CONTRIBUTIONS ON BI SYSTEMS BASED ON BIG DATA AND PREDICTIVE ANALYTICS INTEGRATED WITH AN ETL PROCESS

ANTONIO MANUEL RUBIO SERRANO

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

CONTRIBUTIONS ON BI SYSTEMS BASED ON BIG DATA AND PREDICTIVE ANALYTICS INTEGRATED WITH AN ETL PROCESS

ANTONIO MANUEL RUBIO SERRANO

ORIENTADOR: JOÃO PAULO C. LUSTOSA DA COSTA

COORIENTADOR: RAFAEL TIMÓTEO DE SOUSA JÚNIOR

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO: 571/14

BRASÍLIA/DF: 30 DE JUNHO DE 2014

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

CONTRIBUTIONS ON BI SYSTEMS BASED ON BIG DATA AND PREDICTIVE ANALYTICS INTEGRATED WITH AN ETL PROCESS

ANTONIO MANUEL RUBIO SERRANO

Dissertação de Mestrado Acadêmico submetida ao Departamento de Engenharia Elétrica da Faculdade de Tecnologia da Universidade de Brasília, como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia Elétrica.

Aprovada por:

_____
Prof. João Paulo Carvalho Lustosa da Costa, Dr.-Ing., ENE/UnB
(Orientador)

_____
Prof. Flavio Elias de Deus, Dr., ENE/UnB
(Examinador Interno)

_____
Prof. Edison Pignaton de Freitas, Dr., INF/UFSM
(Examinador Externo)

_____
Prof. Rafael Timóteo de Sousa Jr., Dr., ENE/UnB
(Coorientador)

BRASÍLIA/DF, 30 de JUNHO de 2014

FICHA CATALOGRÁFICA

RUBIO SERRANO, ANTONIO MANUEL.

CONTRIBUTIONS ON BI SYSTEMS BASED ON BIG DATA AND PREDICTIVE ANALYTICS INTEGRATED WITH AN ETL PROCESS [Distrito Federal] 2014.

xi, 213p, 210 x 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2014).

Dissertação de Mestrado - Universidade de Brasília, Faculdade de Tecnologia. Departamento de Engenharia Elétrica.

REFERÊNCIA BIBLIOGRÁFICA

RUBIO SERRANO, A. M. (2014). CONTRIBUTIONS ON BI SYSTEMS BASED ON BIG DATA AND PREDICTIVE ANALYTICS INTEGRATED WITH AN ETL PROCESS

Dissertação de Mestrado em Engenharia Elétrica, Publicação xxx, Departamento de Engenharia Elétrica, Universidade de Brasília, DF, 213p.

CESSÃO DE DIREITOS

AUTOR: Antonio Manuel Rubio Serrano.
TÍTULO: CONTRIBUIÇÕES PARA UM SISTEMA DE BI BASEADAS EM BIG DATA E ANÁLISE PREDITIVA INTEGRADA EM PROCESSO DE ETL

GRAU: Mestre ANO: 2014

_____

Antonio Manuel Rubio Serrano
CLN 412, Bloco E, Sala 102
Brasilia - DF - CEP: 70867-550
toni.rubio.serrano@gmail.com

ACKNOWLEDGEMENT

I would like to express my special gratitude to my advisor João Paulo Carvalho Lustosa da Costa for giving me the opportunity to work in this project and also for offering invaluable assistance, support and guidance throughout the course of this thesis.

I would also express my gratitude to the professor Rafael Timóteo de Sousa Jr., who has also made possible this project and has helped me always I needed. I would like to extend this acknowledgement to all the workmates and friends of the Latitude laboratory and CGAUD project for their help and wishes for the successful completion of this project. I thank to Paulo H. Rodrigues, Daniel Cunha, Ruben Cruz, Beatriz Santana, Andreia Santana, Kelly Santos, Marco Marinho, Ronaldo Sebastião and the rest of the staff.

I am also grateful to the Brazilian Ministry of Planning, Budget and Management for the support given to this project.

Moreover, I am very grateful to the people that have received me, supported me and accompanied me during all these time in Brasília. Thanks to Gabriela Mosquera, Helard Becerra, Sofia Escudero, Esther Toraño, Danny Walsh, Stephanie Alvarez, Bruna Fonseca, Stefano Galimi, Nana Yung and all the people who has shared so many moments with me.

Finally, I express my special gratitude to my parents Antonio and Genoveva, to my brother Raul and to my grandparents Manolo and Emilia. Without their support and inspiration, even from the distance, this project would not be possible

RESUMO

CONTRIBUIÇÕES PARA UM SISTEMA DE BI BASEADAS EM BIG DATA E ANÁLISE PREDITIVA INTEGRADA EM PROCESSO DE ETL

Autor: Antonio Manuel Rubio Serrano
Orientador: João Paulo Carvalho Lustosa da Costa
Coorientador: Rafael Timóteo de Sousa Júnior
Programa de Pós-graduação em Engenharia Elétrica
Brasília, 30 de junho de 2014.

Esta dissertação apresenta o estudo, aplicação e análise dos conceitos envolvidos num processo de Business Intelligence (BI) em três áreas principais: extração e carga de dados, análise preditiva, e armazenamento de dados usando Big Data.

Primeiro, no processo de extração e carga de dados, as diferentes soluções adotadas no sistema de BI do Ministério de Planejamento, Orçamento e Gestão têm sido analisadas, e uma nova solução tem sido proposta para resolver as limitações detectadas nas soluções anteriores a cumprir com os novos requerimentos do sistema. Esses requerimentos incluem a capacidade de trabalhar com um maior volume de dados e a necessidade de um melhor monitoramento do processo de restituição ao erário nos casos em que um servidor público deve devolver o salário recebido indevidamente.

Na parte de análise preditiva, diversos algoritmos de predição foram estudados e comparados usando os dados do MP. As conclusões deste estudo tem sido úteis para propor um sistema automático de detecção de fraudes e uma metodologia chamada de Extração, Transformação, Predição Adaptativa e Carga (ETAPL) que inclui predição adaptativa com seleção e configuração automática do algoritmo dentro de um processo tradicional de Extração, Transformação e Carga (ETL).

Por último, as novas tecnologias de Big Data têm sido estudadas e comparadas com as atuais, de forma a avaliar a viabilidade destas tecnologias como alternativa futura no contexto do MP.

ABSTRACT

CONTRIBUTIONS ON BI SYSTEMS BASED ON BIG DATA AND PREDICTIVE ANALYTICS INTEGRATED WITH AN ETL PROCESS

Author: Antonio Manuel Rubio Serrano
Advisor: João Paulo Carvalho Lustosa da Costa
Co-Advisor: Rafael Timóteo de Sousa Júnior
Post-graduate Program on Electrical Engineering
Brasilia, 30 June 2014.

This dissertation presents the study, application and analysis of the concepts involved on the process of a Business Intelligence (BI) solution in three main areas: data extraction and loading, predictive analytics and storage systems using Big Data.

First, in the BI data loading, the different previous solutions into the BI system of the Brazilian Ministry of Planning, Budget and Management (MP) has been analysed, and a new solution has been proposed for solving the limitations of the previous ones and for fulfilling the new requirements appeared on the project. Those requirements include the necessity of managing a bigger volume of data or the need for a better monitoring of the reimbursement process that is executed when a public servant has to refund the erroneously received money..

In predictive analytics, several prediction algorithms have been analysed and compared using the data of the MP. The results has been useful for proposing an automatic fraud detection system and a new methodology called Extract, Transform, Adaptive Prediction and Load (ETAPL) that includes predictive analytics into a traditional Extract, Transform and Load (ETL) process.

Finally, the new Big Data technologies have been studied and tested as future alternative for the current storage systems at the MP.

SUMMARY

LIST OF FIGURES

DESCRIÇÃO GERAL DO PROJETO

Inteligência de Negócio, em inglês, *Business Intelligence* (BI), pode-se definir como a capacidade de uma organização para coletar, manter e organizar conhecimentos. O objetivo final dos sistemas de BI é assistir na tomada de decisão. Para isso, os sistemas de BI incorporam funcionalidades tais como criação de relatórios, processamento analítico online, mineração de dados, mineração de processos, processado de eventos complexos, gestão do desempenho da empresa, *benchmarking* e análise preditiva.

Nesta dissertação, o objetivo é analisar e melhorar os processos de BI em três áreas: extração e carga dos dados em um sistema de BI, análise preditiva e armazenamento de dados. O trabalho foi desenvolvido no contexto do sistema de BI do Departamento de Auditoria de Recursos Humanos (CGAUD) do Ministério do Planejamento, Orçamento e Gestão (MP).

A extração e carga dos dados é um processo crucial na criação de um sistema de BI. No contexto do MP, diversas soluções foram adotadas durante o desenvolvimento do projeto, começando por uma simples arquitetura composta somente por um banco de dados relacional até chegar a uma solução mais complexa que inclui um *Operational Data Store* (ODS), um processo de Extração, Transformação e Carga (ETL) e diversos *Data Marts* (DM).

Essa evolução se explica por dois motivos. Primeiro, a aparição de novos requisitos, como uma maior granularidade da informação nos relatórios finais sem perder em desempenho, ou a inclusão de um módulo que permitisse o acompanhamento do processo de reposição ao erário dos servidores que receberam salários indevidos. Em segundo lugar, a aparição de novas limitações por causa do incremento no volume de dados a ser tratado, pois o banco é alimentado todos os meses com novas informações.

Neste projeto, tem se estudado as diversas soluções existentes na arquitetura de sistemas de BI, e tem se implementado a mais apropriada ao cenário específico do MP, conseguindo resolver as limitações expostas.

Outro aspecto importante nos sistemas de BI é a capacidade para implementar análise preditiva. Na literatura existem diversos algoritmos e métodos para implementar estas funcionalidades. Por isso, a primeira fase deste projeto foi o estudo e análise destes

algoritmos. Os algoritmos que foram considerados neste projeto são as Redes Neurais Artificiais (ANN), os Algoritmos Autoregressivos (ARX) e os Processos Gaussianos (GPR). Estes algoritmos podem ter diferentes configurações, e a configuração ótima vai depender de cada tipo de dado. Portanto, uma série de testes deve ser realizada para obter a melhor configuração para os parâmetros de cada um dos algoritmos.

Os testes do estudo de caso apresentados neste projeto foram realizados utilizando os dados da Secretaria do Património da União (SPU). A SPU utiliza diversos indicadores para monitorar os processos internos. Cada indicador se compõe de uma série de valores coletados no tempo. Neste estudo de caso, os valores foram agrupados em meses, formando assim vetores de amostras ordenadas em que cada posição representa o valor de um mês. Esse tipo de dado se conhece como série temporal.

Outro aspecto importante são as avaliações das predições realizadas. Para isso, neste projeto propõe-se o uso de dois indicadores: o Erro Quadrático Médio Normalizado (NRMSE) e o Coeficiente de Determinação (COD). O uso destes dois indicadores permite a avaliação objetiva e automática da qualidade das predições.

No contexto específico do BI, muitas plataformas têm incorporado módulos de análise preditiva. Porém, na maioria dos casos o potencial desse tipo de análise não é completamente explorado. Isto se deve principalmente a dois motivos: por um lado, alguns sistemas de BI utilizam somente alguns algoritmos predefinidos sem levar em conta o tipo de dados; por outro lado, existem soluções de BI mais customizáveis mas que precisam de um maior conhecimento matemático e técnico para configurá-los, o que não acostuma a ser o caso nos usuários finais de um sistema de BI.

Neste trabalho, propõe-se o uso do NRMSE e do COD para identificar automaticamente o algoritmo que apresentou o melhor desempenho para cada tipo de dados. Com base nessas duas métricas, propõem-se duas novas metodologias: um sistema automático de detecção de fraude usando uma ANN como preditor; e um processo de ETL com análise preditiva com seleção e configuração automática dos algoritmos.

O terceiro aspeto estudado neste projeto e o armazenamento de dados. O sistema de BI da CGAUD é baseado nos dados recebidos num arquivo proveniente do SIAPE, um banco de dados que contém as informações dos servidores públicos brasileiros. O tamanho do arquivo é atualmente de 16GB, e vai incrementando mês a mês. Apesar de que o banco de

dados atual é capaz de trabalhar com esse volume de dados, neste trabalho exploram-se as novas tecnologias de armazenamento em Big Data, a forma de avaliar sua viabilidade e comparar o desempenho destas com os sistemas atuais.

Em resumo, as contribuições desta dissertação de mestrado são:

- estudo e comparação de desempenho das diferentes arquiteturas de BI da CGAUD, e proposta de solução melhorando as limitações existentes;
- estudo e avaliação dos diferentes algoritmos de predição usando os dados da SPU;
- proposta de indicadores para avaliar automaticamente a qualidade das predições;
- proposta de um sistema automático de detecção de fraudes em series temporais baseado num preditor;
- proposta de metodologia de ETL incluindo análise preditiva com seleção e configuração automáticas do algoritmo (ETAPL);
- estudo e avaliação do desempenho das tecnologias de Big Data no contexto da CGAUD.

Além das contribuições expostas acima, durante o desenvolvimento desta tese de mestrado, outros resultados têm sido alcançados. Devido a que estes resultados não são o foco principal deste trabalho, serão sumarizados seguidamente.

Em (Campos, et al., 2012), propõe-se a utilização de indexação ontológica através de mapas conceituais para gerar evidências de irregularidades na folha de pagamento do MP. A proposta usa indicadores de auditoria como instrumento para argumentação documental, permitindo acompanhar os dados desde o momento da sua disponibilização e estruturação.

De forma similar, o uso de mapas conceituais num processo de ontologia durante a criação de um sistema de BI é proposto em (Fernandes, et al., Construction of Ontologies by using Concept Maps: a Study Case of Business Intelligence for the Federal Property Department, 2012). A metodologia proposta permite fazer mais rápido o processo de criação de um sistema de BI e reduz a complexidade do processo de validação.

Em um tipo diferente de aplicações, dois artigos propõem o uso das técnicas de estado da arte em Seleção da Ordem de Modelo (MOS) para detecção automática de tráfego malicioso nos dados coletados por um *honeypot*. Em (da Costa, et al., 2012) o uso do algoritmo *Modified Exponential Fitting Test* (M-EFT) aplicado para dados coletados por um único *honeypot* é proposto, enquanto em (da Costa, Freitas, Serrano, & de Sousa Jr., 2012) essa proposta foi melhorada para o caso de uma rede de *honeypot*, em que o uso dos algoritmos *R-D Akaike* e *R-D Minimum Description Lenghts* são mais apropriados.

Por último, foram criadas duas patentes. A primeira, (Rubio Serrano, da Costa, & de Sousa Jr, 2013), consiste em um Sistema de Detecção Cega e Automática de Fraudes em Indicadores Modelados como Séries Temporais Usando Técnicas de Análise Preditiva; enquanto a segunda, (da Costa, Rubio Serrano, Rodrigues, Campos, & de Sousa Jr, 2013), propõe um sistema de Análise Preditiva em Sistemas de Inteligência de Negócios através de um Sistema Multialgoritmo no Processo de Extração, Transformação, Predição Adaptativa e Carga.

1 INTRODUCTION

Business Intelligence (BI) can be defined as the ability of an organization to collect, maintain, and organize knowledge. The final objective of the BI systems is to assist with the decision-making. For this task, BI systems have different functionalities such as reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, and predictive analytics (Rahman, Monzurur, Faisal, & Mushfiqur, 2014).

This master's thesis is focused on the study and analysis of BI systems in order to propose new methodologies that improve its capabilities and make them more useful for assisting the decision-making process. To do that, this work initiates with the observation of traditional BI systems architectures, where three aspects of interest have been identified: architecture of BI environments, predictive analytics and data storage systems.

The work has been developed in cooperation with the General Audit Coordination Department, in Portuguese *Coordenação Geral de Auditoria* (CGAUD) belonging to the Brazilian Ministry of Planning, Budgeting and Management, in Portuguese *Ministério do Planejamento, Orçamento e Gestão* (MP). This master's thesis uses the BI environment of CGAUD as case study, which allows validating the proposals in a real environment.

The CGAUD BI system, as well as several BI environments that the MP has incorporated in the lasts years, uses the open source Suite Pentaho[1], an open source platform which includes the functionalities of data integration, On-Line Analytical Processing (OLAP) services, reporting, dash boarding, data mining and ETL capabilities.

The first area of interest is the configuration of the BI architectures. In the MP context, several solutions have been adopted during the development of this project, starting with a simple one using just one relational database and evolving to a complex one including an

[1] Pentaho is a company that offers Pentaho Business Analytics, a suite of open source Business Intelligence (BI) products which provide data integration, OLAP services, reporting, dashboarding, data mining and ETL capabilities. (See: www.pentaho.com).

Operational Data Store (ODS), an Extract-Transform-Load (ETL) process and several Data Marts (DM). This evolution happened because of two reasons. First, new requirements has appeared, such as a better granularity on the final reports without losing performance, or the inclusion of a new software module for monitoring the reimbursement process of those public servants that has received inappropriate salaries. Second, new limitations have appeared, specially the necessity of managing a bigger volume of a database that is updated each month. The different solutions adopted throughout the project and its advantages and drawbacks are deeply analysed in this master's thesis.

The second important aspect studied in this project is the predictive analytics techniques. In BI systems, their capacity for successfully implementing predictive analysis is crucial for assisting the decision-making process. In the literature there are several algorithms and methods for implementing the predictive functionalities desired for this project. Due to that, the first step required is to study and test the performance of these algorithms. In this project, the algorithms considered are the Artificial Neural Networks (ANN), the Autoregressive Algorithms (ARX) and the Gaussian Processes (GPR). These algorithms can have different configurations, and the optimal one may be different for each kind of data. Thus, a set of tests should be done in order to discover which configuration is the best for the different parameters of each algorithm.

The tests of the case study presented in this project have been done using the data of the Federal Property Department, in Portuguese *Secretaria do Patrimônio da União* (SPU). The SPU uses several indicators for controlling its processes. Each indicator is composed by several values of certain information collected over the time. In this case study, the values are grouped per months, forming vectors of ordered samples where each position represents the value of a certain month. Such modelling is known as a time series.

Also, an objective of this work is to make the predictive analytics systems as automatic as possible. Therefore, a very important issue is the automatic evaluation of the results. Due to the fact that the predictions are annual, but expressed in months, the resulting prediction is a vector of twelve positions, where each value represents a month. Thus, a method for comparing two vectors in order to detect how similar they are is required. In this project, two indicators are used: the Normalized Root Mean Square Error (NRMSE) and the Coefficient of Determination (COD).

In lasts years, BI systems are incorporating predictions modules. However, most of the systems do not exploit all the potential of this kind of analysis. This is mainly because of two reasons: on the one hand, some BI software only uses some predefined algorithms that are not always the optimal option for all types of data; on the other hand, there are BI solutions that includes several prediction algorithms and configurations but they require a quite high mathematical knowledge for configuring it, which is not usually the case of the BI users.

In this work, the use of the two evaluation metrics (NRMSE and COD) has been proposed in order to automatically evaluate the best prediction for each type of data. Also, based on those evaluation metrics, two new methodologies have been proposed: a fraud detection system based on a Neural Network predictor; and an ETL process including a prediction module with automatic algorithm selection and optimization (ETAPL).

The third main aspect studied on this project is the data storage. The BI system of the CGAUD is based on a file coming from the SIAPE database, a database containing the information about all the public servants in Brazil. .This file has information of about two and half million workers, among them – active, inactive and retired. The actual size of each SIAPE file per month is about 16GB and is growing every month. Although the actual database is able to handle this volume of data, the new Big Data technologies are explored in this work, studying its viability and comparing the performance with current storage systems, in order to be able to use in it future scenarios.

In summary, the objectives of this master's thesis are the following:

- to study and compare the performance of different BI architectures in the CGAUD scenario, and propose the best one for solving the requirements;

- to study and compare different prediction algorithms using the data of the SPU, and evaluate its applicability for the BI system of the CGAUD;

- to propose a method for including predictive analytics into the ETL process with automatic algorithm selection and optimization (ETAPL).

- to study and evaluate the performance of new Big Data technologies as storage system for the CGAUD project.

Besides the mentioned contributions, during this master's thesis, other results have been achieved. Since these results are not the focus of this work, we summarize them as follows.

In (Campos, et al., 2012), the usage of the ontology indexation process via concept maps to generate evidences of irregularities in payrolls of the MP is proposed. The proposal uses audit indicators as instrument for documental argumentation, allowing following the data from the moment in which it is provided and structured.

In a similar way, the use of concept maps in the ontology step during the BI system construction process is proposed in (Fernandes, et al., 2012). The proposed methodology makes the implementation of the BI system faster and reduces the time and complexity of the validation process.

In a different kind of application, two papers propose the use of state-of-the-art Model Order Selection (MOS) schemes for automatically detecting malicious traffic on the data collected by a honeypot. In (da Costa, et al., 2012), the use of the Modified Exponential Fitting Test (M-EFT) applied to the data collected by a single honeypot is proposed, while in (da Costa, Freitas, Serrano, & de Sousa Jr., 2012) this proposal is improved for the case of a honeypot network, when the use of the algorithms R-D Akaike and R-D Minimum Description Length is more suitable.

Finally, two new patent has been created. The first one, (Rubio Serrano, da Costa, & de Sousa Jr, 2013), consist on an Autimatic Blind Fraud Detection System on Indicators modelled as Time Series using Predictive Analytics Techniques; while the second one, (da Costa, Rubio Serrano, Rodrigues, Campos, & de Sousa Jr, 2013) proposes a method for including predictive analytics into the ETL process with automatic algorithm selection and optimization (ETAPL).

The remainder of this master's thesis is organized as follows. First, in Chapter 2, a theoretical overview about the main the main concepts of BI architectures, the prediction algorithms and the Big Data technologies used in this case study is done. In chapter 3, the improved BI environment for the CGAUD is explained and the results are exposed. The Chapter 4 contains the results of the case study on Predictive Analytics for the CGAUD and two proposals for its application to fraud detection and BI environments. In Chapter 5 the new Big Data storage systems are analysed in the context of the CGAUD project. Finally, the conclusions and future works are drawn in Chapter 6.

2 THEORETICAL FOUNDATION

This chapter is devoted to establish a theoretical foundation. It contains a general overview about the concepts and technologies involved in this master's thesis. The Chapter is subdivided in three parts: first, an overview of the main concepts and methodologies used for building BI environments is done; then, a review of key concepts of time series analysis is presented; and, finally, the basic concepts of the Big Data technologies used in this master's thesis are exposed.

2.1 Traditional BI environments

The term Business Intelligence is referred to methodologies, architectures, technologies, tools and software devoted to transform raw data into useful information in order to assist the decision-making process (Evelson, Moore, Karel, Kobielus, & Coit, 2009).

In the past two decades it has been seen an explosive growth, both in the number of products and services offered and in the adoption of these technologies by industry. This growth has been fuelled by the declining cost of acquiring and storing very large amounts of data. Enterprises today collect data at a finer granularity, which is therefore of much larger volume (Chaudhuri, Dayal, & Vivek, 2011).

BI systems involve tasks of collecting, organizing, analysing and sharing data. This information should be obtained from a usually heterogeneous set of several relational databases or unstructured data sources. Traditional BI environment uses to implement architecture as the one depicted in Figure 2-1.

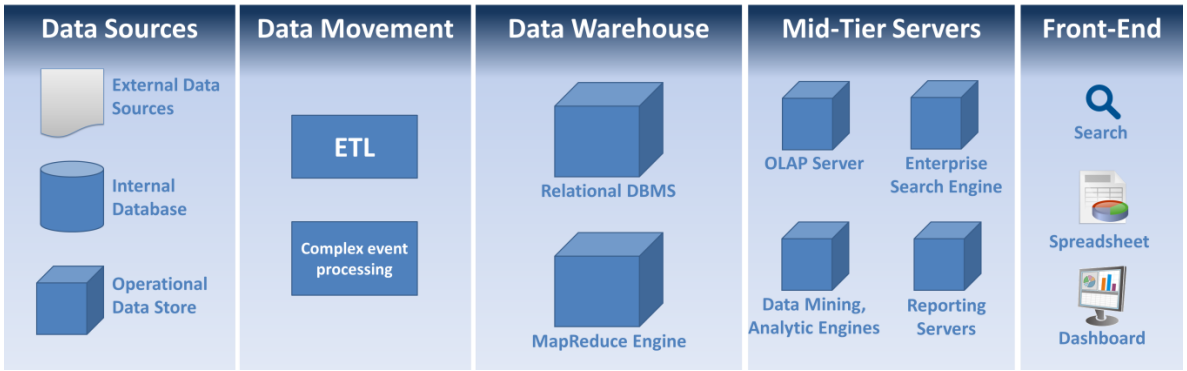

Figure 2-1. Traditional generic architecture for a BI environment

The first step will be to execute an Extract-Transform-Load (ETL) process, although some pre-processing steps can be done when the data is coming from external data source, such as the use of an Operational Data Store (ODS), as explained in Section 5.3.

ETL refers to a collection of tools that play a crucial role in helping discover and correct data quality issues and efficiently load large volumes of data into the warehouse. The accuracy and timeliness of reporting, ad hoc queries, and predictive analysis depends on being able to efficiently get high-quality data into the Data Warehouse (DW) from operational databases and external data sources

DWs store current and historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons. The basic element of those structures is the multidimensional cube. It is a data structure composed by two elements: the dimensions and the fact table. The edges define the context of the cube qualitatively, namely, indicates what the information of the cube is about. The fact table contains only the measures (quantitative values) corresponding to each position of the edges.

Other basic elements on BI systems are the Data Marts (DM). They are subsets of data of the DW, and are devoted to respond for a more specific necessity or for working with a specific population.

How to configure the architecture of the DM and DW of a BI environment is a key concept that will impact on the performance of the system. Two main authors are considered as references in this area: Inmon and Kimball, and they proposed two different approaches for constructing a BI system.

In (Kimball, 1998), a bottom-up structure is proposed. In this approach, the ETL process will load the resulting data into a one or several DM first, and then the DM will be combined in order to build a DW. This kind of configuration is easy to implement and more flexible, due it is possible to start with a unique DM and later add more DM as needed. It makes this approach more suitable for those cases where the questions to be solved may change along the project. However, the use of a unique DW to be sourced by the reporting layer may be a limitation when working with high volumes of data or when a high granularity is required.

On the other hand, (Inmon, 1992) proposes a top-down structure. In this case, the data extracted with the ETL process is consolidated into a DW. Then, this DW is broken down into several DM, each one correspondent to a certain department or knowledge area. The reporting layer will recover the data from each one of the DM. This is a less flexible approach, due to the fact that any change on the load process will require modifying the entire architecture (DW and DM). However, it may be recommended for environments with well pre-defined questions, different knowledge areas at the output and high granularity on reports.

Figure 2-2 shows the architecture model proposed by Inmon, and Figure 2-3 shows the architecture model proposed by Kimball.

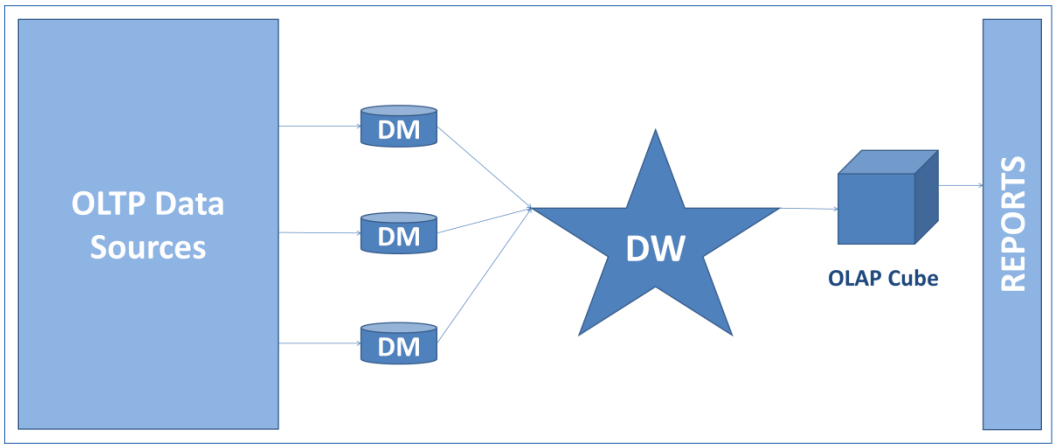

Figure 2-2. Inmon's bottom-up architecture model

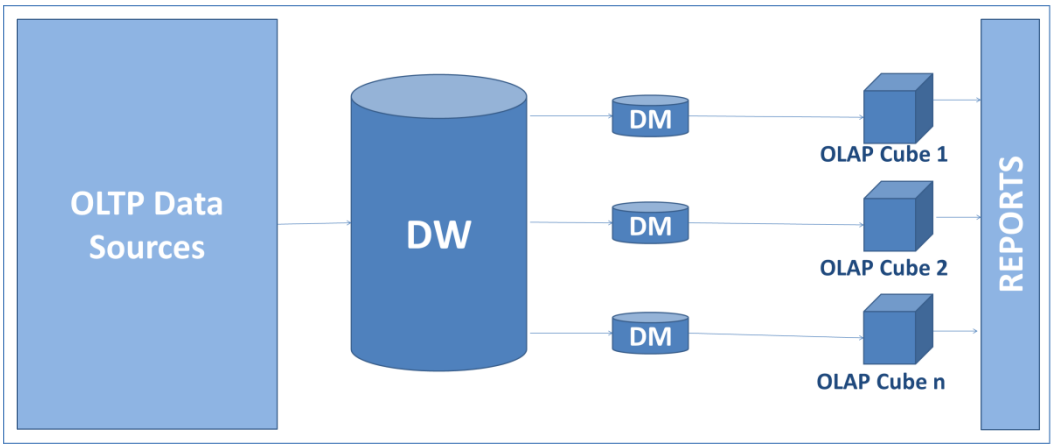

Figure 2-3. Kimball's top-down architecture model

There is no unique solution, and best choice will depend on the conditions of each BI environment. Those conditions are the type of data to deal with, the kind of questions to be solved, but also the characteristics of the client and the maturity its requirement definition. Those questions are analysed for the study case considered on this master's thesis in Chapter 3.

2.2 Time Series Analysis

BI systems are very useful as they provide information about indicators that allow monitoring the status of the relevant processes into an enterprise or organization. Those indicators are sets of sorted values that conforms a time series and, although present and historical data contains relevant information, it may not be enough when dealing with complex time series whose patterns are not intuitively seen by visual inspection. For that reason, predictive analytics is a science that has long been studied, and has become a critical feature of BI environments.

The effectiveness of predictive analytics algorithms may vary a lot depending on the context and the type of data. In this chapter, three prediction algorithms are explored for determining how effective they are in the context of the BI environment of the CGAUD. The three algorithms considered in this study are: Artificial Neural Networks (ANN), Autoregressive Models (AR) and Gaussian Process (GPR).

In the context of time series prediction, ANN are of special interest, due can be useful for nonlinear processes that have an unknown functional relationship and as a result are difficult to fit (Darbellay & Slama, 2000), and has been widely used on forecasting (De Gooijer & Hyndman, 2006).

In order to assess the accuracy of ANN, the results has been compared with those obtained with AR, an older algorithm also widely used for time series analysis. A complete comparison of those algorithms can be found on (Rubio Serrano A. M., Study and implementation of a predictive analytics module for the Business Intelligence System of the Brazilian Ministry of Planning, Budgeting and Management, 2012).

In this Master's thesis, the results on (Rubio Serrano A. M., Study and implementation of a predictive analytics module for the Business Intelligence System of the Brazilian Ministry

of Planning, Budgeting and Management, 2012) have been re-evaluated though an improved configuration of the error measurement, and also GPR has been considered as a third algorithm for comparing the results.

Two properties make GPR an interesting tool for inference. First, a Gaussian process is completely determined by its mean and covariance functions, requiring only the first and second order moments to be specified, which makes it a non-parametric model whose structure is fixed and completely known. Second, the predictor of a Gaussian process is based on a conditional probability and can be solved with simple linear algebra, as shown in (Davis, 2001).

The remainder of the chapter is devoted to, first, review the basic concepts of time series analysis and, second, review de mathematical background of the aforementioned algorithms.

2.2.1 Basic concepts

A univariate time series is a sequence of measurements of the same variable sequentially collected over time. Most often, the measurements are made at regular time intervals. Depending on the kind of information contained on the time series, the time intervals can be different. For example, for some geological process the time intervals may be centuries, while for some biological activity the measurement interval can be seconds.

In this project we will analyse several time series corresponding to different indicators used by the Federal Patrimony Department (in Portuguese *Secretaria do Patrimônio da União*, SPU). The SPU is a department of the Brazilian Ministry of Planning, Budget and Management (MP), and its mission is defined as:

"To know, watch and ensure that each property of the Union fulfils its socio-environmental function in harmony with its tax collecting function in support for the strategic programs for the Nation."

The indicator used for the case study presented in this project is the tax collection in the SPU. This data is expressed in months and starts on January 2005 and goes up to December 2010. To model this data, we assume a random variable x_n with the costs of each month, where for $n = 1, \dots, 12$ indicates the months of the first year, for $n = 13, \dots, 24$ indicates the months of the second year and so on.

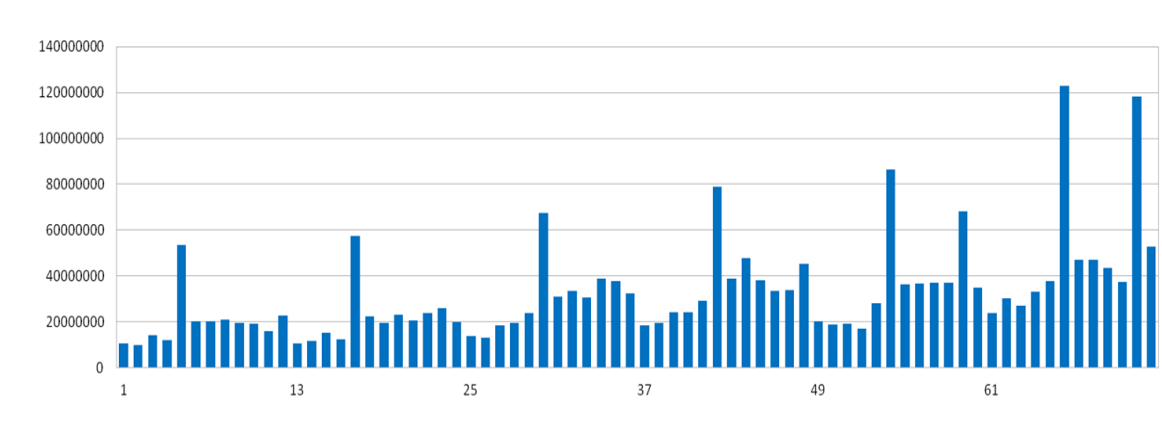

Figure 2-4 Tax collection per month in the SPU

The first step on a time series analysis process is always to plot the data and make a first study by visual inspection. If we look at Figure 2-4, the mean value is increasing with time. Therefore, there is a trend component in this time series. Also, we can detect a repeated pattern each year, due the months have smaller values at the beginning and the end of the year, while the values are higher at the middle of the year. Finally, in all the years there are some peaks or outcomes at the end of the first and second semester. This is because the SPU tries to collect the delayed payments before the end of the semester.

2.2.1.1 Mean, Covariance and Variance

The expected value of a discrete random variable $E(x_n)$ is defined as the weighted average of all the possible values that the variable can take on. It is defined as:

$$E[x_n] = \sum_{n=1}^{\infty} x_n \; p_n \,, \qquad (2.1a)$$

where p_n represents the probability of the outcome x_n.

In practice, only a limited number of samples are available and, normally, the probabilities for each event are unknown. Therefore, in most cases the expected value is approximated by the mean value.

$$E[x_n] \cong \bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \,, \qquad (2.1b)$$

where *N* represents the total number of samples.

Another basic statistical parameter to characterize a time series is the covariance. This parameter measures how much two random variables *x* and *y* change together.

$$Cov(x,y) = \sigma\,(x,y) = E\,[(x - E[x])(y - E[y])]\,. \qquad (2.2)$$

The variance is the special case of the covariance where, instead of comparing two different variables, we compare two identical variables:

$$\sigma\,(x,x) = \sigma^2(x) = E\,[(x - E[x])(x - E[x])] = E\left[\left(x - E(x)\right)^2\right]. \qquad (2.3)$$

If we substitute the expectation operator by the definition in (2.1b), where the expected value is approximated by the mean value, we obtain

$$\sigma^2(x) \cong \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2\,. \qquad (2.4)$$

2.2.1.2 Autocorrelation Function (ACF)

From a statistical point of view, the term correlation is referred to any broad class of dependence between two random variables or sets of data. Let x_n and y_n be two random process where t represent the different realizations or time instants. Then, the correlation between these two processes can be calculated with the Pearson's Product-Moment Coefficient:

$$\rho_{x,y} = corr\,(x,y) = \frac{Cov(x,y)}{\sigma(x)\sigma(y)} = \frac{\sigma(x,y)}{\sigma(x)\sigma(y)}\,. \qquad (2.5)$$

The result of (2.5) will be 0 if x_n and y_n are completely uncorrelated. The maximum and minimum values will be +1 and -1 respectively, where the first case indicates positive (or

increasing) correlation and the second case indicates negative (or decreasing) correlation. The rest of the values will indicate a certain degree of positive or negative correlation.

In a similar way to the correlation, the autocorrelation function (ACF) describes the correlation between the values of a random variable at different time instants. Although various estimates of the ACF exist, in this work the form presented in (Box, Jenkins, & Reinsel, 2008) is adopted. The value of the ACF for a certain lag value h can be defined as

$$r_x(h) = \frac{Cov(x_n, x_{n-h})}{\sigma_n \, \sigma_{n-h}}. \tag{2.6}$$

If we assume that the random process is wide sense stationary and we substitute the definition of the covariance in (2.2), then we can rewrite (2.6) as

$$r_x(h) = \frac{E\,[(x_n - E[x_n])(x_{n-h} - E[x_{n-h}])]}{\sigma_n^2}. \tag{2.7}$$

And replacing the equation (2.1) in (2.7), we obtain

$$r_x(h) = \frac{\sum_{t=0}^{N-h}\left((x_n - \overline{x_1})(x_{n-h} - \overline{x_2})\right)}{\sigma_n^2}, \tag{2.8}$$

where h represents the lag and its value should be in the range $0 < h < N$.

As it can be observed in (2.8), the sample autocorrelation function (ACF) measures the statistical relationship between different observations in a single data series. A very useful way to analyse these relations is to plot the values and make a visual inspection.

In our case study, we have computed the autocorrelation of the Tax Collection time series using MATLAB. The Figure 2-5 shows the ACF with $h_{max} = 36$.

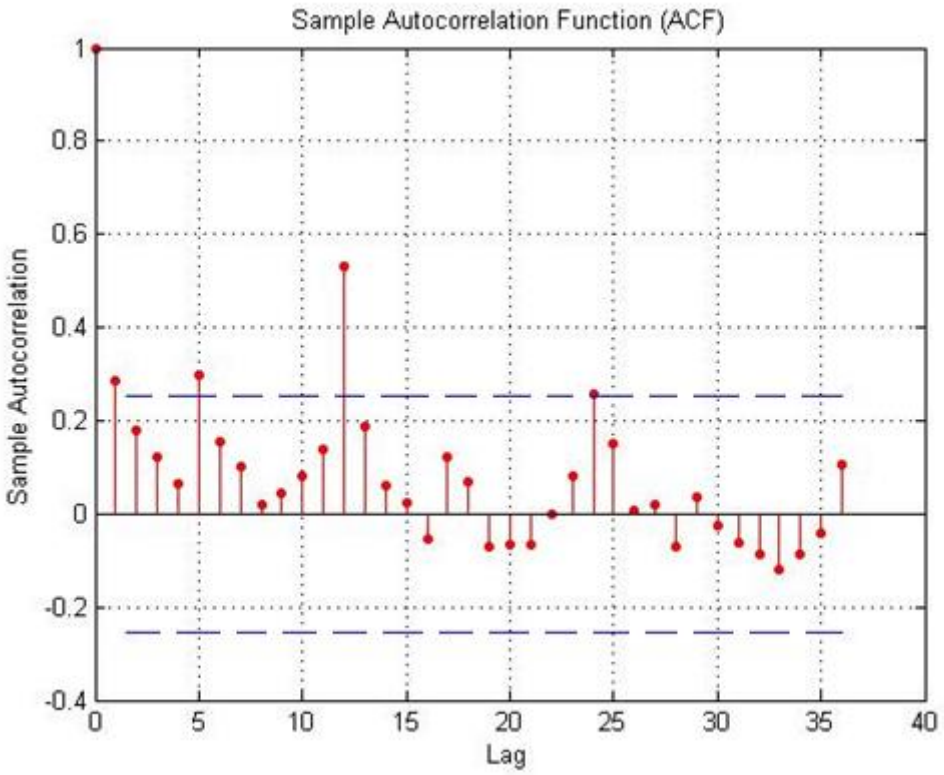

Figure 2-5 Autocorrelation Function (ACF) with h=36 for the Tax Collection data series

In Figure 2-5 the values of the correlation coefficient for different values of lag. It represents how much the values of the time series are correlated with the past values at different lags.

The horizontal boundary dotted lines in the figure are called the confidence bounds, and determine if the correlation from a certain lag is random or not with a 95% of confidence. The approximate calculation that is normally used for these values is $\pm\frac{2}{\sqrt{N}}$, being N the maximum number of samples considered on the time series. Therefore, the values that are higher than the confidence bounds can be considered significant values; while those lags that presents a correlation below the confidence line we will be considered as uncorrelated.

Thus, we can conclude from the Figure 2-5 that a lag value equal to 12 is the one with more correlation.

In order to show, in a visual way, the meaning of the value of the ACF for different lags, it is useful to plot various x_{n+h} values (for $h = 1, 2, \dots$) against the reference observations x_t. For example, in Figure 2-6, the correlation between the original time series and the 1 position-delayed series is shown.

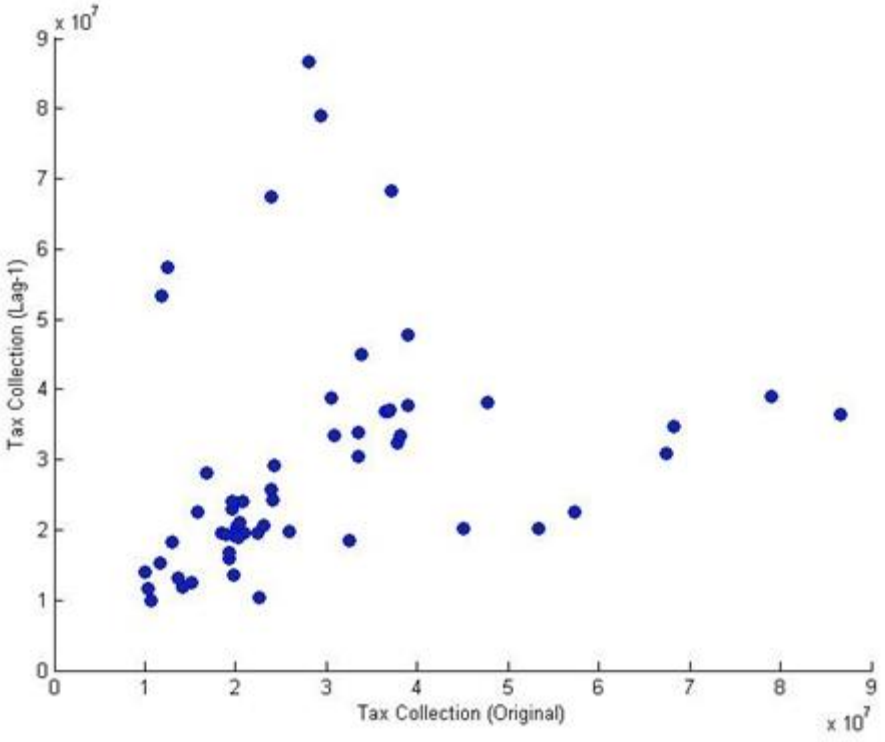

Figure 2-6 Original against lag-1 time series

In Figure 2-6, there is a pattern, since the samples with small values are less spread than the samples with great values. Near to the origin, the values seem to increase in a similar way, showing some positive relationship. However, when the values become higher, they are more spread out, and the linear pattern of the beginning is lost. Intuitively, we could say that probably there is some correlation between these two time periods, but not so important. It agrees with the value of the ACF for $h = 1$, in Figure 2-6, where the value of the function is slightly higher than the confidence bands.

In a similar way, in Figure 2-7 there is a plot of the reference time series against the 2 position delayed time series.

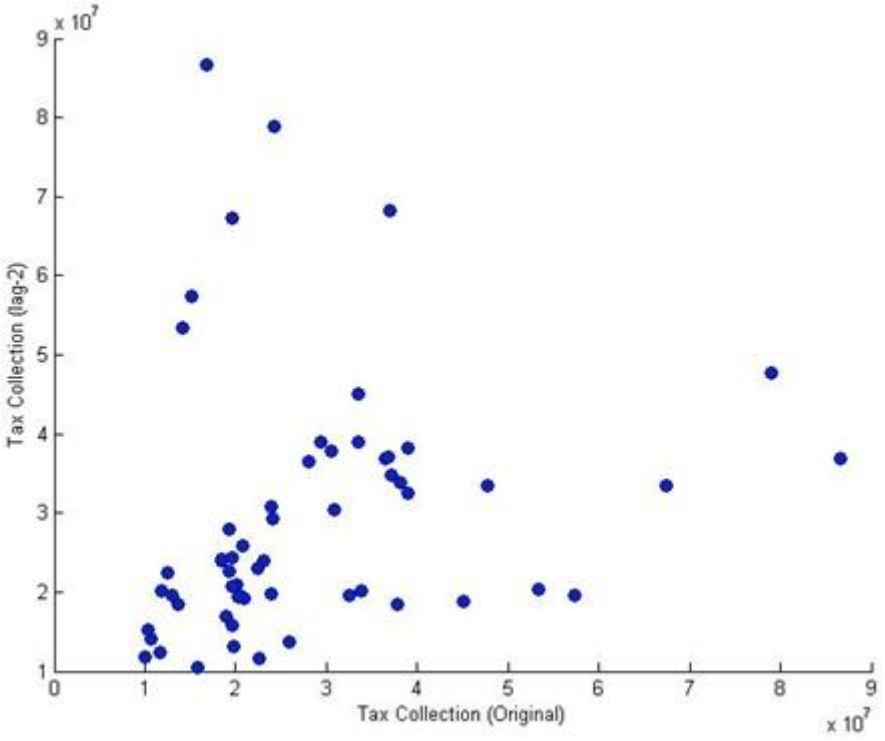

Figure 2-7 Original against lag-2 time series

In this case, the values seem to be spread out in a random way, without any pattern or correlation between the two time series. Such absence of correlation is also represented on the ACF in Figure 2-5, where the value of the function for $h = 2$ is below the confidence boundary.

Unlike Figure 2-6 and Figure 2-7, Figure 2-8 shows a high degree of correlation that can be observed by visual inspection. In this figure, the reference time series is plotted against a 12 position-delayed time series.

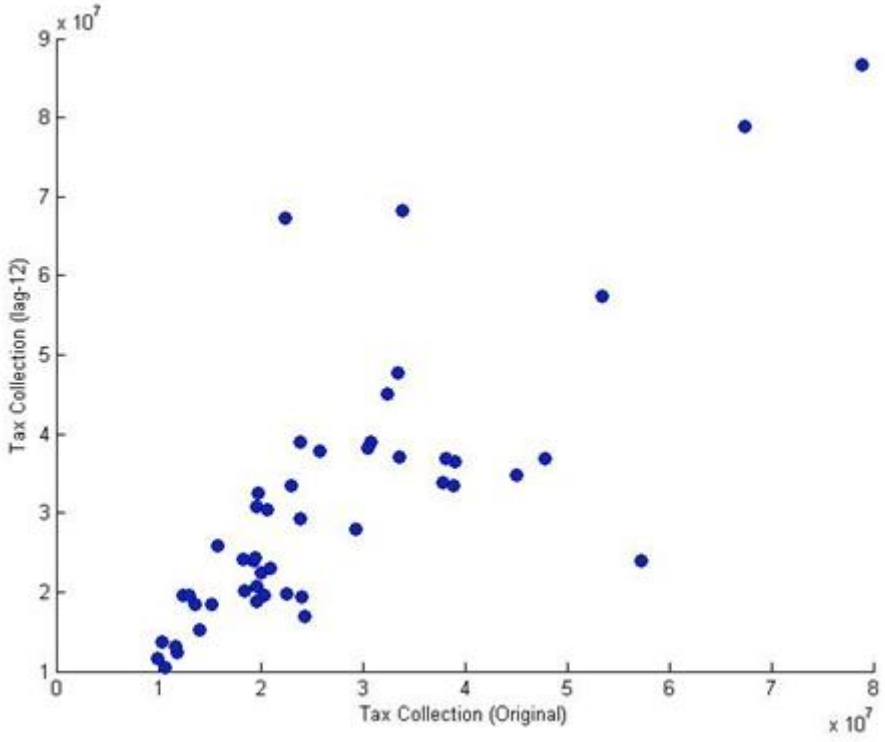

Figure 2-8 Original against lag-12 time series

The high correlations that can be observed in Figure 2-8 corresponds to the higher value on the ACF, which is the case of $h = 12$.

2.2.1.3 Stationary time series

In a formal definition, a strictly stationary stochastic process is one where given $n_1, \dots, n_l$ the joint statistical distribution of $x_{n1}, \dots, x_{nl}$ is the same as the joint statistical distribution of $x_{n1+h}, \dots, x_{nl+h}$ for all l and h. This is an extremely strong definition: it means that all moments of all degrees (expectations, variances, third order and higher) of the process, anywhere are the same. It also means that the joint distribution of (X_n, X_s) is the same as (X_{n+r}, X_{s+r}) and hence cannot depend on s or n but only on the distance between s and n, i.e. $s - n$ (Nason, 2004).

Since the definition of strict stationarity is generally too strict for everyday life, a weaker definition of second order or weak stationarity is usually used. Weak stationarity means that the mean and the variance of a stochastic process do not depend on *n* (that is they are constant) and the auto covariance between X_n and X_{n+h} depend only on the lag h (h is an integer, the quantities also need to be finite). Thus, a time series $\{x_n, n = 0, 1, 2, \dots\}$ is said

to be stationary if their statistical properties are the same or very similar to those of the time series at any other time instant $\{x_{n+h}, n = 0, 1, ...\}$ for each integer h (Brockwell & Davis, 2002).

2.2.1.4 Trend and Seasonal components

The time series can be described by means of four elements or components: the Trend (T), the long term Cycle (C), the Seasonal component (S) and Irregular component or outliers (I). However, not all of them are always present in a time series, and they can be related them through different decomposition models. These elements can be defined in the following way:

- There is a trend (T) when, on average, the measurements tend to increase or decrease over time.
- There is a cycle (C) if exists a long-run cycle or period unrelated to seasonal factors.
- There is a seasonal component (S) if it is possible to find a regularly repeating pattern of highs and lows related to the calendar, such as seasons, quarters, months, etc.
- There are irregular components or outliers (I) when there are abrupt variations present on the time series that cannot be explained according to the previous definitions.

In order to choose an appropriate decomposition model, a graphical analysis of the time series is suitable. In this way, the analyst will examine a graph of the original series and try a range of models, selecting the one which yields the most stable seasonal component. If the magnitude of the seasonal component is relatively constant regardless of changes in the trend, an additive model is suitable. If it varies with changes in the trend, a multiplicative model is the most likely candidate. However if the series contains values close or equal to zero, and the magnitude of the seasonal component appears to be dependent upon the trend level, then pseudo-additive model is most appropriate.

2.2.1.4.1 The Trend. Linear Regression.

The linear regression is a method that determines the straight line that fits better the data in terms of squared error. Given a dataset x_n where $n = \{1, 2, \ldots, N\}$, a linear regression model assumes that the relationship between the dependent variable x_n and the independent variable n is linear.

The expression for the linear regression is

$$x = b_0 + b_1 n \,. \tag{2.9}$$

The coefficients of the equation can be calculated as

$$b_1 = \frac{Cov\,[n, x]}{Var\,[n]} \cong \frac{N \sum_{i=1}^{N} n_i x_i - \sum_{i=1}^{N} n_i \sum_{i=1}^{N} x_i}{N \sum_{i=1}^{N} {n_i}^2 \ (\sum_{i=1}^{N} n_i)^2}, \tag{2.10}$$

$$b_0 = \bar{x} - b_1 \bar{n} \,. \tag{2.11}$$

In Figure 2-9, the original time series is represented, together with a dotted straight line that represents the trend of the data. The trend has been computed with the software MATLAB.

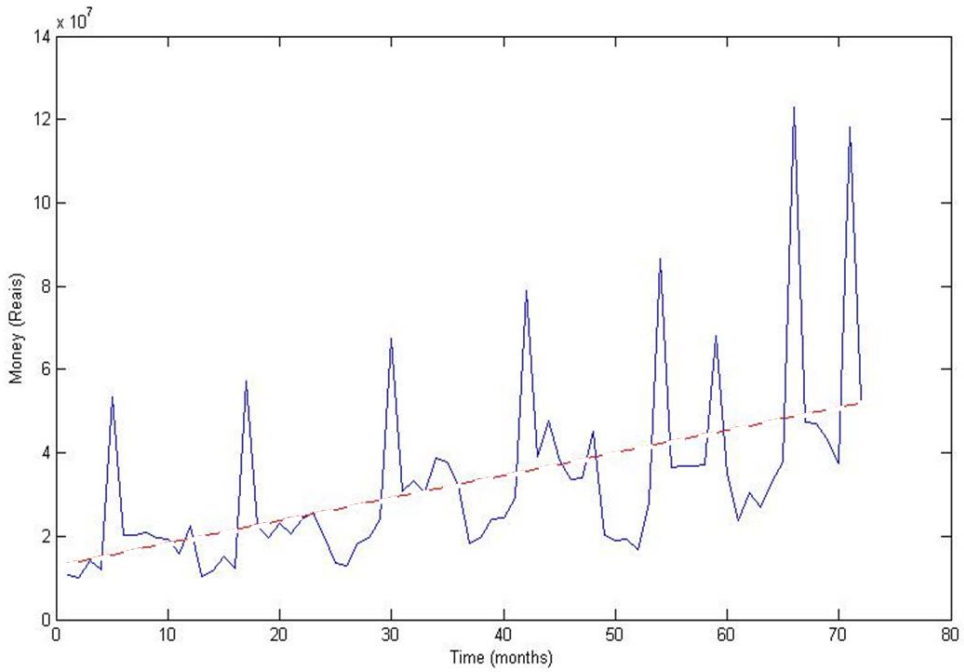

Figure 2-9 Trend and original time series

2.2.1.4.2 Seasonality

Seasonality is a periodic and recurrent pattern caused by factors such as weather, holidays, repeating promotions, as well as the behaviour of economic agents (Hylleberg, 1992).

In other words, a seasonal effect is a systematic and calendar related effect. Some examples include the sharp escalation in most Retail series which occurs around December in response to the Christmas period, or an increase in water consumption in summer due to warmer weather. Other seasonal effects include trading day effects and moving holidays. The number of working or trading days in a given month differs from year to year which will impact upon the level of activity in that month, while the timing of holidays such as Easter varies, so the effects of the holiday will be experienced in different periods each year.

In the case study presented in this project, it is possible to identify a clear seasonal component, being the lasts months of each semester the ones with more tax collection.

In order to identify the stable seasonal component, the first step is to remove the trend of the data. In this case, a 13-term moving average filter is used, because the period of the

time series is 12, as it can be observed on the ACF. Then, we apply a stable seasonal filter to the detrended series. With this filter, we average the detrended data corresponding to each period. That is, average all of the January values (at index 1, 13, 25,...,61), and then average all of the February values (at index 2, 14, 26,...,62), and so on for the remaining months. Finally, the resulting seasonal estimate is centred in order to fluctuate around 0.

Using MATLAB it is easy to identify the seasonal component, which is shown in Figure 2-10.

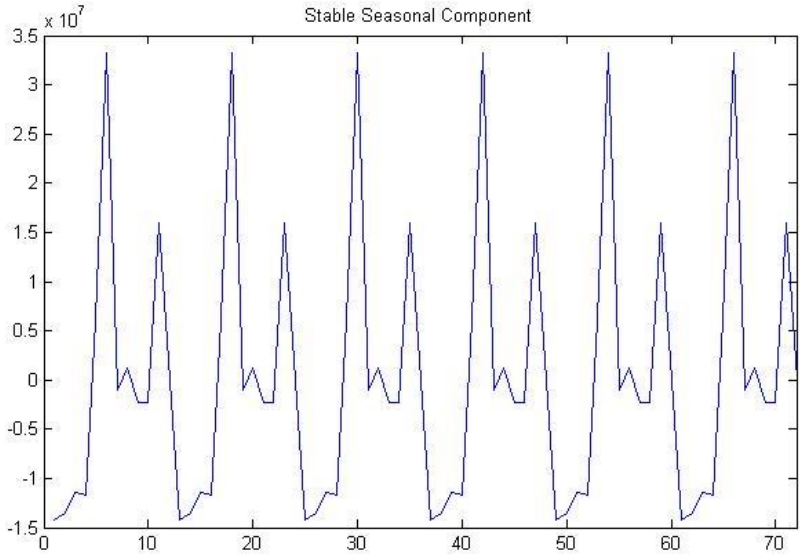

Figure 2-10 Seasonal component

Figure 2-9 and Figure 2-10 represent the stable components of the data series, formed by the Trend and the Seasonal components. The Trend shows that the mean value of the time series in Figure 2-9 is increasing in time, while the seasonal component shows that there is a clear pattern repeated each year, where at the end of the first and second semester of each year is where the tax collection is higher.

A part of the Trend and the Seasonal component, there are Outliers that makes each year different from the previous one and that cannot be detected or predicted by this method.

2.2.2 Artificial Neural Networks

This section contains a theoretical overview about Artificial Neural Networks and how they are used in the case study presented in this project.

2.2.2.1 Introduction to Artificial Neural Networks

An Artificial Neural Network (ANN) is a computational model that is loosely based on the neuron cell structure of the biological nervous system. In most cases a ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. ANN have a potential for intelligent systems because they can learn and adapt, they can approximate nonlinear functions, and they naturally model multivariable systems (Jantzen, 1998).

An ANN is composed of neurons. These neurons are grouped in layers, where the last one is called the output layer, and the previous ones are called hidden layers. The connection of the neurons can be done in different ways, originating different kinds of ANN. In this work, we will focus on Feed-Forward Networks, which are the most widely used for time-series prediction (Frank, Davey, & Hunt, 1999) (Patterson, Chan, & Tan, 1993) (Koskela, 1996).

2.2.2.2 Multi-Layer Perceptron

In Feed-Forward Networks, the information travels only in one direction, from the input to the output, without any feedback or horizontal connection between neurons. The most commonly Feed-Forward Network used is the Multilayer Perceptron (MLP), where all the neurons of a layer are connected to all the neurons of the following layer. The Figure 2-11 shows an example of MLP. The network has three inputs, four units in the first layer, which is called hidden layer, and one unit in the second layer, which is called the output layer.

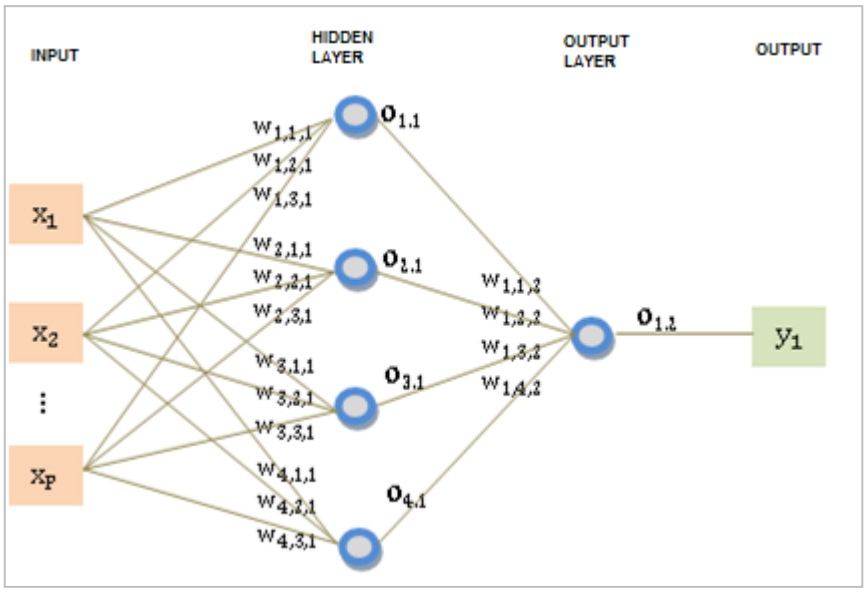

Figure 2-11 A two-layer feed-forward Neural Network

As depicted in Figure 2-11, the network receive the input vector x_n. Each element of the input vector is connected to all the neurons in the hidden layer. In the hidden layer, each neuron assigns a certain weight to each input. According to that, we can define the weight matrix W, containing all the weights for every layer connection.

$$W_{c,p,l} = \begin{bmatrix} w_{1,1,l} & w_{1,2,l} & \dots & w_{1,P,l} \\ w_{2,1,l} & w_{2,2,l} & \dots & w_{2,P,l} \\ w_{C,1,l} & w_{C,2,l} & \dots & w_{C,P,l} \end{bmatrix}, \qquad (1.12)$$

Where the column indices on the elements of matrix $W_{c,p,l}$ indicate the source neuron in the previous layer, and the row indices indicate the destination for that weight in the current layer l. Thus, the indices in $w_{1,2,1}$ refers to the weight connecting the second element of the previous layer, which in the example above is the input vector, with the first element of the actual layer l.

It is considered that the weights that the neurons assigns to their inputs represents the knowledge of the networks, due that these weights will define how important is a certain input.

Taking into account the inputs and the weights, each neuron generates an output. This output will also depend on the activation function of the neuron *h()*, defined later. The Figure 2-12 shows a neuron model in detail.

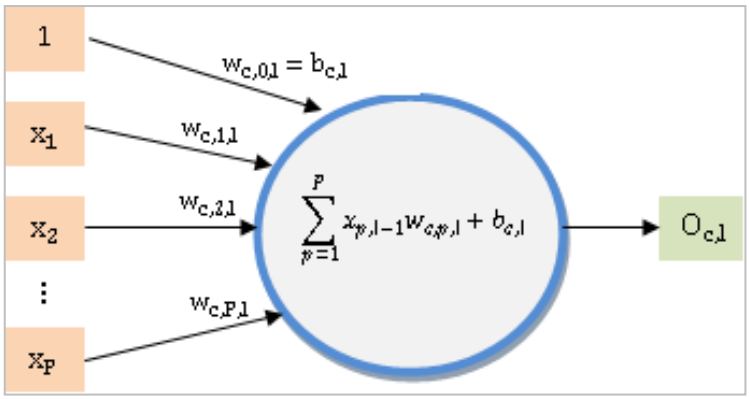

Figure 2-12 Schematic representation of a neuron

The output produced by the *c* neuron of the *l*-th layer, $o_{c,l}(k)$ can be calculated as

$$o_{c,l}(k) = h\big(z_{c,l}(k)\big), \tag{2.13a}$$

where

$$z_{c,l}(k) = \sum_{p=1}^{P} o_{p,l-1}(k) w_{c,p,l}(k) + b_{c,l}(k)\,, \tag{2.14}$$

where k is referred to the iteration or to the input pattern, $o_{p,l-1}(k)$ is the output generated by the neuron p of the previous layer *l-1*, $w_{c,p,l}(k)$ is the weight that connects the neuron p

of the previous layer with the neuron c in the current layer and $b_{c,l}(k)$ is the bias for the neuron c. It is also possible to write the equation in the matrix form:

$$o_{c,l}(k) = h\big(W_{c,l}O_{p,l-1} + b_{c,l}\big). \tag{2.13b}$$

The Figure 2-12 shows how $w_{c,p,l}(k)$ and $o_{c,l}(k)$ are used in the network.

It is evident from (2.13a) and (2.13b) that the output generated by a neuron will depend on its activation function. Neurons can have different activation function in the same network, the more common being the sigmoid function for the hidden layers and the linear function for the output layers. This is also the configuration used in this study. We can define the activation functions $h_{Output}(z)$ for the output layer and $h_{Hidden}(z)$ for the hidden layers as

$$h_{Output}(z) = c_1 z + c_2 \ , \tag{2.15a}$$

$$h_{Hidden}(z) = \mathrm{sgm}(x) = \frac{2c_1}{1 + e^{-c_2 z}} - c_1 \ . \tag{2.15b}$$

In both cases, c_1 and c_2 are constants. The derivatives of the previous equations are:

$$h'_{Output}(z) = c_1 \ , \tag{2.16a}$$

$$h'_{Hidden} = \mathrm{sgd}(z) = \frac{c_2}{2\,c_1}[\,{c_1}^2 - \mathrm{sgm}^2(z)]. \tag{2.16b}$$

In order to build an ANN, several neurons as the one shown in Figure 2-11 can be interconnected.

Once the first output of the network is defined, the network should be able to learn in order to behave as expected, modifying the weights for each input. There are two ways for making the network learn: supervised or unsupervised learning.

In supervised learning a training dataset with the labelled response is given to the classifier. The classifier is trained for obtaining best results in the training dataset and it is supposed that the system will have good generalization properties for non-trained inputs.

In unsupervised learning the classifier is given a set of samples and the classes are unknown a-priori. The algorithm has to learn the classes in the data-set.

Also, inside these two groups, many different learning algorithms have been developed. In this study, the backpropagation learning algorithm, which is a specific implementation of supervised learning, is considered.

2.2.2.3 Learning process: Backpropagation algorithm

Backpropagation is a common method for training artificial neural networks so as to minimize the objective function. The term is an abbreviation for "backward propagation of errors". To make meaningful forecasts, the neural network has to be trained on an appropriate data series. Examples in the form of <input, output> pairs are extracted from the data series, where input and output are vectors equal in size to the number of network inputs and outputs, respectively. A detailed description of the Backpropagation algorithm can be found on (Haykin, 1999).

The error of the output is measured as:

$$e_{c,l}(k) = d_c(k) - y_c(k), \qquad (2.17)$$

Also, we define the error energy E(k) as

$$E_l(k) = \frac{1}{2}\sum_{c=1}^{C} e_c^2(k)\,, \tag{2.18}$$

where $e_{c,l}(k)$ is the error at the output of the neuron *c* in the current layer *l*.

In order to minimize the above objective function, the backpropagation algorithm uses a steepest descent updating, with the gradient calculated from the output layer to the input layer. The main equations of the process are presented below.

The weights will be updated according to the value of $\partial w_{c,p,l}$:

$$\Delta w_{c,p,l}(k) = -\gamma\,\frac{\partial E_l(k)}{\partial w_{c,p,l}}, \tag{2.19}$$

Developing the equation (2.19), we obtain:

$$\Delta w_{c,p,l}(k) = \gamma\, e_{c,l}(k)\, h_{c,l}{}'\left(z_{c,l}(k)\right) o_{p,l-1}(k). \tag{2.20}$$

Where γ is the learning rate, $e_{c,l}(k)$ is the error at the output of the neuron *c* in the current layer *l* and $o_{p,l-1}(k)$ is the output of the neuron *p* in the previous layer *l-1*.

If we define de local gradient $\partial_{c,l}(k)$ as

$$\partial_{c,l}(k) = -\frac{\partial E_l(k)}{\partial z_c(k)} = h_{c,l}{}'\left(z_{c,l}(k)\right) e_{c,l}(k), \tag{2.21}$$

Then, it is possible to write (2.21) as

$$\Delta w_{c,p,l} = \gamma\,\partial_c(k) o_{p,l-1}(k)\,. \tag{2.22}$$

For the output layer, the local gradient can be easily computed as well as obtain directly the error term from (2.17). For the case of hidden layers, the local gradient can be obtained by deriving the equation (2.21). The expression of the local gradient for the hidden layers is

$$\partial_{c,l}(k) = h'_{c,l}\left(z_{c,l}(k)\right)\sum_{n=1}^{N}\partial_{n,l+1}(k)w_{n,c,l+1}(k). \qquad (2.23)$$

Where *n* indicates the neuron unit in the next layer *l+1*, $\partial_{n,l+1}(k)$ is the local gradient of the next layer *l+1*, $w_{n,c,l+1}(k)$is the weight connecting the *c* neuron of the layer *l* with the *n* neuron of the layer *l+1*.

Once the local gradient is computed, the increment of the weights can be calculated with (2.22). The weights of the network in the next iteration of the training process will be updated according to (2.24):

$$w_{c,p,l}(k+1) = w_{c,p,l}(k) + \Delta w_{c,p,l}(k). \qquad (2.24)$$

The Figure 2-13, following the example on Figure 2-11, shows how the Backpropagation Algorithm works.

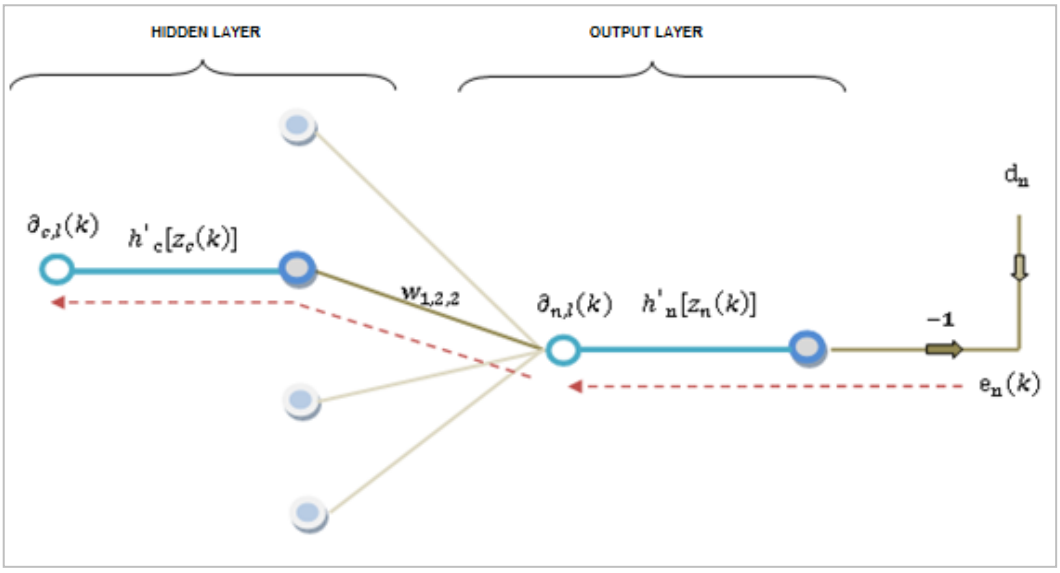

Figure 2-13 Example of Backpropagation algorithm

In Figure 2-13, the error term is computed at the output of the network and it allows calculating $\Delta w_{c,p,l}(k)$ fot the output layer neurons. Then, the error is propagated backwards by computing the local gradient $\partial_{n,l+1}(k)$ and multiplying by the weight connecting to the previous layer neurons. This weight is $w_{1,2,2}(k)$ in the example above.

2.2.2.4 Training process

As exposed along this master's thesis, one of the most important points that determine the performance of an ANN is the training process. In order to select the best configuration for the training process, different tests have been done. Also, different types of ANN have been tested (see Appendix).

Figure 2-14 represents the best configuration found for training the ANN. As shown in the figure, the data is separated by months, so that for making a prediction of one month only the past samples of the same months are used.

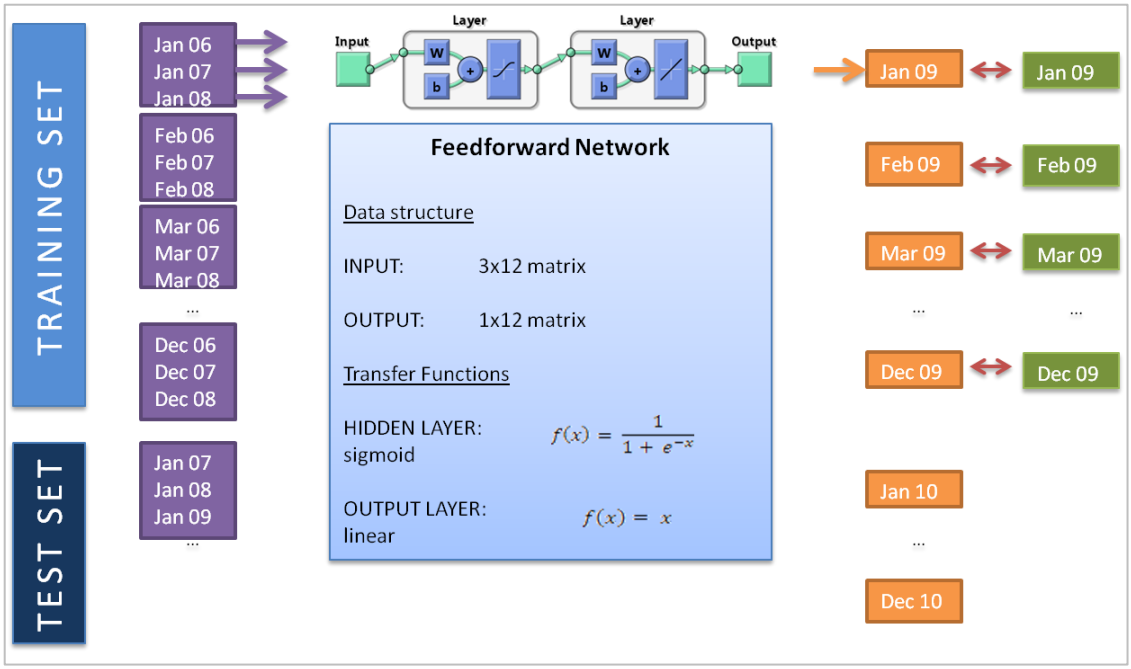

Figure 2-14 Training process for the ANN

In Figure 2-14, the training and testing process is represented. The input vectors contains three elements, and there are only twelve different input patterns for training the network (one per month), which is a really small number for training the network. In order to make a prediction, the algorithm should be executed twelve times, one for each month.

2.2.3 Autoregressive Algorithms

Also, a part of NN, some implementations of the autoregressive algorithms (AR) models have been considered. As well as with ANN, an overview of these models is exposed below.

In temporal series, the Auto-Regressive (AR) model is defined in a way that the variable of interest at time *t* is expressed as a sum of the samples of the same variable as is shown below:

$$y_n = c + \sum_{i=1}^{p} a_i y_{n-i} + \varepsilon_n, \tag{2.25}$$

where P represents the number of steps, also known as the model order, a_i is the coefficient corresponding to the i-th past sample and ϵ_t is the stochastic error component. The equation (2.25) can be also written in the matrix form:

$$\hat{Y} = AX + \varepsilon\,. \tag{2.26}$$

For determining the coefficients A, the Yule-Walker method has been used. This method is based on the direct correspondence between the coefficients and the covariance function of the process. In this way, for each time instant it is possible to define an equation as (2.27)

$$r_x(n) = \sum_{i=1}^{p} a_i\, r_x(n-i)\,. \tag{2.27}$$

The equation (2.28) is the matrix representation of (2.27). Therefore, it can be verified in (2.28) that it is possible to obtain the coefficients a_i by multiplying by the inverse of the covariance matrix from the left side. It is possible to define up to P equations in order to determine all the coefficients:

$$\begin{bmatrix} r_x(1) \\ r_x(2) \\ r_x(3) \\ \vdots \\ r_x(P) \end{bmatrix} = \begin{bmatrix} r_x(0) & r_x(-1) & r_x(-2) & \cdots \\ r_x(1) & r_x(0) & r_x(-1) & \cdots \\ r_x(2) & r_x(1) & r_x(0) & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ r_x(P-1) & r_x(P-2) & r_x(P-3) & \cdots \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix}. \tag{2.28}$$

The equations in (2.28) can be solved using the symmetry of the ACF, which implies that $r_x(n) = r_x(-n)$.

A detailed mathematical analysis of the Yule-Walker equations can be found in (Dimitryou-Fakalou, 2011).

The AR algorithms should be executed many times with the well-known data in order to adapt the value of the A coefficients. Only when the value of these coefficients is adjusted, a real prediction can be done. In the case study presented in this dissertation, the objective is to make a prediction of the lasts twelve months.

Unlike the ANN, two different configurations have been considered. In the first case, the time series is divided in months and only the past sample of a certain month is used for making a prediction of this month. In a second case, all the past samples of all the month are used as a sequence. Also, the optimal value of the model order P has been tested empirically.

2.2.4 Gaussian Process

Gaussian processes belong to the family of stochastic processes schemes that can be used for modelling dependent data observed over time and/or space (Rasmussen & Williams, 2006). In order to be able to make predictions, as in previous cases, a certain function that maps the known past dataset with the prediction outputs is needed. The underlying characteristics of this function can be defined by using Gaussian Process.

A Gaussian process is completely determined by its mean and covariance function, which reduces the amount of parameters to be previously specified since only the first and second order moments of the process are needed. Second, the predicted values are a function of the observed values, where all finite-dimensional distributions set have a multivariate Gaussian distribution.

In a BI environment, the fact that GPR returns a complete statistical description of the predicted variable can add confidence to the final result and help the evaluation of its performance.

Although there are different ways for using GPR, in this work the main interest is regarded to supervised learning, which can be characterized by a function that maps the input-output relationship learned from empirical data, i.e. a training data set. In this study, the output

function is the amount of tax to be collected at any given month by SPU, and hence a continuous random variable.

Since a Gaussian process returns a distribution over functions, each of the infinite points of the function y have a mean and a variance associated with it . The expected or most probable value of y is its mean, whereas the confidence about that value can be derived from its variance.

2.2.4.1 Regression model and inference

Let $S = f\{(x_i, y_i)\}_{i=1}^{m}, x \in \mathbb{R}^n \; and \; y = \mathbb{R}$ be a training set of independent identically distributed (iid) (Pfeifer, 1997) samples from some unknown distribution. In its simplest form, GPR models the output nonlinearly by:

$$y_i = h(x_i) + v_i; i = 1, \dots, m \tag{2.29}$$

where $h(x) \in \mathbb{R}^m$. An additive iid noise variable $v \in \mathbb{R}^m$, with $\aleph(0, \sigma^2)$, is used for noise modelling. Other noise models can be seen in (Murray-Smith & Girard, 2001). Assume a prior distribution over function $h(\cdot)$ being a Gaussian process with zero mean:

$$h(\cdot) \sim GP\ (0, k(\cdot,\cdot),) \tag{2.30}$$

for some valid covariance function $k(\cdot,\cdot)$ and, in addition, let $T = \{(x_i^*, y_i^*)\}_{i=1}^{m}$, $x \in \mathbb{R}^n$ and $y^* \in \mathbb{R}$ be a set of iid testing points drawn from the same unknown distribution S. Defining, for notational purposes:

$$X = \begin{bmatrix} (x_1)^T \\ (x_2)^T \\ \vdots \\ (x_m)^T \end{bmatrix} \in \mathbb{R}^{m\ x\ n};\ X^* = \begin{bmatrix} (x_1^*)^T \\ (x_2^*)^T \\ \vdots \\ (x_m^*)^T \end{bmatrix} \in \mathbb{R}^{m^*\ x\ n} \tag{2.31}$$

and

$$h = \begin{bmatrix} h(x_1)^T \\ h(x_2)^T \\ \vdots \\ h(x_m)^T \end{bmatrix} ; \; v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix} ; \; y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m \tag{2.32}$$

$$h^* = \begin{bmatrix} h(x_1^*)^T \\ h(x_2^*)^T \\ \vdots \\ h(x_m^*)^T \end{bmatrix} ; \; v^* = \begin{bmatrix} v_1^* \\ v_2^* \\ \vdots \\ v_m^* \end{bmatrix} ; \; y^* = \begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_m^* \end{bmatrix} \in \mathbb{R}^{m^*} \tag{2.33}$$

Recalling that, for any function h() drawn from a zero mean Gaussian process prior with covariance function k(;), the marginal distribution over any finite set of input points belonging to X must have a joint multivariate Gaussian distribution:

$$\begin{bmatrix} h \\ h^* \end{bmatrix} | \, X, X^* \sim \aleph \left(0, \begin{bmatrix} K(X,X) & K\,(X,X^*) \\ K\,(X^*,X) & K(X^*,X^*) \end{bmatrix} \right), \tag{2.34}$$

Considering the idd noise model assumed,

$$\begin{bmatrix} v \\ v^* \end{bmatrix} \sim \aleph \left(0, \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \sigma^2 I \end{bmatrix} \right) \tag{2.35}$$

and taking into account that the sum of independent Gaussian random variables are also Gaussians, it yields:

$$\begin{bmatrix} y \\ y^* \end{bmatrix} | \, X, X^* = \begin{bmatrix} h \\ h^* \end{bmatrix} + \begin{bmatrix} v \\ v^* \end{bmatrix} \sim \aleph \, (\mu^1, \Sigma^1), \tag{2.36}$$

Where

$$\mu^{[1]} = 0, \tag{2.37}$$

$$\Sigma^{[1]} = \begin{bmatrix} K(X,X) + \sigma^2 I & K\,(X,X^*) \\ K\,(X^*,X) & K(X^*,X^*) + \sigma^2 I \end{bmatrix} \tag{2.38}$$

Deriving the conditional distribution of y^* results in the predictive equations of GPR:

$$y^* | \, y, X, X^* \sim \aleph\,(\mu^2, \sigma^2) \tag{2.39}$$

Where

$$\mu^{[2]} = K(X^*,X)[K(X,X) + \sigma^2\, I]^{-1} y, \tag{2.40}$$

$$\Sigma^{[2]} = K(X^*,X^*) + \sigma^2\, I - K(X^*,X)[K(X,X) + \sigma^2\, I]^{-1} K(X,X^*). \tag{2.41}$$

Since a Gaussian process returns a distribution over functions, each of the infinite points of the function y^* have a mean and a variance associated with it . The expected or most probable value of y^* is its mean, whereas the confidence about that value can be derived from its variance.

2.2.4.2 Covariance functions and hyperparameters

The power of the Gaussian process to express a rich distribution on functions rests solely on the shoulders of the covariance function (Snoek, Larochelle, & Adams, 2012), if the mean function can be set or assumed to be zero. The covariance function defines similarity between data points and its form determines the possible solutions of GPR (Pérez-Cruz, Vaerenbergh, Murillo-Fuentes, Lázaro-Gredilla, & Santamaria, 2013).

A wide variety of families of covariance functions exists, including squared exponential, polynomial, etc. See (Rasmussen & Williams, 2006) for further details. Each family usually contains a number of free hyperparameters, whose value also need to be determined. Therefore, choosing a covariance function for a particular application involves the tuning of its hyperparameters.

Considering the training SPU data set in Figure 2-4, a pre-processing stage normalized that data set by a mean subtraction - transforming it into a zero mean data set - and an amplitude reduction by a factor of one standard deviation. Thus, the mean function can be set to zero and the focus of the GPR modelling can be fully relied on the covariance function.

Some features of the training data are noticeable by visual inspection, such as the long term rising trend and the periodic component regarding seasonal variations between consecutive years. Taking those characteristics into account, a combination of some well-known covariance functions is proposed in order to achieve a more complex one, which is able to handle those specific data set characteristics.

The uptrend component of the data set was modelled by the following linear covariance function:

$$k_1\,(x, x') = x^T x' \tag{2.42}$$

A closer examination of the data set reveals that, yearly, there is a peak in the tax collection. Additionally, for the years of 2005 and 2006, the peak occurred in the fifth month (May), whereas from 2007 to 2010 the peak occurred in the sixth month (June). The shift of this important data signature makes the seasonal variations not to be exactly periodic. Therefore, the periodic covariance function

$$k_2{}'\,(x, x') = \sigma^2 \exp(-\frac{2\sin^2[\pi/\sigma_2(x - x')]}{\sigma_1^2}) \tag{2.43}$$

was modified by the squared exponential covariance function

$$k_2{}''(x,x') = \exp(-\frac{(x-x')}{2\sigma_3^2}) \tag{2.44}$$

resulting in the following covariance function to model the seasonal variations:

$$\begin{aligned} k_2(x,x') &= k_2' x\, k_2'' \\ &= \sigma_1^2 \exp(-\frac{2\sin^2[\pi/\sigma_2(x-x')]}{\sigma_1^2} - \frac{(x-x')}{2\sigma_3^2}) \end{aligned} \tag{2.45}$$

Finally, the sum of the characteristic components in (2.42) and (2.43) leads to the proposed noiseless covariance function:

$$\begin{aligned} k(x,x') &= k_1(x,x') + k_2(x,x') \\ &= \sigma_1^2 \exp\left(-\frac{2\sin^2\left[\frac{\pi}{\sigma_2(x-x')}\right]}{\sigma_1^2} - \frac{(x-x')}{2\sigma_3^2}\right) + x^T x' \end{aligned} \tag{2.46}$$

In (2.46), the hyperparameter 1 gives the magnitude, or scaling factor, of the covariance function. The σ_1 and σ_3 gives the relative length scale of periodic and squared exponential functions, respectively, and can be interpreted as a "forgetting factor". The shorter the values of $\sigma_{1,3}$, the more uncorrelated two given observations x and x' are. The σ_2, on the other hand, controls the cycle of the periodic component of the covariance function, forcing that underlying function component to repeat itself after σ_2 time indexes.

To complete the modelling profile, the measured noise is assumed to be additive white

Gaussian with variance σ_n^2, which leads to the final noisy covariance function:

$$k(x,x') = k_1(x,x') + k_2(x,x') + \sigma_n^2 I \tag{2.47}$$

$$k\,(x,x') = \sigma_1^2 \exp\left(-\frac{2\sin^2\left[\frac{\pi}{\sigma_2(x-x')}\right]}{\sigma_1^2} - \frac{(x-x')}{2\sigma_3^2} \right) + x^T x' + \sigma_n^2\, I \tag{2.48}$$

Regarding the initial choice of the hyperparameters and its tuning, that learning problem can be viewed as an adaptation of the hyperparameters to a collection of observed data. Two techniques are usual for inference their values in a regression environment: i) the cross-validation and ii) the maximization of the marginal likelihood. As already discussed, GPR can infer the hyperparameters from the training data naturally through a Bayesian framework, unlike other kernel methods such as SVM and KRR that usually rely on cross-validation schemes, which are computational intensive procedures.

Since our observed data possess a trend, splitting it would require some de-trending approach in the pre-processing stage. Also, the number of training data points in this work is small, and the use of cross-validation would lead to an even smaller training set (Rasmussen & Williams, 2006). Therefore, the marginal likelihood maximization was chosen to optimize the hyperparameter's set.

With the covariance function defined in (2.48) and a set of training points given by the first 60 months of the normalized SPU data of Figure 2-4, it is possible to write a multivariate jointly Gaussian distribution using (2.34). This joint distribution can then be conditioned to return the expected value of our random process at a point or a time interval of interest using (2.44).

As well as in previous scenarios, considering the cross-correlation profile of our data shown in Section 2.2.1.2, we will assume that only the amount of tax collected on January of 2005/6/7/8/9 will influence the predictive amount of tax collected in January of 2010, and analogously to the other months.

2.3 New storage technologies: Big Data

Another interesting step of the BI environments to be studied in this master's thesis is the storage of the data. Over the past years, the storage and management of Big Data has become a serious problem not only for BI systems, but the new digital systems in general. According to (Russom, 2011), when the volume of data started to grow exponentially in the early 2000s, storage and processing technologies where overwhelmed by hundreds of terabytes of data. In addition, the heterogeneous nature of the data presents problems that must be considered. This characteristic can be observed in social networks, gene sequences or protein concentrations from cells (Andrew, Huy, & Aditya, 2012). Moreover, improved internet connections and new technologies like smart phones or tablets make it necessary to store and query the data more quickly. For these reasons, organizations and enterprises are becoming more and more interested in Big Data technologies.

Aside from well-known IT companies such as Google[2] and Facebook[3], governments are also interested in Big Data technologies in order to process information related to education, health, energy, urban planning, financial risks and security. Processing all this data makes possible to reduce operating costs and to invest the taxes in a more rational way (Office of Science and Technology Policy of The United States, 2012). Along with other governments; Brazil is starting to employ Big Data technologies in its IT systems.

In this work, we propose the use of Big Data technology to solve the limitations observed on the SIAPE (SIAPE, 2013) database using the Extract, Transform and Load (ETL) tool (Tang J., 2009) and NoSQL data storage. The SIAPE system controls the payroll information of all federal public sector employees in Brazil. Since amount of data on SIAPE is growing at a rate of 16GB per month it can be characterized as Big Data. The SIAPE database is used as case study in this work in order to validate our proposal.

[2] https://developers.google.com/bigquery/

[3] https://www.facebook.com/data/info

2.3.1 Big Data technologies

The literature usually defines Big Data based on the size of the data (Stacy C., 2011). In this work, Big Data is not just defined by the size, but according to (Russom, 2011), which take into account the 3Vs criteria: – Volume, Variety and Velocity. The storage, manipulation and analysis of the enormous datasets are being becoming cheaper and faster than ever with Big Data technologies (Stacy C., 2011). These technologies permit to extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis. In this context, database storage systems have main importance and the development of new databases technologies like Not-only SQL (NoSQL). These databases are characterized by storing and retrieving big quantities of data. In general, NoSQL systems are used on a distributed system that allows offers of scalability and taking profit from multiple nodes.

Unlike traditional relational models, NoSQL does not provide a strict consistency for the data, defined by the Atomicity, Consistency, Isolation and Durability (ACID) features. However, this definition of consistency may be too strict and not necessary in some cases, especially if we want to work in a distributed environment. Instead of that, NoSQL systems are based on the Consistency, Availability and Partition Tolerance (CAP) theory (O'Brien, 2012), which is applied to all the database store systems. One of the most wide used NoSQL databases is Hbase. Hbase implements the column-oriented model as well as Bigtable (Chang F, 2008). Besides that, Hbase essentially offers consistency and partition tolerant according to CAP theorem (O'Brien, 2012).

3 IMPROVEMENT OF THE CGAUD BI ENVIRONMENT

This master's thesis uses the BI environment of the CGAUD as a case study in order to validate the proposed improvements. In this chapter, we study the different solutions adopted on the BI environment of the CGAUD and analyse advantages and drawbacks of each of them. Then, the final adopted solution is explained and the reasons of the choice are exposed. Finally, the most interesting results of the BI system are shown.

3.1 CGAUD BI environment

The CGAUD was previously known as AUDIR, and it is managed directly by the Secretary for Public Management, in Portuguese *Secretaria da Gestão Pública* (SEGEP), which belongs to the Brazilian Ministry of Planning, Budget and Management, in Portuguese *Ministério do Planejamento, Orçamento e Gestão* (MP).

One of the main responsibilities of the CGAUD is to audit the payroll of the federal public sector staff. This payroll is monthly elaborated, and includes the information about the active, retired and pensioners of the Federal Public Administration. 16 Gb of information are generated each month, containing the data about 6,200,000 public servants and summing a total value around R$ 12.5 billion. The basic element of the payroll is called rubric. Each rubric stands for a positive or negative value that is included on the payroll according to the characteristics of the position of each public employee. There are 2,200 different rubrics.

The information of the payroll is stored on the database called Integrated System for the Administration of Human Resources, in Portuguese Sistema Integrado de Administração de Recursos Humanos (SIAPE).

The legal base that regulates the payroll is the Federal Constitution of Brazil, although there are some other specific laws and decrees that complement it. According to this legislation, the responsibility of the CGAUD department is to analyse the rubrics of every payroll in order to detect incompatibility of benefits, inconsistencies and irregularities.

This auditing process is done using audit trails. An audit trail consists on a set of rules based on the legislation that has the goal to identify the irregularities on the data. Once the inconsistency on the payment is detected, it will be notified to the auditors in order to take the required actions to fix the situation.

3.2 Original Audit Process

Before the deployment of the solution proposed in (Campos, et al., 2012) and (Fernandes, et al., Construction of Ontologies by using Concept Maps: a Study Case of Business Intelligence for the Federal Property Department, 2012), the audit process was manually performed. The payrolls were generated and the salary was transferred to the employees, the CGAUD personnel run a query on the payroll database with a set of filters that maps the current legislation. The result set containing the possible irregular cases was exported to an Excel document where each register was analysed in detail by the auditors as shown in Figure 3-1.

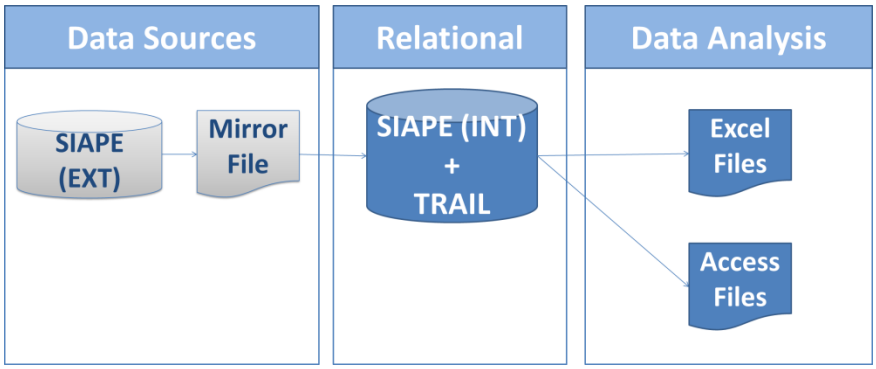

Figure 3-1. Original Audit process

Such method had several limitations that implied a low efficiency on the audit process. The first and more evident is the fact that it was executed after the payment was done. In case an irregularity was detected, another bureaucratic process was required to recover the money paid due to this irregularity. Also, this manual process required an intensive human intervention, which implies the expenses of human resources and possible errors on data manipulation. Finally, the presentation of the data and the possibility of elaborating statistics and historic views were very limited.

Considering this scenario, a BI system with preventive control, automatic process and access to the statistics has been proposed in (Campos, et al., 2012). The proposed BI system should be able to detect irregularities before the payment was done. In addition, the audit process in the BI system should have almost no human intervention in order to reduce the expenses in human resources and to increase its reliability. Finally, a different and clearer presentation of the data, including different types of graphics, lists, reports and dashboards should be added.

In order to address these requirements, in this work, we propose the implementation of a BI solution using the open source software suite Pentaho. The details for the design and implementation of the proposed solution are exposed in the following section.

3.3 The previous BI Proposed Solution

In a first approach for the solution, it was proposed a BI system considering the federal payroll information and the audit trails in order to automatically monitor the audit process.

The first step of the conception of a BI system is the identification and definition of the indicators required by the decision-making managers. In a previous work described in (Fernandes A. , 2012), it is shown how this process can be improved by the use of concept maps. This approach was applied to the concept of Audit Trails. Figure 3-2 shows the BI system considered in (Campos, et al., 2012) and (Fernandes, et al., Construction of Ontologies by using Concept Maps: a Study Case of Business Intelligence for the Federal Property Department, 2012) based on concept maps.

Once the indicators have been defined, a BI solution for providing that information to the decision-making responsible manager is built. However, from the sources of information to the final reports, the data have to be processed in several steps.

As shown in Figure 3-2, the payroll data is generated and stored on the SIAPE database. Each row of the database is identified by the field called *matricula_servidor*, a unique id for each public sector staff member. Moreover, the information contained on the SIAPE database is divided in two groups: personal information and financial information. This information is sent to the CGAUD department through a mirror file.

The 16 Gb data is received every month, and it is directly loaded onto a relational database. This relational database contains also the result of the execution of the audit trails.

This data is extracted from the data sources through an Extract-Transform-Load (ETL) process (Do & Rahm, 2000), and passes to the staging area. In this step, the data is transformed and prepared for being loaded on the Data Warehouse (DW).

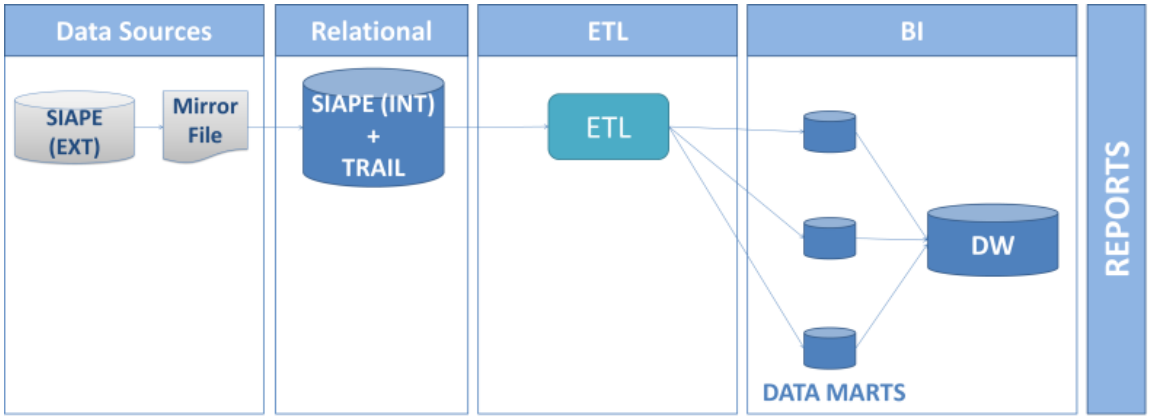

Figure 3-2 Previous BI Solution (Fernandes, et al., Construction of Ontologies by using Concept Maps: a Study Case of Business Intelligence for the Federal Property Department, 2012) (Campos, et al., 2012)

A general problem of ETL tools is their limited interoperability due to proprietary application programming interfaces (API) and proprietary metadata formats making it difficult to combine the functionality of several tools (Oracle, 2005). This problem has been solved by using the Pentaho Data Integration (PDI) tool. Its connection is based on Java Database Connectivity (JDBC) technology that establishes a set of classes and interfaces (API) in Java that allows sending and receiving SQL instructions from any relational database.

The BI step is composed by a Data Warehouse (DW) and several Data Marts (DM). A DW is a database used for reporting and data analysis. It is a central repository of data which is created by integrating data from one or more different sources. Data warehouses store current as well as historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons. On the other hand, DM are smaller slices of the DW containing the data of a specific area of the system, according to the semantics of the application under concern.

Although the previous solution solved several problems of the original system, new requirements were identified and solved in Section 3.3.

.

3.4 The proposed improved BI solution for the new CGAUD scenario

Based on the new requirements, a new BI solution was needed to fulfil the above mentioned needs while keeping the enhancements provided by the previous works. In the following the details of this new solution are presented.

3.4.1 The proposed SIGAWEB application

One of the limitations of previous BI solution was the low efficiency on the tracking of the detected irregularities or inconsistencies. Therefore, we propose here the SIGA application, which supports the auditor along the tracking process of an irregularity or an inconsistency detected on the Audit Trails.

According to Figure 3-3, the implemented SIGA functionalities are (1) Audit Trails, (2) Document Management, (3) Process Registration, (4) Process Monitoring, (5) Exception Register, (6) Reimbursement Tracking and (7) Reporting.

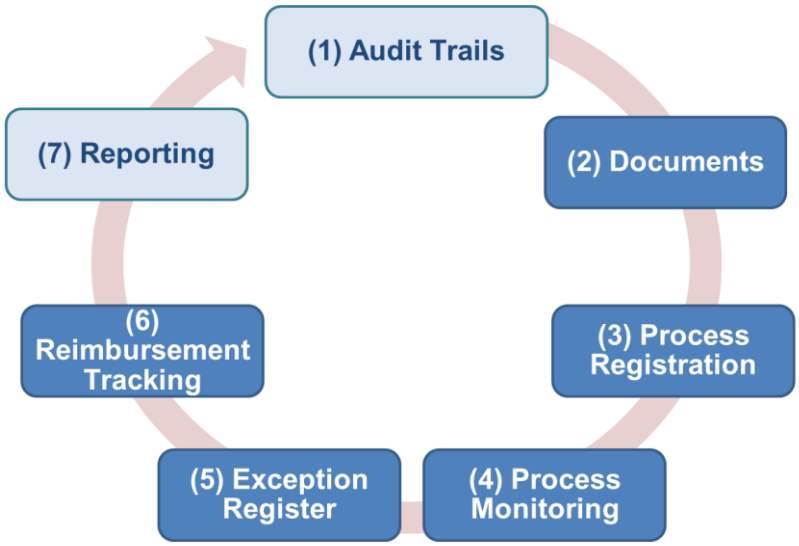

Figure 3-3 The proposed SIGAWEB structure

As shown in Figure 3-3, the process starts with the execution of the Audit Trails (1). Once an irregularity is detected, several documents have to be created in order to register it and

notify the department charged of the payment of the public employee. These documents can be automatically created and easily monitored through the Documents module (2) of the SIGA. With the detected irregularity and the correspondent documents, a new process will be created into the application using the Process registration module (3).

Also, for each irregularity an internal process on SIGA is created. This process can be linked to one or several documents and is related to the starting user. As the audit process requires several analyses by different auditors, there is a Process Monitoring module (4) that allows changing the responsible user, the status of the process and tracks the historical evolution of the process.

Continuing with the flux on Figure 3-3, at the end of the process, it is possible that the analysed public employee's payroll is in an exceptional situation that allows him or her to receive all the rubrics into the payroll, even if it was previously detected as irregular. The Exception Register module (5) permits to register those cases and avoids detecting it again on the next analysis of the next month.

In the case that the process is confirmed as irregular, the public employee will have to refund the received value to the public treasury. In this case, the Reimbursement Tracking module (6) will register this process and monitor it until the reimbursement is finished. This module is of special interest as it allows monitoring the complete reimbursement process, something that was very difficult, and also provides a prediction of when the reimbursement will be finished as well as the age the public employee when the reimbursement is completed.

3.4.2 The improved BI solution

Due to the useful amount of the data generated by the SIGA, the BI solution included integration with the SIGA allowing evaluating the efficiency of monitoring of the processes carried out into the MP. Therefore, our proposed BI solution has two data sources: the SIAPE database and the SIGA database as shown in Figure 3-4.

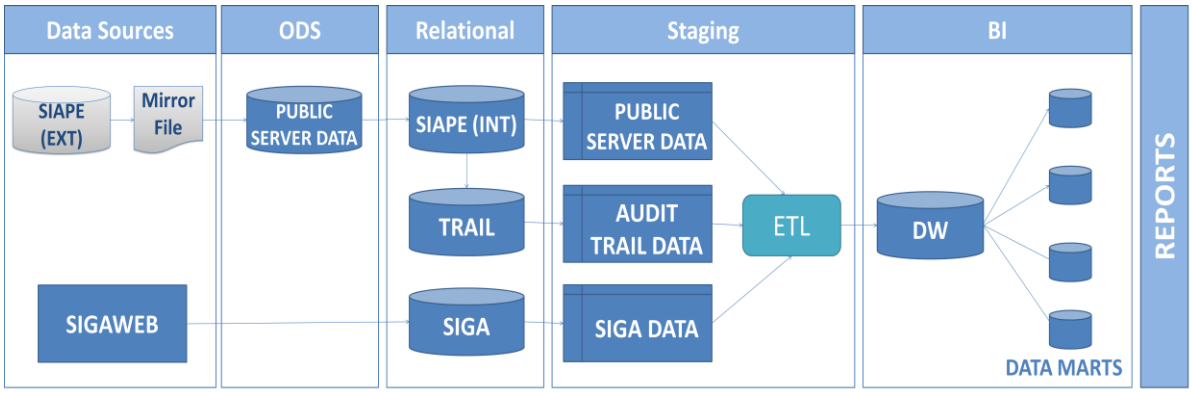

Figure 3-4 Proposed BI Solution

In this improved solution, the data received from the SIAPE is actually loaded onto an Operational Data Store (ODS), as shown in Figure 3-4. The ODS contains all the registers of the mirror file and becomes the internal data source. The inclusion of the ODS facilitates the operational access to the data coming from external entities and provides more reliability to the loading process of the data.

From the ODS, the personal and financial information about the public sector staff is loaded on a relational database called internal SIAPE where the information is consolidated and more reliability and consistency is guaranteed. As implemented in the previous solution, the TRAIL database contains the result of the execution of the audit trails.

Also, two other databases are used in this relational step: SIGA and TRAIL. The SIGA database contains the information generated by the SIGA software.

DM and DW relations can be configured into two manners: top-down and bottom-up. In the first case, exposed on Figure 3-4, several DM are created with the information coming from different sources and then joined in order to build the DW. Top-down environments are architected to deliver precise answers to predefined questions. On the other hand, in bottom-up environments like the one used in Figure 3-4, all the data is first integrated into the DW and then subdivided in several DM.

In this master's thesis, a top-down structure has been used. First, a DW is built with the data obtained on the ETL process, and from this, four DM are created: Audit trails, Restitution to the treasury, Payroll and SIGA.

In a previous work presented in (Campos, et al., 2012), a bottom-up structure was used. However, it showed some limitations mainly due the increasing volume of the data, the addition of new modules of the system with specific indicators and the necessity of increasing the granularity on the final reports. In the previous system, the use of detailed reports including information of the individual public employees implied a too long loading time, lasting around 20 minutes for reporting the data. In this new scenario, where precise and detailed reports in four different knowledge areas are required, the top-down structure has been observed as the most efficient, lasting less than 15 seconds for loading the most detailed reports.

Finally, the information loaded on the four DM can be easily accessed from the Pentaho Report Designer and used to create different types of outputs such as Static Reports, Dashboards, Maps or Web Services.

.

3.5 BI Reports. Presentation of the results

According to the policy of the Brazilian government of using Open Source software solutions, several Federal Entities and Ministries have chosen the Suite Pentaho for implementing their BI solutions. The platform is developed in Java, which allows the system to run in different platforms and can be easily integrated with other existing or even self-developed solutions. In line with this trend, and considering the advantages of this software, the Suite Pentaho has been used for developing the proposed solution in this project. In the following, some of the most relevant results are presented.

Figure 3-5 and Figure 3-6 show the evolution of the volume of the audit process per month, in terms of number of cases and value in BRL, respectively. In both cases we observe a quite stable tendency on time, with two evident exceptions on June and November. This is due to the 13th payment, an extra payment that all the public sector staff receives in Brazil, which divided in two parts.

Also, in comparison with the results presented in (Campos, et al., 2012), in which the value of the audited rubric is 1,5 billion, there is a clear increase in the audited value. This is due to the new approach to define irregularities as well as the capability to manage bigger volumes of data.

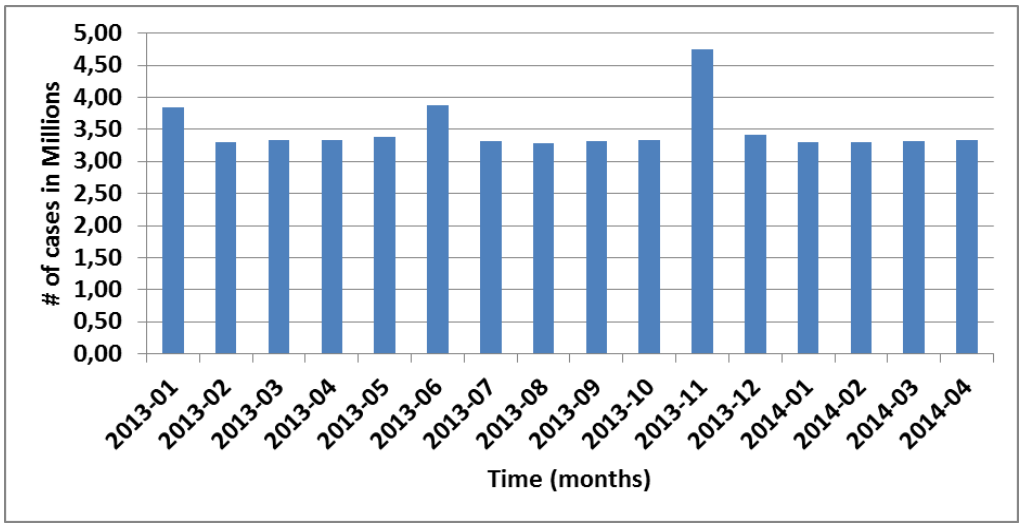

Figure 3-5. Incompatibility of Rubrics, evolution of the audited quantity per month.

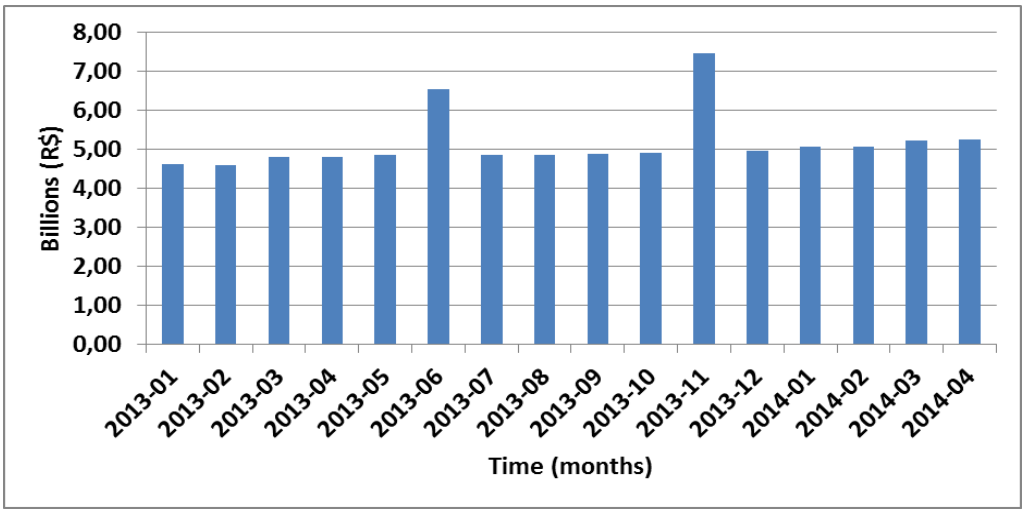

Figure 3-6. Incompatibility of Rubrics, evolution of the audited value per month.

Figure 3-7 and Figure 3-8 show the evolution of the detected irregular cases per month. When an irregular case is detected, the responsible or the audit analyses the case and notifies the correspondent department in order to avoid the irregular payment on the next month. In this case it is possible to observe a clear decreasing tendency, which demonstrates the efficiency of the proposed system in identifying irregularities, which can be fixed and do not appear in the following months

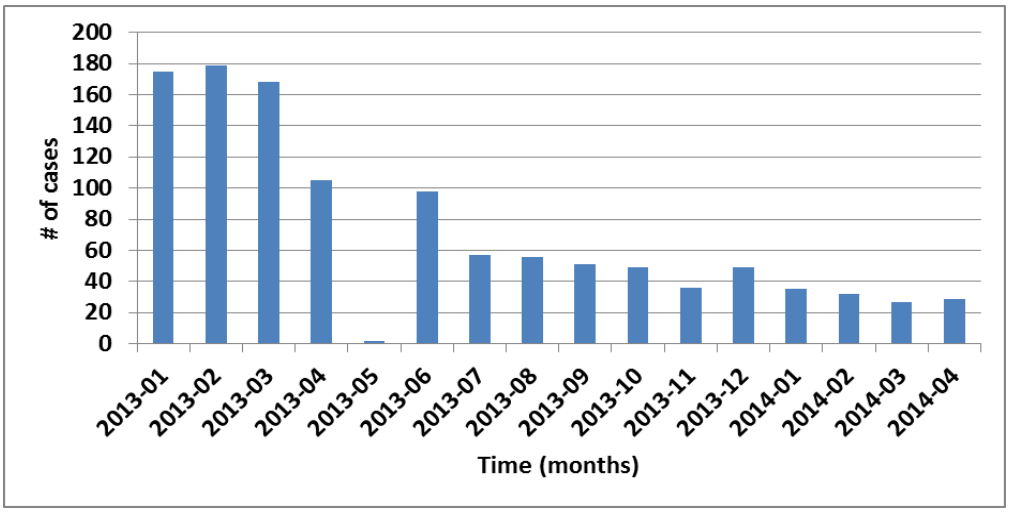

Figure 3-7. Incompatibility of Rubrics, evolution of the amount of irregular cases per month.

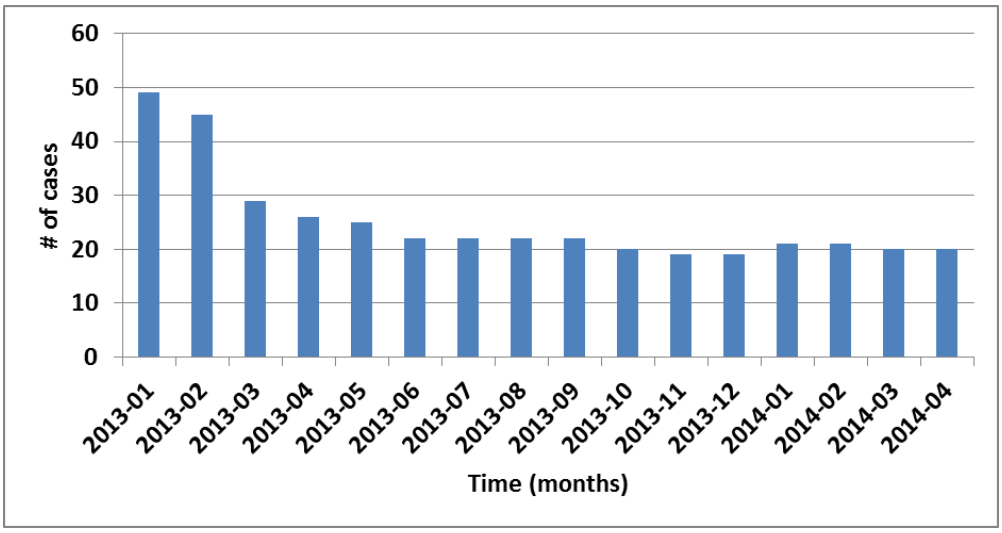

Figure 3-8. Extra salary for alimentation

Figure 3-9 and Figure 3-10 show the evolution in time of the number of cases and value of the restitution to the treasury. In this case it is possible to observe a very irregular evolution because the restitution to the treasury is a complex juridical process. This means that even that the system correctly detects the irregularity, the restitution is not automatically, but it depends on the analysis of a juridical process before the money actually comes back to the treasury.

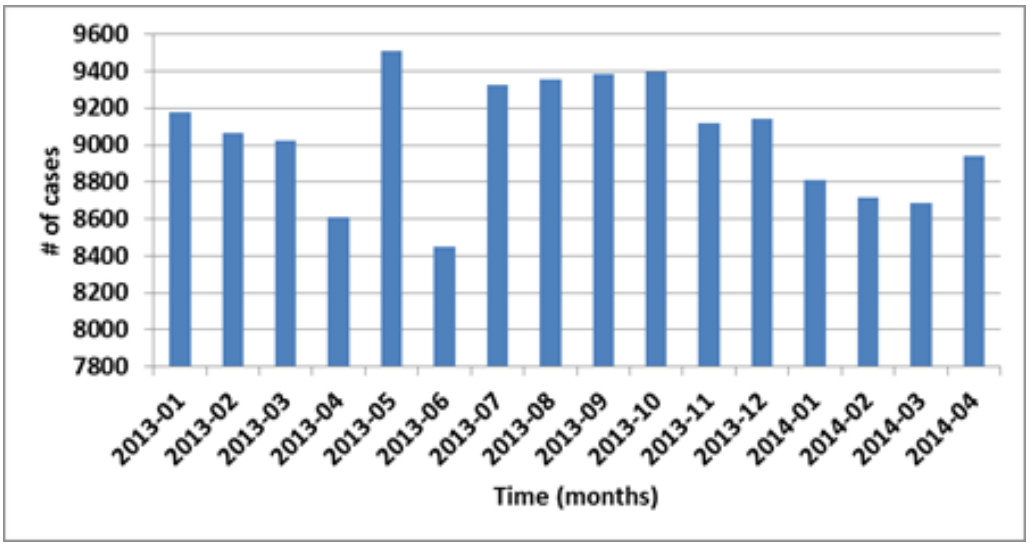

Figure 3-9. Restitution to the public treasury, audited quantity

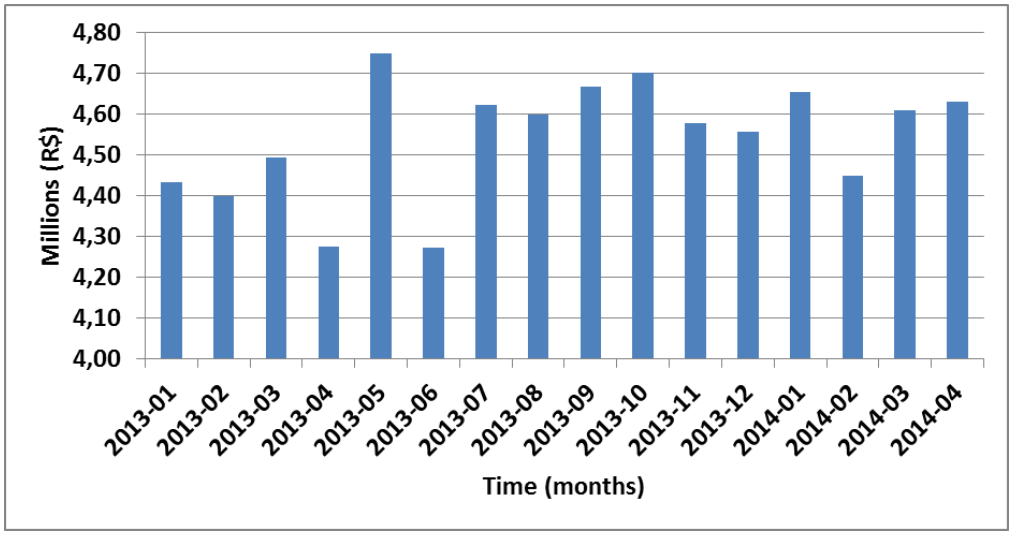

Figure 3-10. Restitution to the public treasury, audited value

Finally, Figure 3-11 shows an example of an analytical report. It is possible to observe the high level of customization on the presentation of the granularity provided by this new solution. As an example, it is possible to analyse results from all staff working in different areas until an individual public sector staff member, by adjusting the column Servidor. This feature allows retrieving very detailed information without losing performance.

					Measures
Grupo Assunto	Orgao	Servidor	Servidor Unico	Situacao Processo	Quantidade Registros
Todos Grupos Assunto	Todos Orgaos	Todos Servidores	Todos Servidores Unico	Todas Situacoes Processo	5.692
ABATE TETO	Todos Orgaos	Todos Servidores	Todos Servidores Unico	Todas Situacoes Processo	2
				Pronto Para Importação	2
ACUMULAÇÃO	Todos Orgaos	Todos Servidores	Todos Servidores Unico	Todas Situacoes Processo	391
				Pronto Para Importação	391
ACÓRDÃO TCU	Todos Orgaos	Todos Servidores	Todos Servidores Unico	Todas Situacoes Processo	99
				Concluído	99
ADICIONAIS	Todos Orgaos	Todos Servidores	Todos Servidores Unico	Todas Situacoes Processo	199
				Cadastro de Exceção	99
				Pronto Para Importação	100
CADASTRO	Todos Orgaos	Todos Servidores	Todos Servidores Unico	Todas Situacoes Processo	96
CONSIGNAÇÃO	Todos Orgaos	Todos Servidores	Todos Servidores Unico	Todas Situacoes Processo	288
TRILHAS DE AUDITORIA	Todos Orgaos	Todos Servidores	Todos Servidores Unico	Todas Situacoes Processo	4.617

Figure 3-11. Analytical report.

4 IMPROVING PREDICTIVE ANALYTICS FEATURES FOR THE CGAUD

In this chapter, the proposal of application of predictive analytics into the BI environment at CGAUD is presented. First, two parameters are proposed as automatic evaluation methods for assess the quality of the predictions: the Normalized Root Mean Square Error and the Coefficient of Determination. Then, the most relevant results of the case study performed with the CGAUD data are exposed. Finally, two methodologies of application of predictive analytics are proposed: (1) an automatic fraud detection system based on an ANN predictor, and (2) a new architecture for incorporating predictive analytics with automatic algorithm selection and configuration into an ETL process.

4.1 Prediction evaluation methods

When a prediction is done, it is normal and useful to plot the predicted and the expected series, in order to compare them by visual inspection. This gives us a subjective evaluation of the results. However, it is much more desirable to have a more quantitative measure of how good the prediction is (Drossu & Obradovic, 1996). For this reason, in this study two indicators are considered:

- A normalized Root Mean Square Error

$$\mathrm{NRMSE} = \sqrt{\frac{1}{N}\frac{\sum_{n=0}^{12}(x_n - y_n)^2}{\sum_{n=0}^{12}(x_n)^2}}. \tag{4.1}$$

- Coefficient of determination

$$\mathrm{COD} = 1 - \frac{\sum_{n=0}^{12}(x_n - y_n)^2}{\sum_{n=0}^{12}(x_n - \bar{x}_n)^2}. \tag{4.2}$$

In both equations, x_n and X are referred to the actual or real values and vector respectively, while y_n and Y are referred to the predicted values and vector. In the first case, the optimum value would be NRMSE = 0, while for the second case we desire $\mathrm{COD} = 1$.

In general, the measurement of the NRMSE is the most used manner for computing the goodness of a prediction. The NRMSE gives us an idea about how the two time series are

similar on the mean value. However, in this project the objective is to make predictions of a complete year with a resolution of months. Thus, to know that the mean value of the two series is similar is not enough information, due to the fact that the mean values can be very similar but not the values of each month separately.

Therefore, we need also to consider the shape of the predictions before saying if a prediction is good or not. In order to make this comparison, we use the COD, which provides information about how the variance of the two series is similar.

As exposed in the Case study section, the use of these two indicators is needed for making a useful comparison between the prediction and the real values.

4.2 Case Study with SPU data. Tests and Results

In this chapter, a case study using the data of the MP is presented. The objective is to compare the performance of different algorithms and its possible configurations in order to obtain a prediction with the maximum accuracy, with the aim of discover whether these algorithms can be successfully applied to the CGAUD environment or not.

The tests have been done using the data series containing the amount of tax collection per month by the MP expressed in BRL. Moreover, the tests have been done using two different software: MATLAB[4] and WEKA[5]. The first is a well-known powerful mathematical software developed by MathWorks, while the second one is an open source suite of machine learning algorithms developed by the University of Waikato (New Zealand).

[4] MATLAB is a numerical computing environment and fourth-generation programming language developed by MathWorks. It allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages. (See: www.mathworks.com).

[5] Weka is an open source software developed by the Univ. Of Waikato that contains a collection of machine learning algorithms for data mining tasks. (See: www.cs.waikato.ac.nz/ml/weka/).

Several tests using ANN and ARX with different software have been carried out on (Rubio Serrano A. M., Study and implementation of a predictive analytics module for the Business Intelligence System of the Brazilian Ministry of Planning, Budgeting and Management, 2012). In this work, only the most interesting results are exposed in Section 5.7.1.1 and Section 5.7.1.2, while other algorithms and applications has been incorporated

The results of these tests are exposed in the following sections.

4.2.1.1 Algorithm: Multilayer Perceptron

In order to discover the optimal configuration for the Neural Network, several tests has been done varying the number of layers and the number of neurons per layer. The complete set of test is shown in (Rubio Serrano A. M., Study and implementation of a predictive analytics module for the Business Intelligence System of the Brazilian Ministry of Planning, Budgeting and Management, 2012). Figure 4-1 and Figure 4-2 contain the best results obtained with Matlab and Weka respectively.

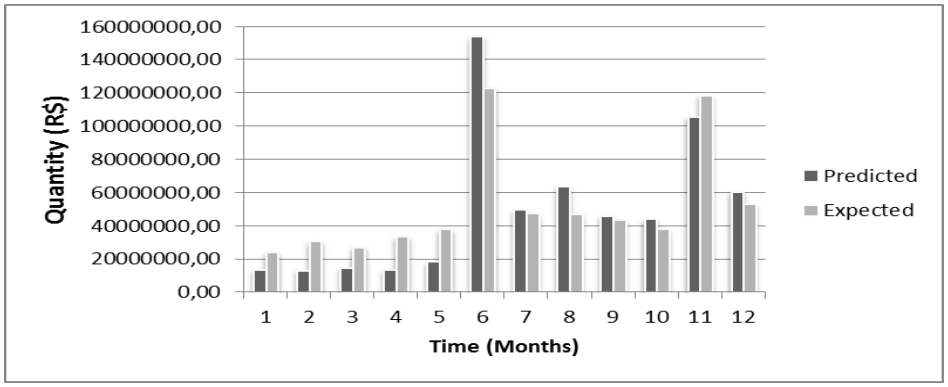

Figure 4-1 MATLAB. 8-1 MLP. NRMSE=0.25, COD=0.76.

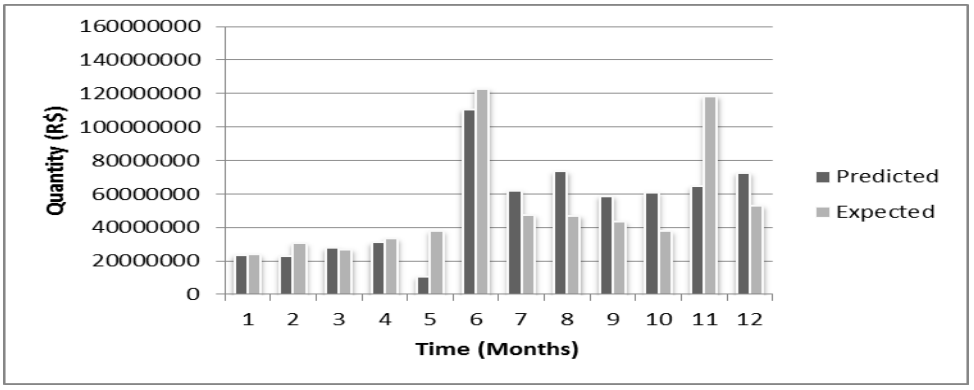

Figure 4-2 WEKA. 8-1 MLP. NRMSE=0.25, COD=0.56

4.2.1.2 Autoregressive Models

The following graphs show the most significant results for the AR models.

In a first test, the model has been applied to each month separately. It means, considering the past samples of January for making a prediction of January, using the past samples of February for making a prediction of February and so on. With this configuration, the best result was obtained with one past sample ($p = 1$). The results are shown in Figure 4-3.

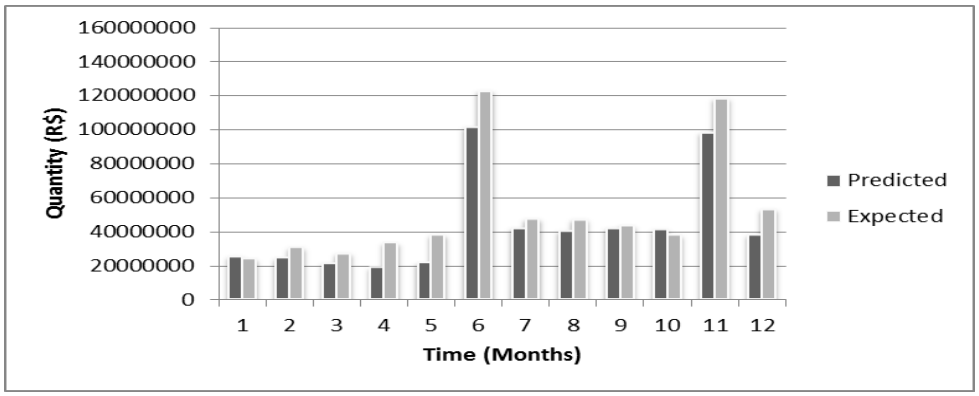

Figure 4-3 Prediction and error using ARX, month by month and $p = 1$. NRMSE=0.18, COD=0.85.

In another test, presented in Figure 4-4, the past samples of all the months together are used. In this case, the best result was obtained for the case of $p = 16$.

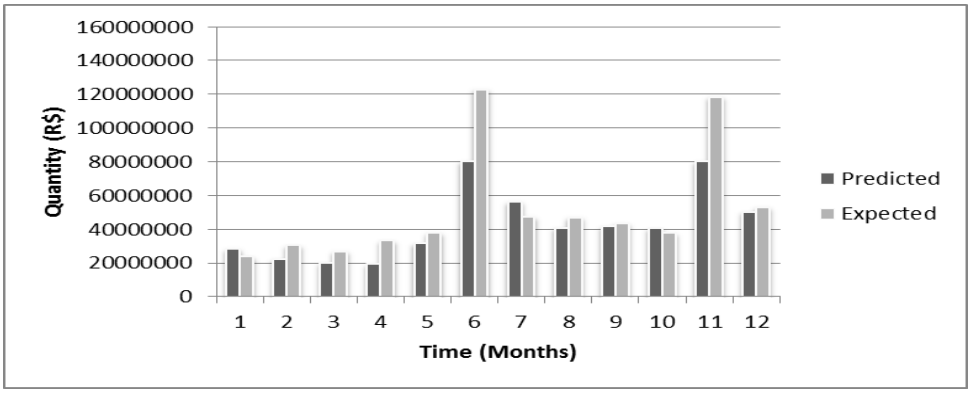

Figure 4-4 Prediction and error using ARX, considering all the months and $p = 16$. NRMSE=0.17, COD=0.67.

According to (Rubio Serrano A. M., Study and implementation of a predictive analytics module for the Business Intelligence System of the Brazilian Ministry of Planning, Budgeting and Management, 2012), quite good results can be obtained with ARX when the data has a more linear behaviour. However, even in this scenario, ANN are able to achieve better predictive capacity.

4.2.1.3 Gaussian Process

Figure 4-5 shows a plot of the predicted values using the optimized hyperparameters, where it can be seen that the uncertainty of May's prediction is quite higher, mainly because the tax collection profile changed drastically in the training data. This behaviour contradicts the linear increasing trend that was used to model the covariance function, since the linear regression of this specific month shows a clear downtrend. However, in spite of the uncertainty level, the prediction of this month turned out to be precise.

Also, it can be noted that November was the only month whose target value fell of the uncertainty predictive interval delimited in this section. In spite the fact that the predicted value is larger than the last year's value for this month, the rate of growth from 2009 to 2010 could not be estimated by this model based only on the information of the training data.

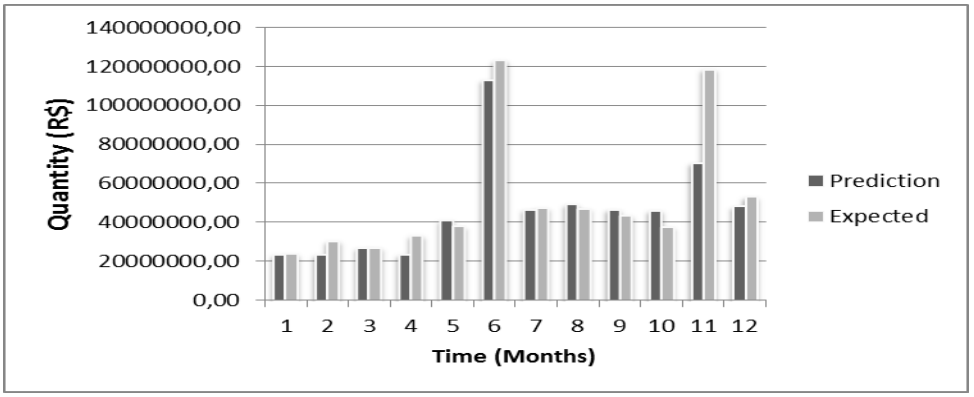

Figure 4-5 Prediction using Gaussian Process. NRMSE = 0.24, COD=0.78.

4.2.2 Analysis of the results

The graph in Figure 4-6 shows a comparison between the best predictions with Neural Networks and AR algorithms.

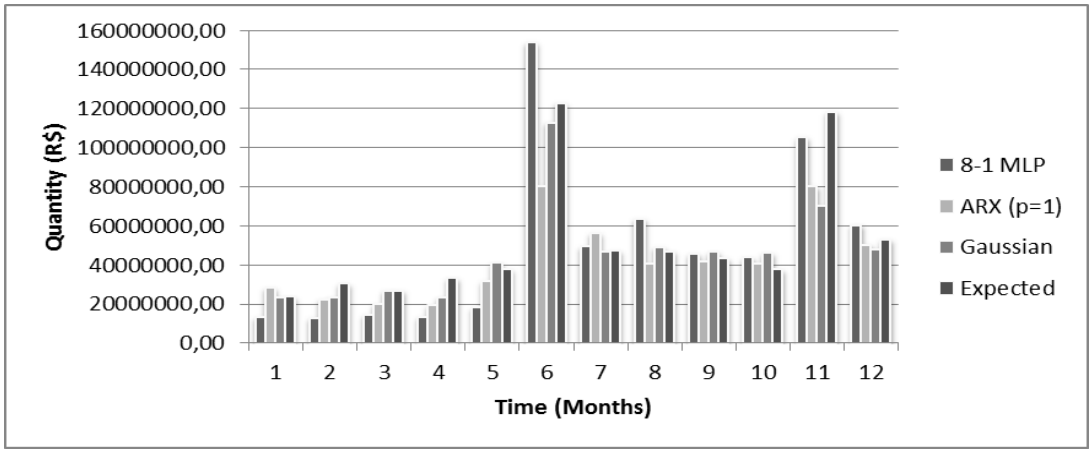

Figure 4-6 Comparison of the results

Also, the best results obtained with each algorithm are summarized on Table 4-1.

Table 4-1 Best results obtained with each algorithm

Algorithm	NRMSE	COD
ARX	0.18	0.85
ANN	0.25	0.76
GPR	0.24	0.78

From these observations it is possible to conclude that the three algorithms has a good performance and can be used as a predictor in this context, although ARX may be a bit superior when using only the previous month as an input.

However, if it is desired to apply this solution to other environments, it is important to notice that the performance of the algorithms will vary according to each type of data. In this way, it is necessary to look for a more general solution (Rubio Serrano A. M., Study and implementation of a predictive analytics module for the Business Intelligence System of the Brazilian Ministry of Planning, Budgeting and Management, 2012).

Interestingly, ANN shows very good performance in this scenario with low number of samples. These conclusions are consistent with (Zhang, 2003). According to the mentioned paper, an important advantage of the ANN with respect other algorithms is that the ANN

are universal approximators that can approximate different kinds of functions with high precision. However, the optimal configuration of the network will vary a lot depending on the kind of data, and it is important to be aware to the very high sensibility to the initial conditions (Rubio Serrano A. M., Study and implementation of a predictive analytics module for the Business Intelligence System of the Brazilian Ministry of Planning, Budgeting and Management, 2012).

Also GPR showed acceptable performance for prediction with low number of training samples. The hyperparameters of GPR's covariance function were optimized by maximum likelihood, i.e. the proposed model let the data speaks for itself by learning the hyperparameters only with information obtained from the data.

It is relevant to notice that the optimization algorithm can converge to a local minimum, making the initial choice of hyperparameters a critical part of the optimization task. Another positive point of GPR is related to the complete statistical description of the predicted data, which gives an powerful tool of confidence. Using this feature, a classification method can be built to trigger trusted and possibly fraudulent tax collection data based on the confidence interval of the prediction.

With respect to the software, both Weka and Matlab have been able to make very accurate predictions. The differences on the results for the same data and same configuration can be explained by the different manner of initializing the algorithm. In this way, we conclude that both softwares are useful for implementing a predictor based on a MLP. Weka has the advantage that it is distributed under an open source license and is written in Java, which makes very simple to extract functions and integrate it into the Pentaho environment.

4.3 Proposal of Fraud Detection System

In this chapter the application of the ANN predictor applied for fraud detection over indicators that can be modelled as a time series is presented. This proposal has been also presented in (Rubio Serrano A. , da Costa, Cardonha, & de Sousa Jr., 2012).

4.3.1 Current approach to fraud detection

Fraud has been very common in our society, and affects private enterprises as well as public entities. However, in recent years, the development of new technologies have also provided criminals more sophisticated way to commit fraud and requires more advanced techniques to detect and prevent such events.

One way to detect fraud is to apply statistical fraud detection schemes. In general, statistical methods can be classified as supervised or unsupervised. The supervised techniques require both fraudulent and non-fraudulent samples in order to construct models that allow classifying future behaviour patterns. On the other hand, unsupervised methods simply seek those accounts, customers and so forth which are most dissimilar from the norm. One drawback of most of unsupervised techniques is that fraudsters adapt to new prevention and detection measures. In this sense, fraud detection needs to be adaptive and to evolve over time in order to gradually change their behaviour over some period of time and to avoid spurious alarms (R.J. Bolton, 2002).

Nowadays there are different types of solutions for facing fraud. Traditional statistical classification is still applied with a high detection probability (R.J. Bolton, 2002) (Hand, 1985). For instance, rule-based methods are classifiers based on a set of conditions with the form *If{condition},Then{consequence}* that represents the behaviour of the fraudulent actions. Examples of these algorithms are BAYES (R.J. Bolton, 2002) (P. Clark, 1989), FOIL (R.J. Bolton, 2002) (Quinlan, 1990) and RIPPER (R.J. Bolton, 2002) (Cohen, 1995).

Also, Artificial Neural Networks (ANN) have been applied for fraud detection, mainly in the context of supervised classification. In (Dorronsoro, Ginel, Sanchez, & Cruz, 1997), ANN is applied for credit card fraud detection.

4.3.2 Proposed solution

In this dissertation, we propose the use of a neural network-based predictor to identify possible fraudulent patterns in the Federal Patrimony Department, in Portuguese *Secretaria de Patrimônio da União* (SPU). Our proposed method is based on the fact that the artificial neural network (ANN) can be used in the recognition of statistical characteristics on a time series and make predictions.

In the case of an ANN classifier, the ANN is used to classify an input into a group from a set of predefined groups (for example, fraudulent and non-fraudulent). In the case of ANN predictor, the ANN returns the prediction of a certain input data. The advantages and the limitations of both systems are exemplified through the case study presented below, that uses the tax collection data of the SPU.

For our case study, we use data from the SPU starting from 2005 until 2010 and divided into months. The simulations try to detect if 2010 is a fraudulent year or not based on the predicted data.

We propose the use of ANN prediction algorithms applied to fraud detection in time series data. Our ANN makes the prediction and the real results of this year are compared with the prediction. If the results presents a great difference with the prediction it will mean that this sample should be deeply investigated. As shown in our simulations results, the proposed methodology based on ANN offers a very high accuracy for predicting time series.

The block diagram of the proposed methodology is shown in Figure 4-7

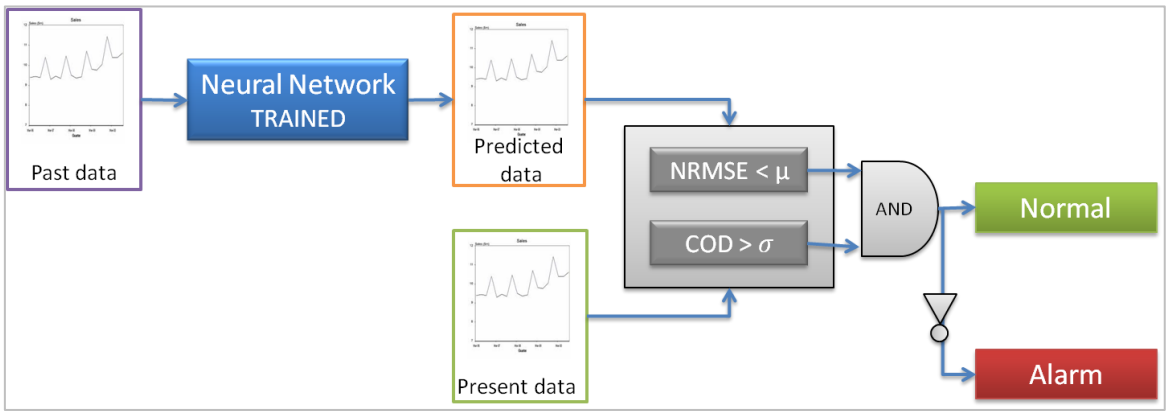

Figure 4-7 Schematic representation of the proposed solution for fraud detection

4.3.3 Experimental results

In this section, the experimental results are presented. First, we search empirically the best ANN configuration, i.e. with best number of neurons and of layers, for our data. We find out that the Multi-Layer Perceptron of two layers, containing 8 neurons on the hidden layer and 1 neuron on the output layer, presents the smallest error.

In Figure 4-8, the proposed error metrics are applied to compare the predicted and the actual data, obtaining an NRMSE = 25 % and a COD = 0.76. Therefore the similarity is very high between the predicted and the actual data. Such similarity can be also easily visualized.

Once the good performance of the chosen ANN configuration is confirmed, the real data is corrupted in order to simulate the fraudulent data in the year 2010. We corrupt the data by adding a 10 % error distributed over all the months of the year.

In Figure 4-9, the real and predicted time series are still very similar. The NRMSE = 34% and the COD = 0.86. Depending on the threshold level, this fraud could be detected. For instance, if the threshold for the NRMSE is 10 %, the fraud is detected, if the threshold is higher the fraud will not be detected.

In Figure 4-10, the 10 % error is concentrated on the third month only, the fraud becomes visible. In this case, we obtain a NRMSE of 40% and a COD equal to 0.32.

In Figure 4-11, we divide a 15 % error through four months randomly. First, we consider a difference of 15% on the overall year, with the increments divided in 4 months. The NRMSE = 28% and the COD = 0.33. Due to the values of NRMSE and COD, most probably frauds have taken place in this year.

Another interesting test is presented in Figure 4-12. In this case, there is no increase or decrease on the overall value, but a redistribution of the values among the year. This anomaly will also be detected using the COD (0.38) and the NRMSE (25%).

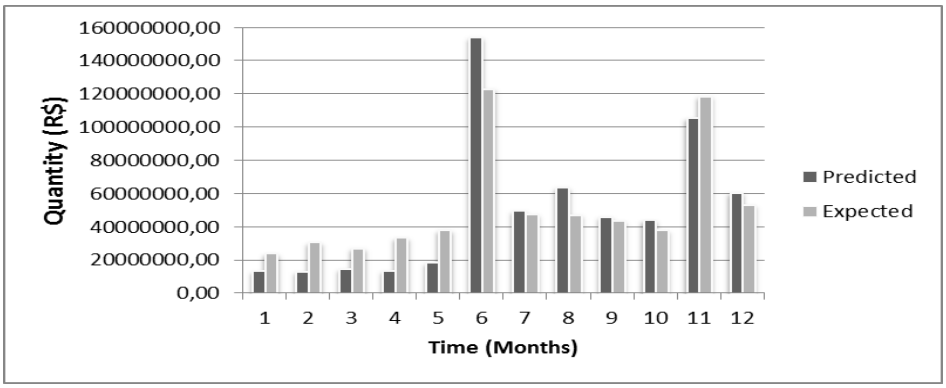

Figure 4-8 Predicted and Expected values using an 8-1 MLP. NRMSE = 25% and COD = 0.76.

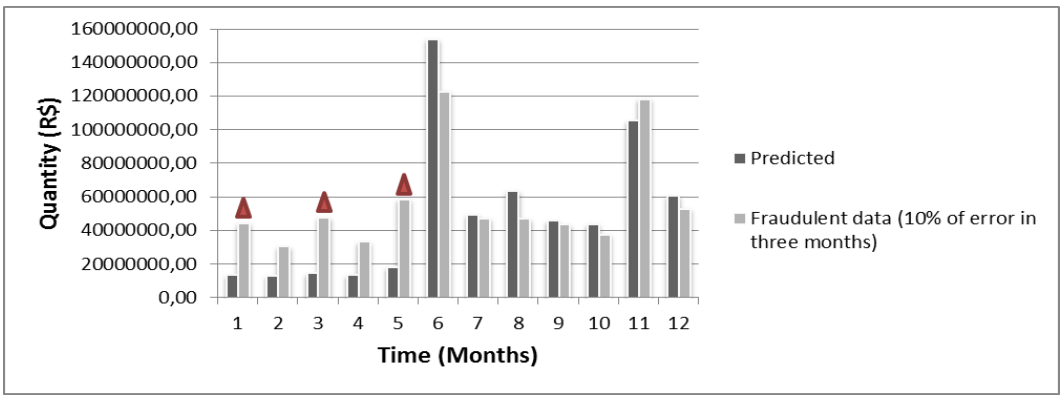

Figure 4-9 10% of error distributed in 3 months. NRMSE = 34% and COD=0.86.

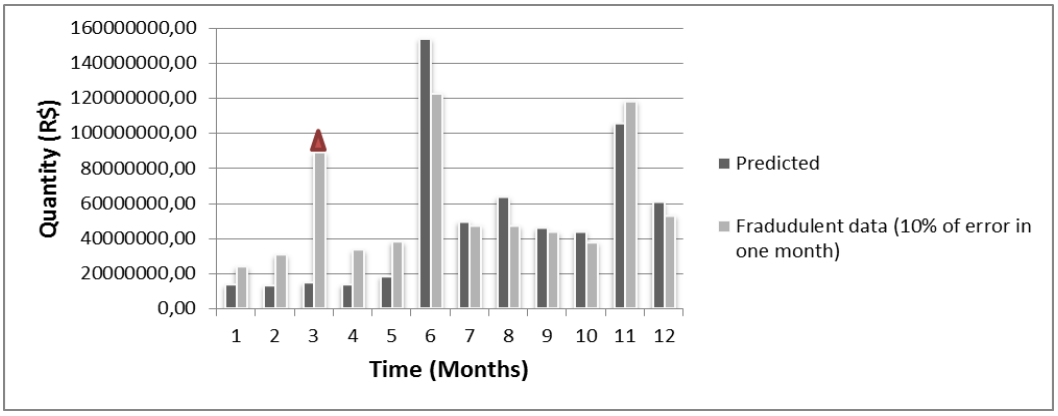

Figure 4-10 10% of error in 1 month. NRMSE = 40% and COD=0.32.

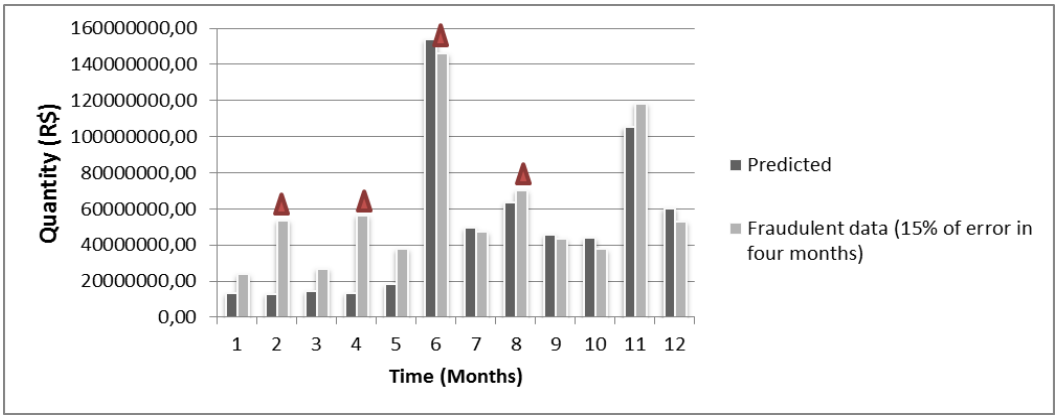

Figure 4-11 15% of error distributed in 4 months. NRMSE=28% and COD=0.33.

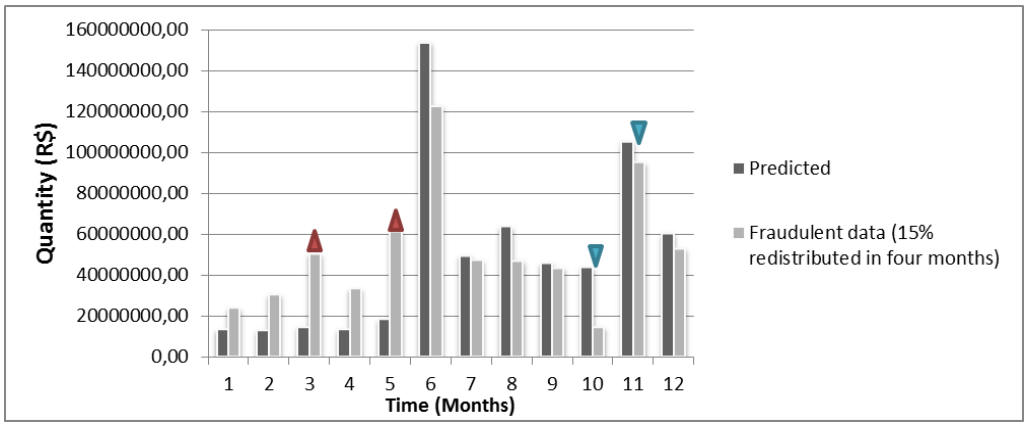

Figure 4-12 15% redistributed in different months. NRMSE=38% and COD=0.38.

Comparing the proposed solution with the traditional ANN classifier, it is possible to observe the better performance of ANN predictor for this scenario. In a traditional ANN classifier, instead of predicting the data, the ANN gives a hard output, which can be '1' for fraudulent sample and '0' for non-fraudulent samples. Due to the small number of samples available as shown in Figure 2-4, the traditional ANN has not been able to correctly classify fraudulent and non-fraudulent samples. Therefore, for small number of data, our proposed scheme should be applied for fraud detection.

4.3.4 Analysis of the results

In this chapter, we have proposed the application of ANN for time series prediction together with NRMSE and COD in order to detect frauds in financial systems.

Our proposed solution is particularly interesting when only a small number of samples are available. Since the traditional solution does not work for such scenarios, our proposed scheme should be applied. Therefore, our method based on ANN predictor should be considered as an alternative to ANN classifiers when the data is a time series which is evolving with time and when the number of sample for training the network is small.

To validate our study, we have considered data from the tax collection of the SPU.

4.4 Proposal of a new methodology for BI systems: ETAPL

The application of the predictive analytics studies exposed in previous section to the BI systems is of special interest. In this work, it is proposed a new approach to include the predictive analytics into the Extract, Transform and Load (ETL) process.

The traditional ETL process aims to extract data from several heterogeneous storage systems, transform it accorder to business rules previously established and load it into a Data Warehouse (DW) or a Data Mart (DM) for a posterior use of this data.

Traditional BI systems already incorporate some predictive functionalities. Usually, BI systems incorporate a prediction module that is executed after the ETL process, using the data loaded on the OLAP storage systems (Chaudhuri, Dayal, & Vivek, 2011). A standard BI process is shown in Figure 4-13.

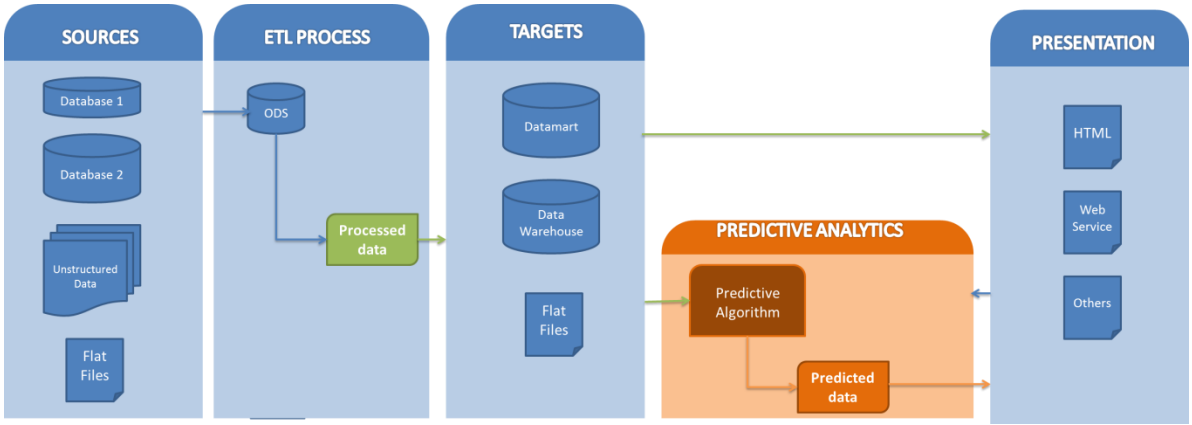

Figure 4-13 Traditional BI architecture with predictive analytics module

This kind of systems, although useful for some scenarios, has some drawbacks that should be considered. First, the predictive module is executed on-demand, when the users select it from the user interface. It implies a delay due to the recovering and processing of the data.

Also, in most of the cases there is a limitation of the capabilities of the predictive analytics. This is because while some BI systems only offers one or two predictive algorithms, other has a more powerful prediction module but requires a complex manipulation and configuration of it, what difficult its use by non-specialized people. This is the case of the Pentaho suite used in the CGAUD and several other departments among the MP. Pentaho incorporates the software Weka, which contains a huge set of prediction and classification

algorithms. However, its configuration and use will require considerable technical and mathematical skills, what normally does not match with the profile of the users of BI applications.

In order to solve those limitations, the Extract, Transform, Algorithm-Selection, Prediction and Load (ETAPL) method is proposed. This method is based on the incorporation of the predictive analytics into the traditional ETL process, as well as the use of prediction evaluation methods exposed in Chapter 4.2 for automatically select and configure the best algorithm for each data series. The proposed ETAPL method is shown in Figure 4-14.

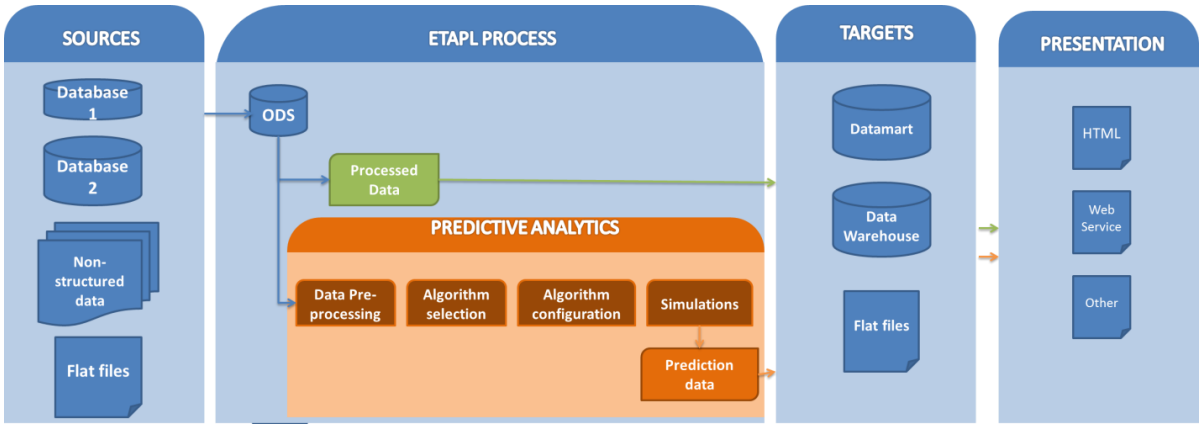

Figure 4-14 Proposed ETAPL methodology

In the proposed solution, the predictive analytics module is executed in parallel with the traditional ETL process.

First, the data is pre-processed in order to be prepared for the prediction algorithm. As shown in Chapter 2.2, remove the trend and to use of the Auto Correlation Function is suitable in order to better adapt to the characteristics of the data.

Then, in the algorithm selection phase, several algorithms are executed in order to make predictions of a subset of the known data (known as testing set). The accuracy of the predictions will be evaluated using the NRMSE and the COD. The algorithm with the best result will be selected for that type of data.

In the same way, it should be considered that each algorithm may have several configuration parameters. The algorithm configuration will be adjusted using the same method.

Once the best algorithm and its best configuration is chosen, the prediction of the future data is computed. Finally, this prediction data is loaded on the DW and DM.

This proposed method allows solving the limitations exposed below on traditional BI systems. First, the incorporation of the predictive module on the ETL process allows loading the predicted data into the DW and the DM, making it available when required by the user. It results on a faster BI system, avoiding the delay of computing the prediction each time the user requires it. Second, the use of the prediction evaluation methods allows to select and optimize the predictive algorithms for each type of data automatically, increasing the accuracy of the predictions without require human intervention.

5 NEW DATA STORAGE SYSTEMS: BIG DATA

In this chapter, the case study for using Big Data technologies as main database into the BI environment of the CGAUD is presented. The objective is to evaluate the viability of using the new Big Data technologies as main database as well as to compare its performance with the current relational database.

5.1 Case Study

The case study presented in this work consists of the implementation of a NoSQL storage system using Hadoop[6] and Hbase[7] technologies. The performance of the proposed solution will be compared to the PostgreSQL (PostgreSQL, 2013) implementation that is currently used on the SIAPE database.

The hardware setup used for the relational database system consists of a Dell Inc. PowerEdge R610 with a 2xCPU Xeon 2.80GHz, 32GB RAM and a HDD of 500GB. The Database Management System (DBMS) used in this case is the PostgreSQL v8.4 optimized for this hardware.

In the case of the NoSQL approach, two types of configurations are used: the Master Server and the Region Server. A single system has been configured as a Master Server, while three systems are used as Region Server.

In the remainder of this section we explain, first, how the file is formatted before beginning with the loading data and, secondly, the process of loading data after the file formatting process. Finally, we discuss some aspects related to data storage.

[6] http://hadoop.apache.org/

[7] http://hbase.apache.org/

5.2 SIAPE File/Database

The SIAPE (SIAPE, 2013) is a national system to manage payroll of Brazilian federal workers. This system manages every paying unit at national level. Besides this, it ensures the availability of data on the page siapenet.gov.br. In this context, the SIAPE file contains a sequential copy of SIAPE database for each month including personal, functional and financial data of federal public workers in Brazil. This file has information of about two and half million workers, among them – active, inactive and retired. The actual size of each SIAPE file per month is about 16GB and is growing every month. Besides that, this file contains 212 fields between personal, functional and financial data. The current state of the database that imports this data is:

- More than 27 million rows of the public workers table.
- Financial data table has about 200 million rows.

This information covers 2012 exclusively, so in both short-term and long-term time frames this database will have serious problems for storing and querying data.

5.3 Modelling

The first challenge was to adapt the relational model of the SIAPE database (See Figure 5-1(a)) to NoSQL (HBase) model. The relational model of SIAPE database is composed mainly by the tables “servidor_historico”, “servidor_dado_financeiro”, auxiliary tables and the related indexes for the main queries. For NoSQL model it was defined the “Worker” table and its row key composed by the fields “Year/Month”, ”MatSiape”, “CodOrgao”, “Upage” and “SeqSdf” to guarantee a unique identifier for every employee. The creation of this row key was defined to take advantage of HBase data structure. This row key was created according to the more common questions to be answered in the queries.

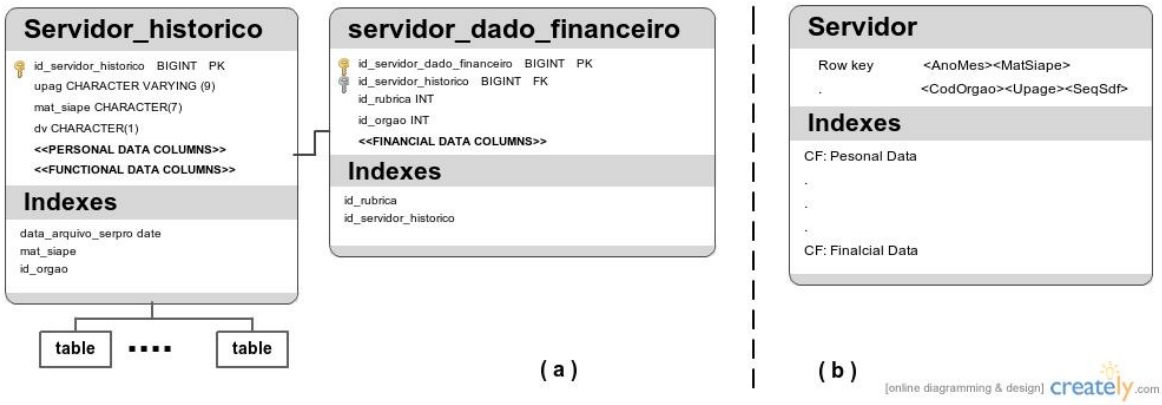

Figure 5-1 (a) Postgres structure for SIAPE Database and (b) HBase Database Structure for SIAPE Database.

In our implementation, two column families were defined: "Personal Data" and "Financial Data". We define "Personal Data" as a column family to group all columns related to the employee's personal data, while the "Financial Data" column family is defined to store the employee's financial data. In this sense, the last attribute (SeqSdf) of the row key contains a sequence value to represent every financial data of a specific employee. Besides this, the use of column families in Hbase for storing data allows the disk usage optimization. The data structure designed for store SIAPE data in the HBase system is shown in Figure 5-1(b).

5.4 Implementation

First, we chose Cloudera Hadoop Distribution v4.0 (CHD4) for easy use and free distribution. So it was configured CHD4, Zookeeper (only for the Master Server), HDFS and Hbase. Second, we defined two steps for loading the SIAPE file: formatting the SIAPE file in a CSV format and load the CSV files using the ETL Pentaho Data Integration (PDI) (Sergio & María, 2011).

To format, we use a shell script to separate the personal and functional data in a CSV file called "servidor". Besides that, another CSV file contains financial data of employees called "servidor_dados_financeiros". The "servidor" file has 192 fields. After the formatting stage this file contains almost 2.4 million rows (or employees and their personal and functional data). The "servidor_dados_financeiros" file has 20 fields and after formatting, this file contains almost 20 million rows.

To load: we use the PDI (Sergio & María, 2011) to load data into “Worker” Hbase table for “servidor” and “servidor_dados_financeiros” data. Figure 5-2 (a) and (b) show the sequence followed by the loading process for the “Worker” table in the Hbase database. In the first step, we treat the CSV file of “servidor” concatenating the year and month fields. In the Second step, it was generated the row key for every employee with the fields Year/Month, MatSiape, CodOrgao, Upage and SeqSdf. In the third step were removed unnecessary or temporary fields. Finally, we insert into the “Worker” Hbase table. For financial data it follows the same sequence as showed in Figure 5-2(b).

It is important to be emphasized that formatting and loading processes happen every month when the SIAPE file is provided to the human resources department.

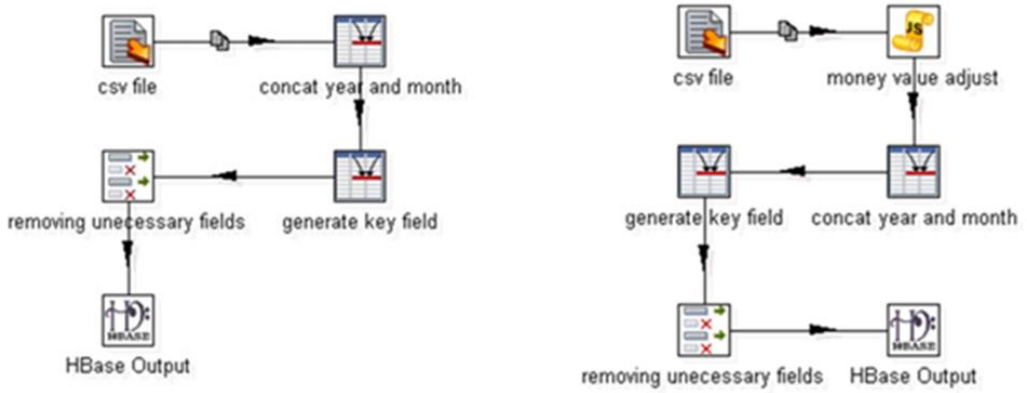

Figure 5-2 (a) Loading Data Sequence for Personal Data; (b)Loading Data Sequence for Financial Data.

5.5 Results and Discussion

The loading process was repeatedly tested using the resulting CSV file. This process lasted about 30.0 minutes with Postgres and 05.18 minutes with Pentaho/Hbase. This means that Pentaho/Hbase was 82.72% more quickly than Postgres database. Besides that, a brief comparison of querying data is shown in Table 5-1. The latency is shown when querying data using a relational and NoSQL (Hbase) technologies.

Table 5-1 Result of querying data

	Pentaho-HBase	Postgres	Improvement%
Single result by key	16 ms	23 ms	30.43
All data about a worker filtered by month and year	17 ms	35 ms	51.43
All data filtered by month and year	0.440472 min	24.9 min	98.23
Audit Trail: Incompatibility of rubric	8.7833 min	26.666 min	67.06

Comparing the Hbase and Postgres results for loading data we notice that Hbase technology obtained better results for many reasons. One reason is because Hbase does not use the ACID principle. While Postgres spends more time and resources trying to fit the ACID properties, the Hbase is interested in fitting consistency and partition tolerance according to CAP theorem (Vora, 2011). This is less expensive in terms of time and resources (O'Brien, 2012). Another aspect that may help to reduce the loading time was to use the HDFS to load the SIAPE formatting file. This means before beginning with loading we store the CSV file into the HSDF with the objective of reducing the O/I during the loading process.

The query time tests obtained good results also for Hbase technology. The used model influences the speed of query. First, the Hbase model uses the key to get an entire row (value). Thus, the way you build the key will impact the speed of your query. Secondly, the Hbase key table is ordered, so if the key is built with the more used fields the query responses will be faster. This is because every row is stored in an ordered fashion according to the used key and the probability of the row being stored in the same HDFS DataNode is very high. This will help to query data very quickly because the more fields (more used field to query data) contained in the key the faster the query will be. This is the case unless the number of fields is large because this may compromise the query performance. And finally, the Hbase columns are lexicographically oriented. This allows "scan" operations very fast in a specific range without the necessity of using secondary indexes.

The Hbase works differently from a relational database. It is a column oriented database where the columns are lexicographically oriented. It allows one to make scan operations

within a specific range in a more direct and rapid way, without the necessity of using secondary indexes.

Additionally it is worth to point that the results can be improved by tuning the data store, for instance, via table scan. Currently the scan is done in a sequence fashion. This configuration can be done in a parallel fashion for every DataNode. This can deliver the data processing to every DataNode and it will improve the scan table. Furthermore, the scalable property of Hbase helps manage data growth that will be expected in the future years.

5.6 Discussion

It is undeniable that Big Data is going to grow in importance in all fields. In this context we examine a relatively big database (SIAPE database), which is actually facing storage, process and query problems. In addition, in the short-term and the long-term the current system will be untenable. In this work is shown how the performance improves by using Big Data technologies. In this way, we improved the loading and querying data using Pentaho Kettle and Hbase as a data storage. The results showed an improved performance in terms of time for loading data in the SIAPE database and improving the latency for querying data. Moreover, we notice that good modelling of data can help to get good query latency.

In the future we intend to test the loading data in a larger cluster including desktop computers with less processing and storage capacities. A comparison with other NoSQL technologies can be done. This will show Hbase performance comparing it to other data storage like Cassandra (Cassandra , 2013).

6 CONCLUSION

In this project, an analysis and improvement for a BI solution has been proposed in three main areas: data loading for a BI system, predictive analytics and data storage using Big Data. The conclusions for each subject are exposed below.

6.1 BI environment

In the present work, we proposed an improved BI solution for the CGAUD department of the MP. The previous BI system was only focused on the case of the Audit Trails for a limited amount of data.

In this project, a new software module including the complete Reimbursement Tracking System has been proposed.

Moreover, an improved BI solution based on a top-down structure has been proposed for fulfilling the new requirements of the project: to manage a bigger volume of data, including the restitution to the treasury module on the BI and providing more customization on the granularity of the final reports.

The results were validated using the real data of the public sector staff payroll from the Brazilian Ministry of Planning, Budget and Management, showing evidence of the efficiency in providing valuable and accurate data that converts in enormous savings for the government.

6.2 Predictive analytics

Analysing the results of the case study, it is possible to observe that ANN, ARX and GPR show good performance in time series prediction with a small number of samples. Indeed, ARX seemed a slightly better in this specific scenario.

However, it is important to consider that ANN are universal approximators, and usually they can adapt better to different types of data.

Also, GPR has the interesting advantage of giving not only a prediction, but also a confidence interval of that prediction, what can be very useful in BI environments and fraud detection systems.

In all cases, the use of the Auto Correlation function appears to be very useful for configuring the algorithm in time series with a clear seasonal component.

In this master's theses, two applications have been proposed for these prediction algorithms: a fraud detection system; and a prediction system integrated on a BI process (ETAPL).

In the first case, all algorithms seem to be precise enough for implementing a fraud detection system for this scenario. However, ANN and GPR show more interesting characteristics due its versatility of ANN and the information about the prediction confidence in GPR as mentioned before.

In the second case, the proposed ETAPL methodology should provide a faster and most efficient BI environment increasing the prediction capabilities of a BI environment in a transparent manner a reducing the loading time of the reports. It has been shown that this proposal would be especially useful in environments with well pre-defined indicators and even with a small number of past samples.

6.3 Data storage using Big Data

In this work is shown how Big Data technologies have become a real alternative to traditional relational databases.

In this way, we improved the loading and querying data using Pentaho Kettle and Hbase as a data storage. The results showed an improved performance in terms of time for loading data in the SIAPE database and improving the latency for querying data.

Moreover, should be noticed that good modelling of data can help to get good query latency.

6.4 Future Works

The Business Intelligence area is in constant development and can be improved from different approaches. Following the research line of this master's thesis, several possibilities remain open for improving the current proposals:

- Study the possibility of use of classification algorithms able to suggest correlation between different indicators automatically.
- Implementation of multivariate prediction using correlated indicators.
- Explore the use of MOS techniques for faster determining the optimal configuration of the prediction algorithms.
- Uses of machine learning techniques for automatically determine the threshold value of NRMSE and COD in the fraud detection system.
- Analysis and testing of other Big Data technologies.

7 REFERENCES

Own Publications

Campos, S. R., Fernandes, A. A., de Sousa Jr., R. T., de Freitas, E. P., da Costa, J. P., Serrano, A. M., de Sousa, R. T.; Rodrigues, C. T. (2012). Ontologic Audit Trails mapping for detection of irregularities in payrolls. *International Conference on Next Generation Web Services Practices (NWeSP 2012).* São Carlos, Brazil.

da Costa, J. P., Freitas, E. P., David, B. M., Amaral, D., Serrano, A. R., & de Sousa Jr., R. T. (2012). Improved Blind Automatic Malicious Activity Detection in Honeypot Data. The International Conference on Forensic Computer Science (ICoFCS), Best Paper Award. Brasilia.

da Costa, J. P., Freitas, E. P., Serrano, A. M., & de Sousa Jr., R. T. (2012). Improved Parallel Approach to PCA Based Malicious Activity Detection in Distributed Honeypot Data. *International Journal of Forensic Computer Science (IJoFCS)*.

da Costa, J. P., Rubio Serrano, A. M., Rodrigues, D. C., Campos, S. R., & de Sousa Jr, R. T. (2013). *Patent No. BR 10 2013 011211 9 (INPI).* Brazil.

Fernandes, A. A., Amaro, L. C., da Costa, J. P., Martins, V. A., Serrano, A. M., & de Sousa Jr., R. T. (2012). Construction of Ontologies by using Concept Maps: a Study Case of Business Intelligence for the Federal Property Department. *Proc. International Conference on Business Intelligence and Financial Engineering (BIFE).* Lanzhou, China.

Rubio Serrano, A. M., da Costa, J. P., & de Sousa Jr, R. T. (2013). *Patent No. BR 10 2013 003266 2 (INPI).* Brazil.

Rubio Serrano, A., da Costa, J., Cardonha, C., & de Sousa Jr., R. (2012). Neural Network Predictor for Fraud Detection: A Study Case for the Federal Patrimony Department. *The International Conference on Forensic Computer Science (ICoFCS).* Brasilia.

Other Authors

Andrew, C., Huy, L., & Aditya, P. (2012). Big Data. XRDS. *The ACM Magazine for students, Vol. 19, No 01*, 7-8.

Box, G., Jenkins, G., & Reinsel, G. (2008). In *Time Series Analysis: Forecasting and Control* (pp. 21-45). Wiley & Sons.

Brockwell, P., & Davis, R. (2002). *Introduction to time series and forecasting.* Springer, 2nd Edition.

Cassandra . (2013). *Cassandra*. Retrieved from http://cassandra.apache.org

Chang F, e. a. (2008). Bigtable: A Distributed Storage System for Structured Data. *ACM Transactions on Computer Systems. Vol. 26, No. 2*, 1-26.

Chaudhuri, S., Dayal, U., & Vivek, N. (2011). An Overview of Business Intelligence Technology. *Communications of the ACM, Vol. 54*, 88-89.

Cohen, W. (1995). Fast effective rule induction. *Proc. of the 12th International Conference on Machine Learning.* Morgan Kaufmann, Palo Alto, CA.

Darbellay, G. A., & Slama, M. (2000). Forecasting the short-term demand for electricity: Do neural networks stand a better chance? *International Journal of Forecasting, 16*, pp. 71– 83.

Davis, R. A. (2001). Gaussian Process. In D. Brillinger, *Encyclopedia of Environmetrics, Section on Stochastic Modeling and Environmental Change.* New York: Willey.

De Gooijer, J., & Hyndman, R. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, pp. 443-473.

Dimitryou-Fakalou, C. (2011). Yule Walker estimation for the Moving-Average model. *International Journal of Stochastic Analysis*.

Do, H., & Rahm, E. (2000). *On Metadata Interoperability in Data Warehouses. Techn. Report 1-2000 (http://dol.uni-leipzig.de/pub/2000-13).* University of Leipzig, Department of Computer Science.

Dorronsoro, J., Ginel, F., Sanchez, C., & Cruz, C. (1997). Neural fraud detection in credit card operations. *IEEE Transactions on Neural Networks, 8*(4), 827-834.

Drossu, R., & Obradovic, Z. (1996). Rapid design of Neural Networks for Time Series Prediction. *Computational Science and Engineering, IEEE, 3*, 78-89.

Evelson, B., Moore, C., Karel, R., Kobielus, J., & Coit, C. (2009, September 11). *Forrester's BI Maturity Assesment Tool.* Retrieved from Forrester Research Inc.: http://pt.scribd.com/doc/76597995/Forresters-Bi-Maturity-Assessment-Tool

Fernandes, A. (2012). *Proposta de aplicação de ontologia através de mapas conceituais e uso de algoritmos de preditividade para uma solução de Business Intelligence.* Brasilia: UnB Master's Thesis.

Frank, R., Davey, N., & Hunt, S. (1999). *Time Series Prediction and Neural Networks.* Hertfordshire, UK: University of Herdfordshire.

Hand, D. (1985). *Discrimination and Classification.* New York: J. Wiley & Sons.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation.* Englewood, NJ: Prentice Hall, 2nd Edition.

Hylleberg, S. (1992). *General Introduction in Modelling Seasonality.* Oxford, UK: Oxford University Press, pp. 3-14.

Inmon, B. (1992). *Building the Data Warehouse. 1st Edition. .* New York: Wiley and Sons.

Jantzen, J. (1998). *Introduction to Perceptron Networks.* Copenaghe: Technical University of Denmark.

Kimball, R. (1998). *The Data Warehousing Lifecycle Toolkit: expert methods for designing, developing, and deploying data warehouses.* New York: John Wiley & Sons.

Koskela, T. (1996). Time Series Prediction with Multilayer Perceptron. *Proceedings of the World Congress on Neural Networks.* Chicago.

Luhn, H. (1958). *A Bussiness Intelligence System.* IBM Journal.

Murray-Smith, R., & Girard, A. (2001). Gaussian process priors with arma noise models. *Irish Signals and Systems Conference*, (pp. 147-152). Maynooth.

Nason, G. (2004). *Statistics in Vulcanology. Chapter 11.* Bristol, UK: Dept. of Mathematics, University of Bristol.

O'Brien, C. J. (2012). CAP Theorem in the Cloud. *MSCS 6350 Project*.

Office of Science and Technology Policy of The United States. (2012). Retrieved from www.WhiteHouse.gov/OSTP

Oracle. (2005). *Oracle Database Data Warehousing Guide, 10g Release 2 (10.2), B14223-02, Copyright © 2001.*

P. Clark, T. N. (1989). The CN2 induction algorithm. *Machine Learning, 3*, 261-285.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*, 1345-1359.

Patterson, D., Chan, K., & Tan, C. (1993). *Time Series Forecasting with Neural Networks: A comparative study.* Singapore: Proc. the International Conference on Neural Networks Applications to Signal Processing.

Pérez-Cruz, F., Vaerenbergh, S. V., Murillo-Fuentes, J. J., Lázaro-Gredilla, M., & Santamaria, I. (2013). Gaussian processes for nonlinear signal processing. *IEEE Signal Processing Magazine, vol. 30, no. 40*, 40-50.

Peters, E. (1994). *Fractal Market Hypothesis: Applying Chaos Theory to Investment and Economics.* New York: Wiley & Sons.

Pfeifer, P. E. (1997). *A Brief Primer on Probability Distributions.* Darden Business Publishing, Univ. of Virginia.

PostgreSQL. (2013). *PostgreSQL: The world's most advanced open source database.* Retrieved from http://www.postgresql.org/

Quinlan, J. (1990). Learning logical definitions from relations. *Machine Learning, 5*, 239-266.

R.J. Bolton, D. H. (2002). Statistical Fraud Detection: A Review. *Statistical Science, 17*(3).

Rahman, S. M., Monzurur, M., Faisal, K., & Mushfiqur, R. (2014). Integrated Data Mining and Business Intelligence. In J. Wang, *Encyclopedia of Business Analytics and Optimization* (pp. 1234-1253). Hershey: IGI Golbal, doi:10.4018/978-1-4666-5202-6.ch11.

Rasmussen, C. E., & Williams, C. K. (2006). Gaussian Processes for Machine Learning. *The MIT Press*.

Rubio Serrano, A. M. (2012). *Study and implementation of a predictive analytics module for the Business Intelligence System of the Brazilian Ministry of Planning, Budgeting and Management.* Barcelona, Spain: Master's Final Thesis, Politecnic University of Catalonia (UPC).

Rud, O. (2009). *Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy.* Hoboken, N.J.: Wiley & Sons, 4th Edition.

Russom, P. e. (2011). Big Data Analytics. *TDWI Best Practices Report*.

Sergio, P., & María, R. (2011). *Pentaho Data Integration 4 Cookbook. First Edition.* London: Packt Pub Limited.

SIAPE. (2013). *Siape - Sistema Integrado de Administração de Recursos Humanos .* Retrieved from https://www.serpro.gov.br/conteudo-solucoes/produtos/administracao-federal/siape-sistema-integrado-de-administracao-de-recursos-humanos

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *NIPS*, 2960-2968.

Stacy C., e. a. (2011). Why big data is a big deal. *Computerworld, Vol. 45, No. 20*, 1-6.

Tang J., e. a. (2009). The Research & Application of ETL Tool in Business Intelligence Project. *IEEE International Forum on Information Technology and Applications (IFITA), Vol. 2*, 620-623.

Vora, M. N. (2011). Hadoop-HBase for large-scale data. *International Conference of Computer Science and Network Technology (ICCSNT), IEEE , Vol. 1*, 601-605.

Xiao, J., He, C., & Wang, S. (2012). Crude Oil Price Forecasting: A Transfer Learning based Analog Complexing Model. *Fifth International Conference on Business Intelligence and Financial Engineering.* Lanzhou, China.

Zhang, G. (2003). Time-series forecasting using a hybrid ARIMA and Neural Network model. *Neurocomputing, 50*(1), 159-175.

8 APPENDIX

MATLAB CODES

Coefficient of Determination (COD)

```
function [ans] = coefofdet(Y,f)



%Y - prediction

%f - expected(real)



   numerator=0;

   denominator=0;

   meanf = mean(f);

   for i=1:12

      numerator=numerator+(f(i)-Y(i)).^2;

      denominator=denominator+(f(i)-meanf).^2;

   end



ans = 1-(numerator/denominator);
```

Normalized Root Mean Square Error (NRMSE)

```
function [ans] = nrmse(Y,f)



%Y - prediction

%f - expected(real)



NRMSE =

sqrt(mean((double(x) - double(y)).^2)/mean(double(x).^2))



ans = NRMSE;
```