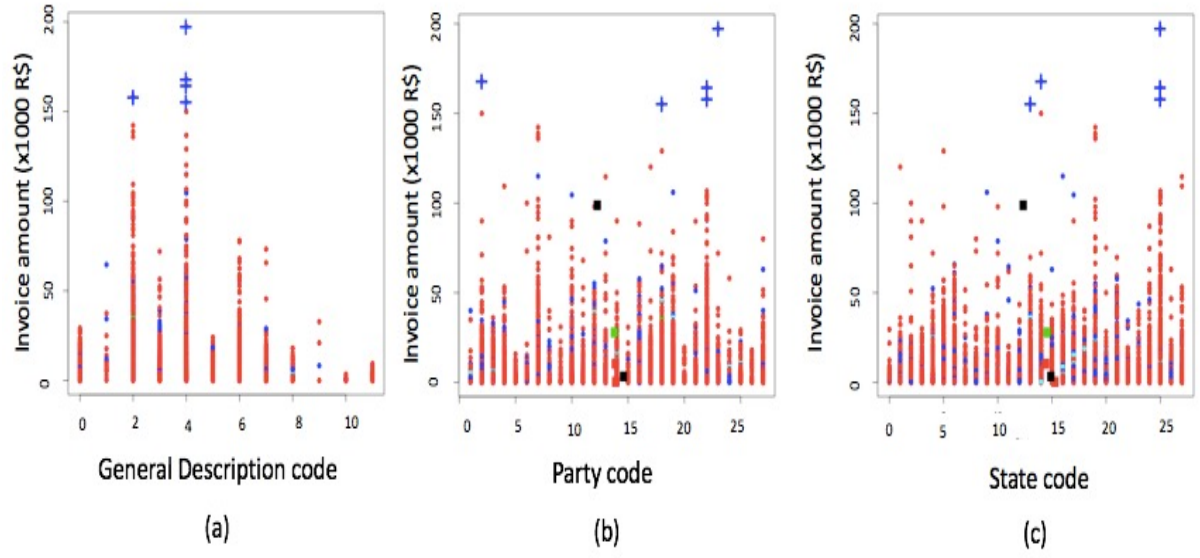

Figure 3. Total expenditure clustering output for year 2013 in terms of (a) General description, (b) Party code and (c) State code.

Regarding the expenditures of 2014, Figure 4-a shows that all outliers are in the same category of divulgation of the parliamentary activity (code 4), belonging to three parties (Figure 4-b), PSD (code 18) and PTB (code 23), with one observation each, and PT (code 22) with three observations. Figure 4-c shows the outliers are spread over four states, Mato Grosso do Sul (code 12), Minas Gerais (code 13) and Rio Grande do Sul (code 21) each with one outlier and So Paulo (code 25), presenting two outliers.

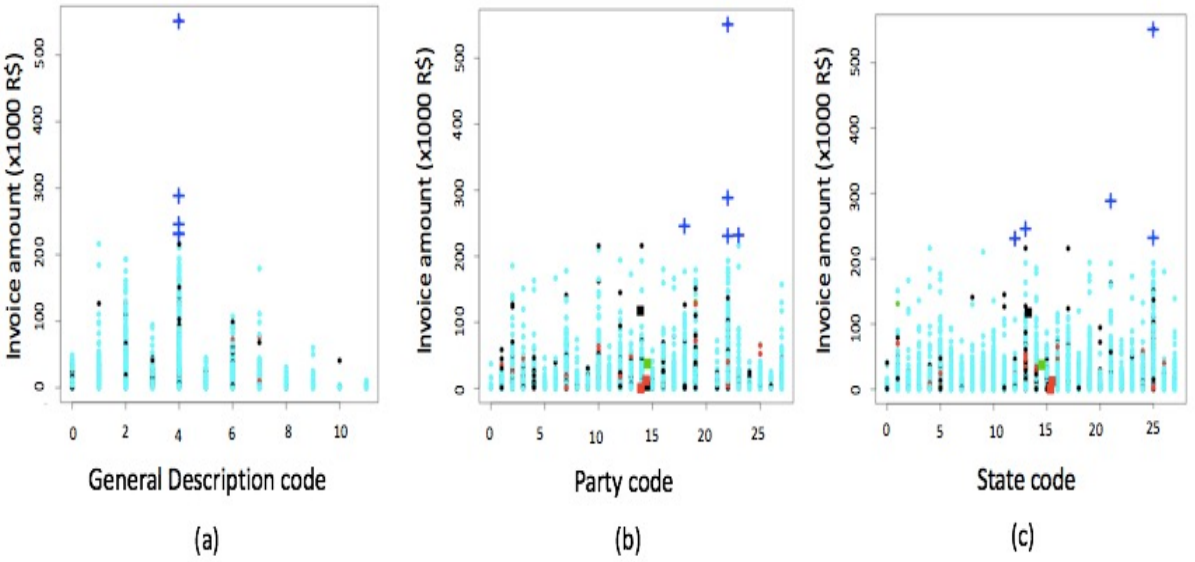

Figure 4. Total expenditure clustering output for year 2014 in terms of (a) General description, (b) Party code and (c) State code.

Considering the outputs for 2015 (Figure 5), all five observations correspond to the category of divulgation of the parliamentary activity (Figure 5-a). Figure 5-b illustrates that the outliers are spread in four different parties, where PODE (code 9), PSD (code 18) and PSDB (code 19) have one observation each and PMDB (code 7) has two, corresponding to four different states (Figure 5-c), Para (code 14), Parana (code 16), Rio de Janeiro (code 19) and Sao Paulo (code 25).

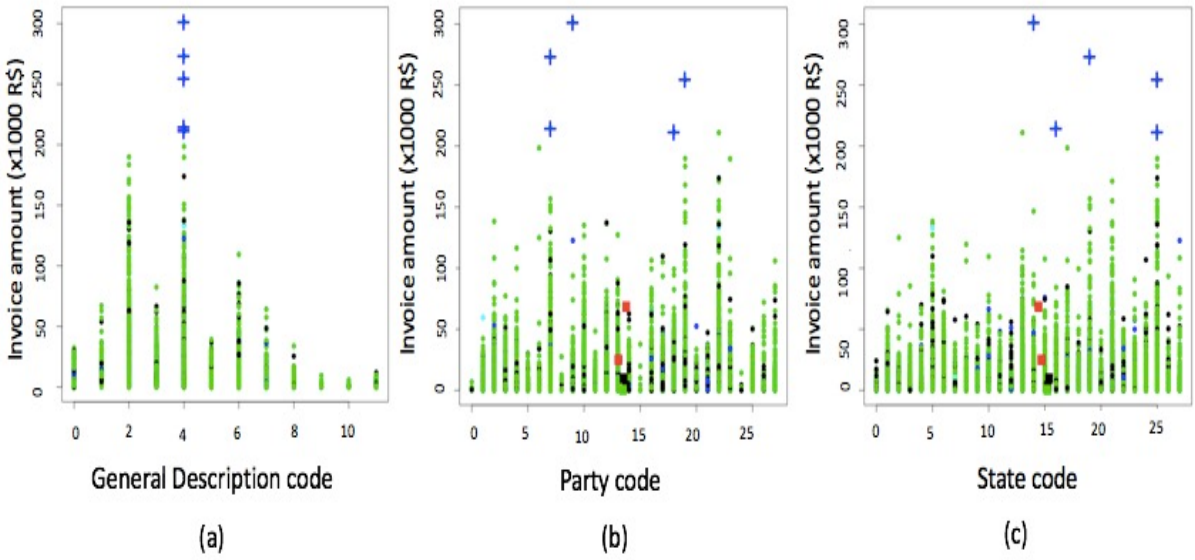

Figure 5. Total expenditure clustering output for year 2015 in terms of (a) General description, (b) Party code and (c) State code.

Finally, the outputs for 2016 (Figure 6) show five outliers falling into the category of divulgation of the parliamentary activity (code 4, Figure 6-a). Figure 6-b demonstrates that the outliers are spread

in five parties, PODE (code 9), PMDB (code 7), PSC (code 17), PT (code 22) and PTB (code 23), belonging to the states Rio de Janeiro (code 19), Sao Paulo (code 25), Sergipe (code 26) and Tocantins (code 27) as per Figure 6-c.

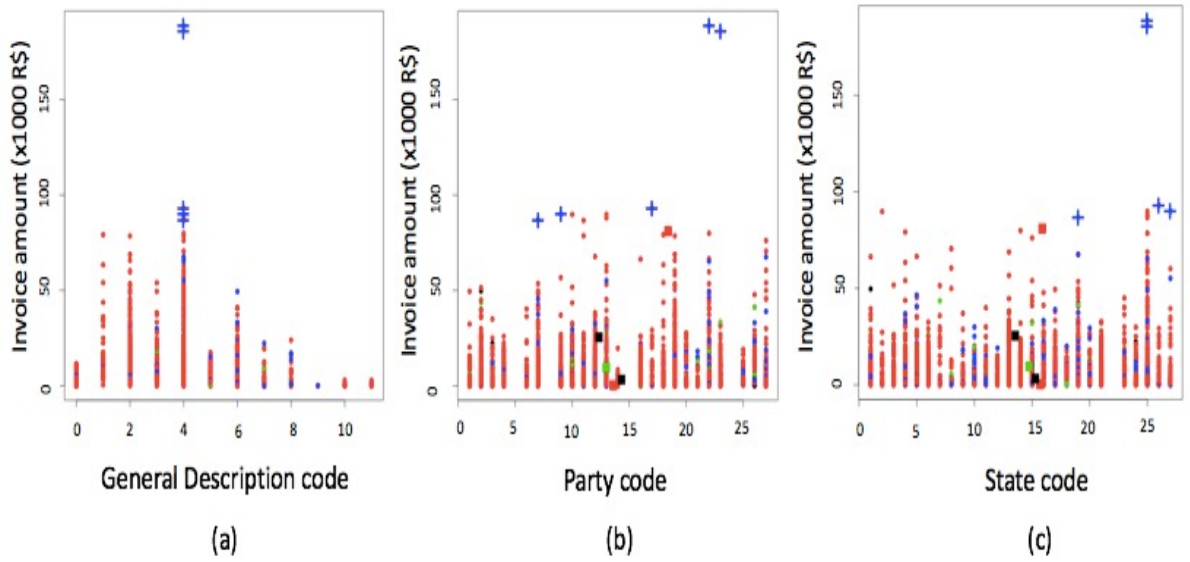

Figure 6. Total expenditure clustering output for year 2016 in terms of (a) General description, (b) Party code and (c) State code.

5.2 Time Series

5.2.1 Holt-Winters

As previously mentioned, a time series model from January/2013 to December/2015 was created, revealing seasonal observations (Figure X1), and used to fit the Holt Winters model. Figure 7 illustrates the result of the forecast fitted model for a period of 36 months. Holt Winters creates two distinct forecast confidence levels, of 80% and 95%, which are represented, in Figure 7, as the blue and grey area respectively. The blue line corresponds to the fitted model, giving a linear prediction of expenses for 2016 and 2017, around R$9.5 million.

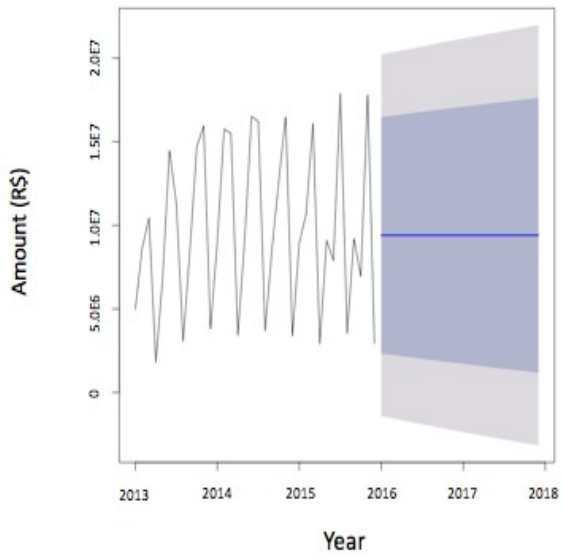

Figure 7. Holt-Winters forecast model

5.2.2 Artificial Neural Network

The same time series from January/2013 to December/2015 (Figure 1) was used to fit the ANN model and the model prediction. The results, shown in Figure 8, reveal a projection for the next 24 months that is also seasonal. The pattern of the fitted model, displayed as a blue line, is visually similar when compared to the previous three years. The model projects the highest pick for the beginning of 2017. Moreover, it appears that the claimed expenses show a decreasing pattern approaching the end of 2017. In sum, the model projected the smallest amount of R$2.5 million on April/2016 and the highest amount of R$21 million on February 2017.

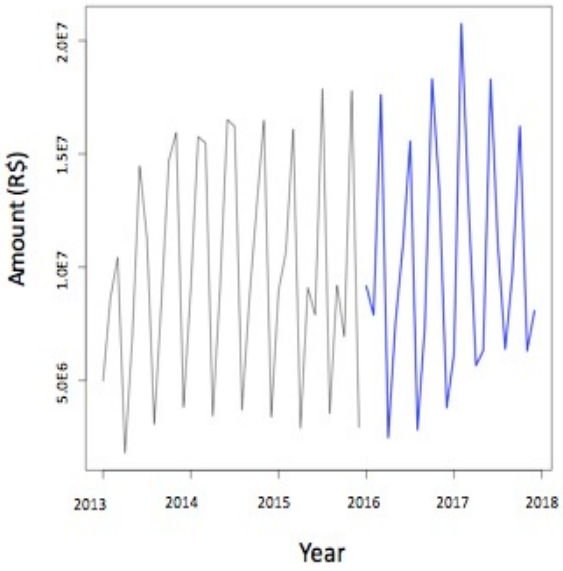

Figure 8. ANN forecast model

5.3 Discussion

Applying k-means in the claimed expenses dataset made possible to identify outliers that lie far from their cluster centres. Since this an unsupervised method was used, exclusively for outlier identification, i.e. there are no known outliers identified beforehand, it was not necessary to apply an evaluation method for such task. 25 different outliers were identified (5 for each analysis), throughout the analysis of the aggregated period of 2013 to 2016 as well as of each year individually. Interestingly, 23 of the 25 outliers fell into the category of divulgation of the parliamentary activity. Regarding the parties, of those that were more present among the outliers, PT and PTB stands out, with six observations each and outlier occurrence in all years from 2013 to 2016. PODE and PSD are the second ones presenting more outliers, with three observations each, while PMDB presented two outliers and AVANTE, PP, PRB and PSC were pointed out as outliers once. The analysis of the provenience of the outliers shows that Sao Paulo state concentrates the majority, from where twelve outliers belong to. In total, the outliers were spread into eleven different states, including also Amazonas, Goias, Mato Grosso do Sul, Parana, Rio de Janeiro, Rio Grande do Sul, Sergipe and Tocantins, each having one observation. Minas Gerais and Para, in turn, presented two and three observations respectively. Verifying the outliers closely, it was checked the maximum amount allowed per party in each state, considering the number of deputies representing the corresponding party. Among the 5 outliers pointed in 2013, parties PTB and AVANTE from Sao Paulo and Para respectively, exceeded the monthly amount for over R$120,000.00 each. The 2014 verification showed that parties PTB from Sao Paulo and PT from Mato Grosso do Sul, exceeded their month allowance in around R$150,000.00. In 2015 only the party PODE, from Para, spent over R$259,000.00 more than the allowed. Finally, in 2016, PTB from Sao Paulo and PSC from Sergipe over spent R$110.000,00 and R$53,000.00 respectively.

This work aimed to forecast Brazilian deputies expenses for years 2016 and 2017, comparing two different methods, Holt-Winters and ANN. As mentioned in the previous section, Holt-Winters resulted in a linear trend (Figure 9-a), while ANN resulted in a seasonal projection (Figure 9-b). Figure 9 displays a comparison for each fitted model between the forecasted results (blue line) with the actual data (red line) for 2016. Visually, Holt-Winters accommodates the actual data in the forecasted range area, with its

Table 2: Accuracy results

Method	RMSE	MAE	MAPE
Holt Winters	5533808	4771188	76.09234
Neural Network	47496.29	33117.88	0.5989224

most part falls in the blue area, except for the beginning of 2016, when it reaches the 95% of confidence level projection. In turn, the ANN forecasted model estimated values varying with the actual claimed by the deputies during 2016. Even tough demonstrating a pattern closer (visually) to the train data, in 2016 the expenditures were mostly lower than the predicted.

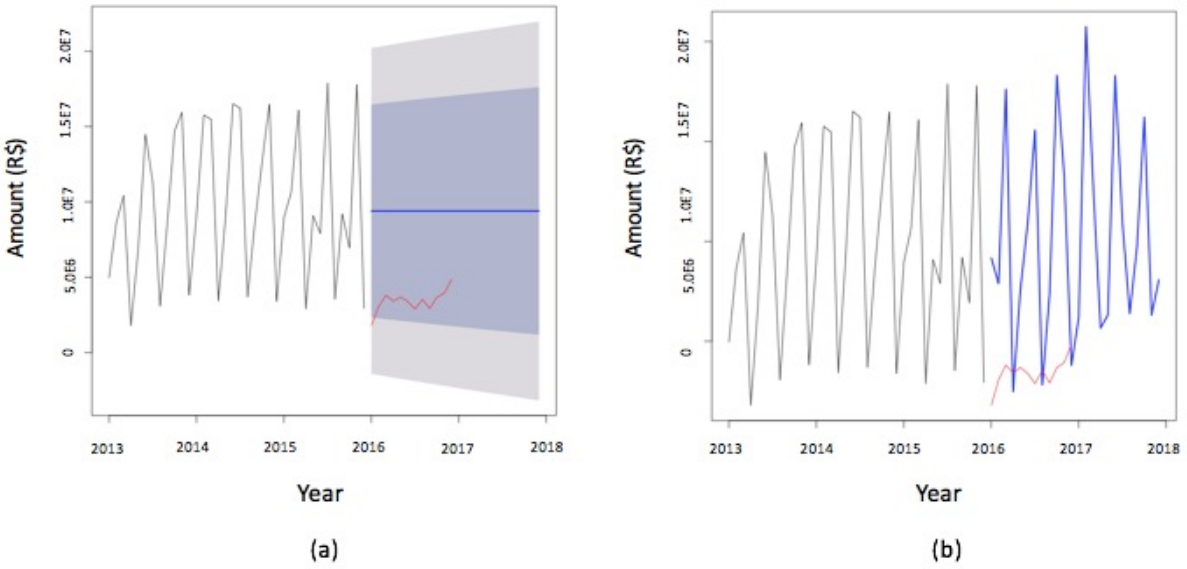

Figure 9. Comparison of Holt-Winters and ANN forecast model with actual data from 2016

To measure the accuracy of both models, it was calculated: the root-mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and the mean absolute scaled error (MASE), which are different measurements of prediction accuracy, comparing the estimation trend with the actual values. The results are displayed on Table 2. Comparing RMSE values, which demonstrates how the predictions deviated from actual values in average, Holt-Winters shows a higher value than ANN, which means that the previous deviated more than the latest. Regarding MAE, which is an average of the absolute errors, Holt-Winters presents a higher number in comparison with ANN and this explains the highest absolute difference when comparing predicted and actual values. MAPE, in turn, measures the error in percentage terms, showing an error of approximately 76.09% for Holt-Winters and only 0.6% for ANN. In other words, one can express these errors in terms of accuracy, then, showing approximately 23.09% of accuracy for Holt-Winters and 99.4% for the ANN.

After analysing the accuracy results of both selected methods for this work, it is possible to state that, considering the analysed data with the presented set up, ANN outperformed Holt-Winters method.

6 Conclusion and Future Work

In summary, an outlier detection method based on k-means clustering was implemented to analyse Brazilian deputies claim expenses, focusing on the identification of outliers regarding the total amount spent per invoice category. Moreover, this work aimed to predict the claim expenses of 2016 and 2017, comparing two popular forecasting methods, Holt-Winters and ANN, and verifying a possible reduction on the deputies allowance.

Intriguingly, after analysing the results provided by the k-means method to spot outliers, it was found that 92% of the pointed outliers corresponded to invoices that fell into the same category of divulgation of parliamentary activity. It is noticeable that this expense description carries a high level of ambiguity,

leaving unclear the exact reason for the claim. Comparing the spent amounts of these outliers with their respective allowances, 7 outliers were above their budget, varying from 1.3 to 6 times the monthly limit. It is worth mentioning that each outlier represents the total amount spent on the respective expenditure of one category per party of each state. Considering that parties have a different number of elected deputies, the monthly claim expense amount allowed for the Brazilian deputies, changes as per the state they represent, going from R$ 30.788,66 to R$ 45.612,53 (for each deputy). The analysis showed that some observations amounts varying from R$ 93.150,00 to R$ 301.400,00 are higher than the limit proposed by the parliament rules, available in the government web page [6]. Among the five different parties spotted in the seven outliers observations presenting amounts over the allowance in this work (AVANTE, PODE, PSC, PT and PTB), only two of them, AVANTE and PODE are not under investigation for corruption [7].

This work compared two popular time series forecasting methods to predict the Brazilian deputies claim expenses for the period of two years (2016 and 2017), comparing Holt-Winter and ANN. For the presented data, ANN outperformed Holt-Winters, but even with a high margin of accuracy of the ANN, the actual data showed a different behaviour for the year 2016. A possibility of this sudden change in the deputies behaviour, could be due to the fact that Brazil was facing a presidential impeachment during the course of 2016 and the population clamour for severe punishment for corrupt politicians. Since in the period from 2013 to 2016 the deputies have not claimed more than R$17,882,625.00, July/2015, and the total monthly summed allowance for all deputies is R$ 20,088,590.29 (available in the configuration manual, table exercise of the parliamentary activity), with the forecast numbers obtained by the applied ANN, this work proposes that the claim expenses allowance for the Brazilian deputies could be reduced in 8%. Reducing the current monthly allowance for the exercise of the parliamentary activity to R$18,481,503.07 would make available the amount of R$ 19,285,046.64 (yearly) that could be applied in other areas such as education, security or health.

In conclusion, this work shows that the used models can be useful in auditing public expenses, focusing on the identification of unusual claims, as well as in the prediction of fiscal year, giving insights with potential use to guide financial manoeuvres. Thus, considering the potential that machine learning presents on the understanding of data, a further investigation, detailed on daily claims, would be beneficial to broaden the comprehension of the deputies behaviour in this matter.

Considering that machine learning techniques can provide good insights and understanding of data, it would be necessary to drill down the research in a daily manner to better understand the deputies claim expenses behaviour and investigate the unusual claimed expenses to make sure that they were used in benefit of the public and not for the personal will of the deputies.

References

Albashrawi, M. and Lowell, M. (2016). Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015, *Journal of Data Science* **14**(3): 553–569.

Bratu, M. (2012). Macroeconomic forecasts accuracy in romania, *Review of Economic Studies and Research Virgil Madgearu* **5**(2): 99.

ÇAPAR, S. (2015). Importance of initial value in exponential smoothing methods, *Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi* **17**(3): 291–302.

Chandola, V., Banerjee, A. and Kumar, V. (2009). Anomaly detection: A survey, *ACM computing surveys (CSUR)* **41**(3): 15.

Chirita, M., Sarpe, D. et al. (2012). Usefulness of artificial neural networks for predicting financial and economic crisis, *Annals of Dunarea de Jos, University of Galati, Fascicle I. Econ Appl Inform* **18**(2): 61–66.

Ding, X., Li, Y., Belatreche, A. and Maguire, L. P. (2014). An experimental evaluation of novelty detection methods, *Neurocomputing* **135**: 313–327.

[6] http://www2.camara.leg.br/camaranoticias/noticias/POLITICA/474313-CONHECA-O-VALOR-DO-SALARIO-DE-UM-DEPUTADO-E-DEMAIS-VERBAS-PARLAMENTARES.html

[7] http://meucongressonacional.com/lavajato/partidos

Fanning, K. M. and Cogger, K. O. (1998). Neural network detection of management fraud using published financial data, *International Journal of Intelligent Systems in Accounting, Finance & Management* **7**(1): 21–41.

Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G. and Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges, *computers & security* **28**(1): 18–28.

Goldstein, M. and Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, *PloS one* **11**(4): e0152173.

Goodson, S., Mory, K. and Lapointe, J. (2012). Supplemental guidance: The role of auditing in public sector governance, *The Institute of Internal Auditors* .

Hamid, S. A. and Habib, A. (2014). Financial forecasting with neural networks, *Academy of Accounting and Financial Studies Journal* **18**(4): 37.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages, *International journal of forecasting* **20**(1): 5–10.

Jayasree, V. and Balan, R. V. S. (2013). A review on data mining in banking sector, *American Journal of Applied Sciences* **10**(10): 1160.

Ji, M. and Xing, H.-J. (2017). Adaptive-weighted one-class support vector machine for outlier detection, *Control And Decision Conference (CCDC), 2017 29th Chinese*, IEEE, pp. 1766–1771.

JuND, S. (2006). *Auditoria: conceitos, normas, técnicas e procedimentos: teoria e 900 questões*, Elsevier.

Kim, J., Ong, A. and Overill, R. E. (2003). Design of an artificial immune system as a novel anomaly detector for combating financial fraud in the retail sector, *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*, Vol. 1, IEEE, pp. 405–412.

Kotsialos, A., Papageorgiou, M. and Poulimenos, A. (2005). Long-term sales forecasting using holt–winters and neural network methods, *Journal of Forecasting* **24**(5): 353–368.

Kotu, V. and Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*, Morgan Kaufmann.

Kumar, V. (2005). Parallel and distributed computing for cybersecurity, *IEEE Distributed Systems Online* **6**(10).

Lecue, F. and Wu, J. (2017). Explaining and predicting abnormal expenses at large scale using knowledge graph based reasoning, *Web Semantics: Science, Services and Agents on the World Wide Web* .

Lin, J., Keogh, E., Fu, A. and Van Herle, H. (2005). Approximations to magic: Finding unusual medical time series, *Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on*, IEEE, pp. 329–334.

Mills, T. C. (2011). *The foundations of modern time series analysis*, Springer.

Ngai, E., Hu, Y., Wong, Y., Chen, Y. and 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**(3): 559–569.

Patrick, J., Okunev, J., Ellis, C. and David, M. (2000). Comparing univariate forecasting techniques in property markets, *Journal of Real Estate Portfolio Management* **6**(3): 283–306.

Paula, E. L., Ladeira, M., Carvalho, R. N. and Marzagão, T. (2016). Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering, *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, IEEE, pp. 954–960.

Phua, C., Lee, V., Smith, K. and Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research, *arXiv preprint arXiv:1009.6119* .

Ribeiro, D. C., Cordeiro, N. and Guimaraes, D. A. (2016). Interface between the brazilian antitrust, anti-corruption, and criminal organization laws: The leniency agreements, *Law & Bus. Rev. Am.* **22**: 195.

Sagaert, Y. R., Aghezzaf, E.-H., Kourentzes, N. and Desmet, B. (2018). Tactical sales forecasting using a very large set of macroeconomic indicators, *European Journal of Operational Research* **264**(2): 558–569.

Santos, M. M. (2015). Desastres naturais no brasil: um estudo das práticas de auditoria adotadas quanto à aderência ao guia intosai, *Revista da Controladoria-Geral da União* **7**(11): 18.

Song, X.-P., Hu, Z.-H., Du, J.-G. and Sheng, Z.-H. (2014). Application of machine learning methods to risk assessment of financial statement fraud: evidence from china, *Journal of Forecasting* **33**(8): 611–626.

Spence, C., Parra, L. and Sajda, P. (2001). Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model, *Mathematical Methods in Biomedical Image Analysis, 2001. MMBIA 2001. IEEE Workshop on*, IEEE, pp. 3–10.

Thiprungsri, S. and Vasarhelyi, M. A. (2011). Cluster analysis for anomaly detection in accounting data: An audit approach., *International Journal of Digital Accounting Research* **11**.

Tkáč, M. and Verner, R. (2016). Artificial neural networks in business: Two decades of research, *Applied Soft Computing* **38**: 788–804.

Tratar, L. F. and Strmčnik, E. (2016). The comparison of holt–winters method and multiple regression method: A case study, *Energy* **109**: 266–276.

Virdhagriswaran, S. and Dakin, G. (2006). Camouflaged fraud detection in domains with complex relationships, *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 941–947.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages, *Management science* **6**(3): 324–342.

Wong, S. and Venkatraman, S. (2015). Financial accounting fraud detection using business intelligence, *Asian Economic and Financial Review* **5**(11): 1187.

Wu, J., Xiong, H., Wu, P. and Chen, J. (2007). Local decomposition for rare class analysis, *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 814–823.