

**PREDICTIVE ANALYTICS VIA GAUSSIAN PROCESSES AND
STATISTICAL AUDIT VIA GAUSSIAN MIXTURES IN
BUSINESS INTELLIGENCE SYSTEMS**

BRUNO HERNANDES AZENHA PILON

**DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA**

Brasília, Abril de 2015

**FACULDADE DE TECNOLOGIA
UNIVERSIDADE DE BRASÍLIA**

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

**PREDICTIVE ANALYTICS VIA GAUSSIAN PROCESSES AND
STATISTICAL AUDIT VIA GAUSSIAN MIXTURES IN
BUSINESS INTELLIGENCE SYSTEMS**

BRUNO HERNANDES AZENHA PILON

**ORIENTADOR: JOÃO PAULO CARVALHO LUSTOSA DA COSTA
COORIENTADOR: JUAN JOSÉ MURILLO-FUENTES**

**DISSERTAÇÃO DE MESTRADO EM
ENGENHARIA ELÉTRICA**

**PUBLICAÇÃO: PPGEE.DM - 591/2015
BRASÍLIA/DF: ABRIL - 2015**

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

PREDICTIVE ANALYTICS VIA GAUSSIAN PROCESSES AND
STATISTICAL AUDIT VIA GAUSSIAN MIXTURES IN
BUSINESS INTELLIGENCE SYSTEMS

BRUNO HERNANDES AZENHA PILON

DISSERTAÇÃO DE MESTRADO 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.

Banca Examinadora


Prof. João Paulo C. Lustosa da Costa, Dr. (ENE-UnB)
Orientador


Prof. Rafael Timóteo de Sousa Júnior, Dr. (ENE-UnB)
Examinador interno


Cristiano Rocha Heckert, Dr. (MP)
Examinador externo

Brasília, Abril de 2015

FICHA CATALOGRÁFICA

PILON, BRUNO HERNANDES AZENHA

Predictive Analytics via Gaussian Processes and Statistical Audit via Gaussian Mixtures in Business Intelligence Systems.

[Distrito Federal] 2015.

xvi, 69p., 297mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2015).

Dissertação de Mestrado – Universidade de Brasília.

Faculdade de Tecnologia.

Departamento de Engenharia Elétrica.

- | | |
|-----------------------------|------------------------|
| 1. Inteligência de Negócios | 2. Análise Preditiva |
| 3. Processos Gaussianos | 4. Misturas Gaussianas |
| I. ENE/FT/UnB | II. Título (série) |

REFERÊNCIA BIBLIOGRÁFICA

PILON, B.H.A. (2015). Predictive Analytics via Gaussian Processes and Statistical Audit via Gaussian Mixtures in Business Intelligence Systems. Dissertação de Mestrado em Engenharia Elétrica, Publicação PPGEE.DM-591/2015, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 87p.

CESSÃO DE DIREITOS

AUTOR: Bruno Hernandes Azenha Pilon

TÍTULO: Predictive Analytics via Gaussian Processes and Statistical Audit via Gaussian Mixtures in Business Intelligence Systems.

GRAU / ANO: Mestre / 2015

É concedida à Universidade de Brasília permissão para reproduzir cópias desta dissertação de mestrado e para emprestar ou vender tais cópias somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação e nenhuma parte dessa dissertação de mestrado pode ser reproduzida sem autorização por escrito do autor.



Bruno Hernandes Azenha Pilon

Universidade de Brasília - Faculdade de Tecnologia

Departamento de Engenharia Elétrica

70910-900 — Brasília-DF — Brasil

*"In an economy where the only certainty is uncertainty,
the one sure source of lasting competitive advantage is knowledge."*

Ikujiro Nonaka

À minha mãe, Angela.

À minha futura esposa, Agnes, e à nossa futura família.

Agradecimentos

A Deus.

À minha mãe, Angela, por ter semeado e cultivado o valor da educação e dos estudos em nossa família. Por nunca ter deixado de nos prover um porto seguro, mesmo diante das adversidades da vida. Por sua incansável dedicação ao nosso bem-estar e à nossa felicidade. Por ter me guiado pelo caminho do bem, ter me ensinado a lutar o bom combate, ter construído as plataformas que me permitiram alçar os voos que voei e ter amortecido as quedas que sofri. Por ser o meu exemplo de vida, mãezinha, a minha eterna gratidão.

A você, Agnes, pelo lindo sorriso acolhedor que tão bem me recebeu nesta cidade fria e por ter abdicado momentaneamente do seu direito de convivência para que este trabalho pudesse ser realizado. Te amo.

Ao meu orientador e amigo, Prof. João Paulo Carvalho Lustosa da Costa, cuja paixão pela academia me inspirou e enriqueceu desde o nosso primeiro contato. Por ter me aberto as portas do mundo de processamento de sinais em arranjos, um campo de pesquisa fascinante. Por sempre se colocar à disposição e ser extremamente acessível. Por não ter desistido deste aluno, mesmo nos momentos mais difíceis. Por ter me dado a oportunidade de trabalhar ao seu lado, uma experiência bastante recompensadora, e por ter me dado a oportunidade de fazer muito mais do que o tempo e as circunstâncias permitiram. Certamente este é só o começo de uma longa parceria.

Ao meu coorientador, Prof. Juan José Murillo-Fuentes, que da longínqua Universidade de Sevilha (ES) contribuiu imensamente com este trabalho. Suas análises rápidas, intervenções cirúrgicas e enorme conhecimento relacionado a processos e misturas Gaussianas foram decisivos para a validação e qualidade deste trabalho.

Aos colegas do Laboratório de Processamento de Sinais em Arranjos (LASP), do Laboratório de Tecnologias da Tomada de Decisão (LATITUDE) e do Grupo de Processamento Digital de Sinais (GPDS) da UnB, em ordem alfabética, Danilo Tenório, Jayme Milanezi, Marco Marinho, Stefano Mozart, Stephanie Alvarez e Toni Serrano, pelo apoio, convivência e ótimos momentos.

Aos meus professores e orientadores de iniciação científica, Profa. Rosangela Gin e Prof. Reinaldo Bianchi, por terem incentivado o meu potencial acadêmico na graduação e por terem criado os bichinhos da pesquisa que já naquela época me picaram.

Ao Ministério do Planejamento, Orçamento e Gestão (MP) que, por meio dos acordos de cooperação com o Laboratório LATITUDE da UnB, cedeu os dados e sistemas de inteligência de negócio utilizados neste trabalho.

RESUMO

ANÁLISE PREDITIVA VIA PROCESSOS GAUSSIANOS E AUDITORIA ESTATÍSTICA VIA MISTURAS GAUSSIANAS EM SISTEMAS DE INTELIGÊNCIA DE NEGÓCIOS

Autor: Bruno Hernandes Azenha Pilon

Orientador: Prof. Dr. João Paulo Carvalho Lustosa da Costa

Coorientador: Prof. Dr. Juan José Murillo-Fuentes

Programa de Pós-graduação em Engenharia Elétrica

Brasília, Abril de 2015

Um sistema de Inteligência de Negócios, do inglês *Business Intelligence* (BI), é um sistema de informação que emprega ferramentas de diversas áreas do conhecimento na coleta, integração e análise de dados para aprimorar e embasar o processo decisório em empresas e instituições governamentais. O Ministério do Planejamento, Orçamento e Gestão (MP), órgão do governo federal brasileiro, possui uma série de sistemas de inteligência de negócios e, neste trabalho, dois destes sistemas foram considerados. O primeiro sistema de BI, mantido pela Secretaria de Patrimônio da União (SPU), contém dados de arrecadação mensal de impostos daquela Secretaria, enquanto o segundo sistema de BI, mantido pela Coordenadoria de Inteligência e Auditoria Preventiva da Folha de Pagamento (CGAUD), contém dados da folha de pagamento dos servidores públicos federais brasileiros. Ambos os sistemas foram construídos objetivando-se a detecção de fraudes e irregularidades como evasão fiscal e pagamentos não autorizados. Ao longo deste trabalho, pretende-se incorporar estágios que adicionem análise preditiva e melhorias de performance aos sistemas de BI existentes. No sistema de BI da SPU, Regressão por Processos Gaussianos (RPG) é utilizada para modelar as características intrínsecas da principal série temporal financeira. RPG retorna uma descrição estatística completa da variável estimada, que pode ser tratada como uma medida de confiança e pode ser utilizada como gatilho para classificar dados em confiáveis ou não confiáveis. Ademais, um estágio de pré-processamento reconfigura a série temporal original em uma estrutura bidimensional. O algoritmo resultante, com RPG em seu núcleo, superou métodos preditivos clássicos como indicadores financeiros e redes neurais artificiais. No sistema de BI da CGAUD, um Modelo de Misturas Gaussianas (MMG) é utilizado para descrever o processo estocástico que governa a distribuição de probabilidades dos contracheques. Rotular uma probabilidade relativa em cada contracheque habilita o sistema de BI a listá-los e filtrá-los com base em suas probabilidades. A inserção de um filtro estatístico em um sistema de BI determinístico resultou em efetiva redução na quantidade de dados a serem analisados pelas trilhas de auditoria.

Palavras-chave: Inteligência de negócios, processos Gaussianos, misturas Gaussianas.

ABSTRACT

PREDICTIVE ANALYTICS VIA GAUSSIAN PROCESSES AND STATISTICAL AUDIT VIA GAUSSIAN MIXTURES IN BUSINESS INTELLIGENCE SYSTEMS

Author: Bruno Hernandes Azenha Pilon

Supervisor: Prof. Dr. João Paulo Carvalho Lustosa da Costa

Co-supervisor: Prof. Dr. Juan José Murillo-Fuentes

Programa de Pós-graduação em Engenharia Elétrica

Brasília, April of 2015

A Business Intelligence (BI) system is an information system that employs tools from several areas of knowledge for the collection, integration and analysis of data to improve and support the decision making process in companies and governmental institutions. The Ministry of Planning, Budget and Management, in portuguese *Ministério do Planejamento, Orçamento e Gestão* (MP), an agency of the Brazilian federal government, possesses a wide number of BI systems and, in this work, two of those systems were considered. The first BI system, maintained by the Federal Patrimony Department, in portuguese *Secretaria de Patrimônio da União* (SPU), contains data regarding the monthly tax collection of that department, whereas the second BI system, maintained by the Human Resources Auditing Department, in portuguese *Coordenadoria de Inteligência e Auditoria Preventiva da Folha de Pagamentos* (CGAUD), contains data regarding the payroll of Brazilian federal employees. Both systems were designed aimed at fraud and irregularities detection such as tax evasion and unauthorized payments. Throughout the present work, we aim to incorporate stages into the existing BI systems in order to add predictive analytics and performance enhancements. In the BI system of SPU, Gaussian Process for Regression (GPR) is used to model the intrinsic characteristics of the core financial time series. GPR natively returns a full statistical description of the estimated variable, which can be treated as a measure of confidence and can be used as a trigger to classify trusted and untrusted data. In order to take into account the multidimensional structure of the original data, we also propose a pre-processing stage for reshaping the original time series into a bidimensional structure. The resulting algorithm, with GPR at its core, outperforms classical predictive schemes such as financial indicators and artificial neural networks. In the BI system of CGAUD, a Gaussian Mixture Model (GMM) is used to describe the stochastic process that governs the probability distribution of payrolls. Attaching a relative probability into each payroll enables the BI system to sort and filter payrolls based on their probabilities. Inserting a statistical filter in a deterministic BI system showed to be effective in reducing the amount of data to be analyzed by rule-based audit trails.

Keywords: Business intelligence, Gaussian process, Gaussian mixtures.

CONTENTS

1	INTRODUÇÃO	1
1.1	CONTEXTO E MOTIVAÇÃO.....	1
1.2	OBJETIVOS E CONTRIBUIÇÕES	4
1.3	ORGANIZAÇÃO DESTE TRABALHO	7
2	INTRODUCTION.....	8
2.1	CONTEXT AND MOTIVATION	8
2.2	OBJECTIVES AND CONTRIBUTIONS.....	10
2.3	ORGANIZATION OF THIS WORK	13
3	THEORETICAL FOUNDATION	15
3.1	BUSINESS INTELLIGENCE	15
3.1.1	KEY COMPONENTS.....	16
3.1.2	FRAUD DETECTION APPLICATIONS	17
3.2	FINITE MIXTURE MODELS	19
3.2.1	ESTIMATION OF PARAMETRIC MIXTURE MODELS.....	20
3.2.2	EXPECTATION MAXIMIZATION ALGORITHM	21
3.3	GAUSSIAN PROCESS FOR REGRESSION	23
3.3.1	MULTIVARIATE GAUSSIAN DISTRIBUTION.....	23
3.3.2	GAUSSIAN PROCESSES	25
3.3.3	REGRESSION MODEL AND INFERENCE	26
3.3.4	COVARIANCE FUNCTIONS AND ITS HYPERPARAMETERS	27
4	BUSINESS INTELLIGENCE SYSTEMS AND DATA	30
4.1	PAYROLLS OF FEDERAL EMPLOYEES	31
4.2	FEDERAL TAX COLLECTION	34

5 STATISTICAL AUDIT	36
5.1 STATYSTICAL ANALYSIS ON A DETERMINISTIC BI SYSTEM	36
5.1.1 STATISTICAL AUDIT MODULE	38
5.1.2 GMM FOR STATISTICAL AUDITING	38
5.2 OPTIMIZATION AND EXPERIMENTAL RESULTS	40
6 PREDICTIVE ANALYTICS.....	45
6.1 UNIDIMENSIONAL PREDICTOR MODEL.....	45
6.1.1 MEAN AND COVARIANCE FUNCTION MODELING	45
6.1.2 UNIDIMENSIONAL PREDICTION RESULTS.....	47
6.2 BIDIMENSIONAL DATASET RESHAPE	48
6.2.1 TIME CROSS-CORRELATION	49
6.2.2 DATASET RESHAPE.....	50
6.3 OPTIMIZATION AND PREDICTION RESULTS	51
6.3.1 HYPERPARAMETERS TUNING	52
6.3.2 BIDIMENSIONAL PREDICTION RESULTS	54
6.3.3 PREDICTION COMPARISON AND ERROR METRICS.....	54
6.3.4 CLASSIFICATION STAGE PROPOSALS.....	56
7 CONCLUSIONS.....	58
PUBLICATIONS FROM THIS WORK	61
REFERENCES	62
A ERROR METRIC FORMULAS	68

LIST OF FIGURES

1.1	Estimativa do volume total de dados armazenados eletronicamente em ZB ao longo dos anos. Com uma taxa de crescimento composta anual de 40 por cento, a estimativa do volume de dados armazenados deve alcançar 45 ZB no ano de 2020 [3]....	2
2.1	Estimation of the total volume of electronically stored data in ZB along the years. With a growing compound annual rate of 40 percent, stored data is estimated to reach nearly 45 ZB by 2020 [3].....	9
3.1	A traditional architecture and components of a generic BI system	16
4.1	Scatter plot of 10,000 samples of payroll data, with gross income in one dimension (ordinate) and total discounts and deductions in the other dimension (abscissa). Both dimensions are plotted in <i>Reais</i> , the Brazilian currency.....	31
4.2	Architecture of the current state-of-the-art BI system of CGAUD	32
4.3	Example of a concept map for an audit trail. Adapted from [14].....	33
4.4	Architecture of the current state-of-the-art BI system of SPU	34
4.5	Monthly tax collected by SPU, in <i>reais</i> (R\$), indexed by the m^{th} month.	35
5.1	Scatter plot of 10,000 samples of payroll data, showed in Fig. 4.1, with a zoom around the origin for a better visualization of the correlation profile.	37
5.2	Block architecture of the proposed statistical audit module solution in the original BI architecture shown in Fig. 4.2.....	38
5.3	Contour plot of the estimated pdf of the dataset presented in Fig. 5.1 with (a) 8 sources (log-likelihood: -174317); (b) 16 sources (log-likelihood: -173282); (c) 24 sources (log-likelihood: -173019) and (d) 32 sources (log-likelihood: -172937). The axis in all subfigures are the same as in Fig.5.1.	41
5.4	(a) Contour plot and (b) surface plot of the resulting pdf of the proposed GMM.	42

6.1	Normalized plot of the posterior inference of the Gaussian process, indexed by a continuous time interval $\mathcal{X} = [0, 80]$, obtained using the covariance function (a) $k_{2,1}(\mathbf{x}, \mathbf{x}')$ in red (the periodic component) and $k_{2,2}(\mathbf{x}, \mathbf{x}')$ in blue (the squared exponential component); (b) $k_2(\mathbf{x}, \mathbf{x}')$ in black (the product of both components)....	47
6.2	Prediction results from conditioning the posterior Gaussian jointly distribution at a continuous time interval $\mathcal{X} = [0, 75]$. The blue dots are the training data, the red dots are the target data, the black tick line is the expected value at a time index and the gray band represents the 95% confidence interval (two standard deviations above and below the expected value).....	48
6.3	Estimated absolute normalized cross-correlation between the target data and the hole SPU data set. The sequence was trimmed due the zero-padding, and the red circles highlights where the lag m is a multiple of 12 months.	50
6.4	Plot of the SPU data set converted in a 2D array.....	51
6.5	Plot of the Gaussian process prediction in blue, target SPU data in red. The error bars corresponds to a confidence interval of two standard deviations with respect to the predictive mean (around 95% of confidence).....	54
6.6	Monthly plot of target data and predictive results, in <i>Reais</i> (R\$), indexed by the m^{th} month.....	56

LIST OF TABLES

1.1	Os cinco V's das propriedades dos dados.....	3
2.1	The five V's of data properties.....	10
3.1	Key components in BI systems framework.....	16
4.1	Legal attributions of MP	30
5.1	Fraud occurrences detected by audit trails, grouped by trail ID # and divided according to their probability of occurrence.	43
5.2	Total fraud occurrences detected by audit trails, divided according to their probability of occurrence.	43
6.1	Optimized set of hyperparameters Θ , σ_1^2 and σ_n^2 after 100 iterations, using the marginal likelihood with the kernel in (6.4).....	53
6.2	Performance comparison by several error metrics	55

LIST OF ACRONYMS

ABS

Absolute Value. 42

ANN

Artificial Neural Network. 19, 59

BI

Business Intelligence. xi, xiii, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 30, 31, 32, 34, 36, 37, 38, 42, 44, 45, 58, 59

CGAUD

Coordenadoria de Inteligência e Auditoria Preventiva da Folha de Pagamentos (Intelligence and Preventive Audit of Payrolls Division). xi, 3, 4, 5, 7, 10, 11, 12, 13, 32, 36, 58, 59

CIO

Chief Information Officer. 58

DW

Data Warehouse. 17

E-Step

Expectation Step. 39

EB

Exabyte = 10^{18} bytes. 1, 8

EM

Expectation Maximization. 13, 21, 22, 38, 39, 40, 41, 42

ETL

Extract, Transform and Load. 16, 17

GB

Gigabyte = 10^9 bytes. 4, 31

GMM

Gaussian Mixture Model. xi, 12, 13, 20, 36, 37, 38, 39, 40, 42, 44, 58

GPR

Gaussian Process for Regression. 12, 13, 26, 27, 28, 45, 47, 48, 51, 52, 55, 56, 59

iid

independently and identically distributed. 19, 20, 26, 27

KRR

Kernel Ridge Regression. 28, 52

M-Step

Maximization Step. 39

ME

Maximização da Esperança (Expectation Maximization). 7

ML

Machine Learning. 15, 17, 28

MMG

Modelo de Misturas Gaussiana (Gaussian Mixture Model). 5, 7

MP

Ministério do Planejamento, Orçamento e Gestão (Ministry of Planning, Budget and Management). 4, 7, 11, 13, 30, 58

ODS

Operational Data Store. 16

OLAP

On-line Analytical Processing. 15, 17

pdf

probability density function. xi, 11, 19, 20, 21, 23, 36, 37, 38, 39, 40, 42, 58

RPG

Regressão por Processos Gaussianos (Gaussian Process for Regression). 6, 7

SIAPE

Sistema Integrado de Administração de Recursos Humanos (Integrated System of Human Resources Administration). 31, 32, 37, 38, 40

SPU

Secretaria de Patrimônio da União (Federal Patrimony Department). xi, xii, 3, 5, 6, 7, 10, 12, 13, 31, 33, 34, 35, 45, 46, 47, 48, 49, 50, 54, 55, 56, 58, 59

SVM

Supported Vector Machine. 28, 52

ZB

Zettabyte = 10^{21} bytes. xi, 1, 8

Capítulo 1

INTRODUÇÃO

1.1 CONTEXTO E MOTIVAÇÃO

Conhecimento é poder. Em corporações e instituições governamentais, informações referentes a inteligência do negócio são vitais para auxiliar a alta administração no processo de tomada de decisão, na condução dos negócios e nas operações institucionais [1]. Neste domínio de conhecimento, BI¹ evoluiu como um importante campo de pesquisa. Ademais, mesmo fora do mundo acadêmico, BI foi reconhecida como uma iniciativa estratégica em aumentar a eficácia e gerar inovações em diversas aplicações práticas no universo dos negócios.

Neste contexto, avanços tecnológicos aumentaram massivamente o volume de dados e informações disponíveis eletronicamente, onde cerca de 2,5 EB de dados são criados a cada dia ao redor do mundo, e este número dobra a cada 40 meses aproximadamente [2]. A Fig. 1.1 ilustra esta tendência de aumento exponencial do volume de dados ao longo dos anos. Por outro lado, grande parte destes novos dados não possuem qualquer estrutura associada. Organizar e analisar este volume de dados em ascensão e encontrar, em seu conteúdo, significado e informação útil são pontos chave para sistemas de BI.

Sobre este tópico, Hal Varian, Economista-Chefe do Google e professor emérito da Universidade da Califórnia, comentou: “Então, o que está se tornando onipresente e barato? Dados. E o que é complementar aos dados? Análise. Então, minha recomendação é frequentar muitos cursos relacionados à manipulação e análise de dados: banco de dados, aprendizado de máquina, econometria, estatística, visualização, e assim por diante.” [4].

Não obstante o volume crescente de dados, a gestão de *grandes dados* também lida com a variedade, a velocidade, a variabilidade e o valor dos dados. Em [5], pela primeira vez, a gestão de dados foi tratada a partir de uma visão tridimensional, onde o volume, a velocidade e a

¹Inteligência de Negócios, do inglês *Business Intelligence* (BI)

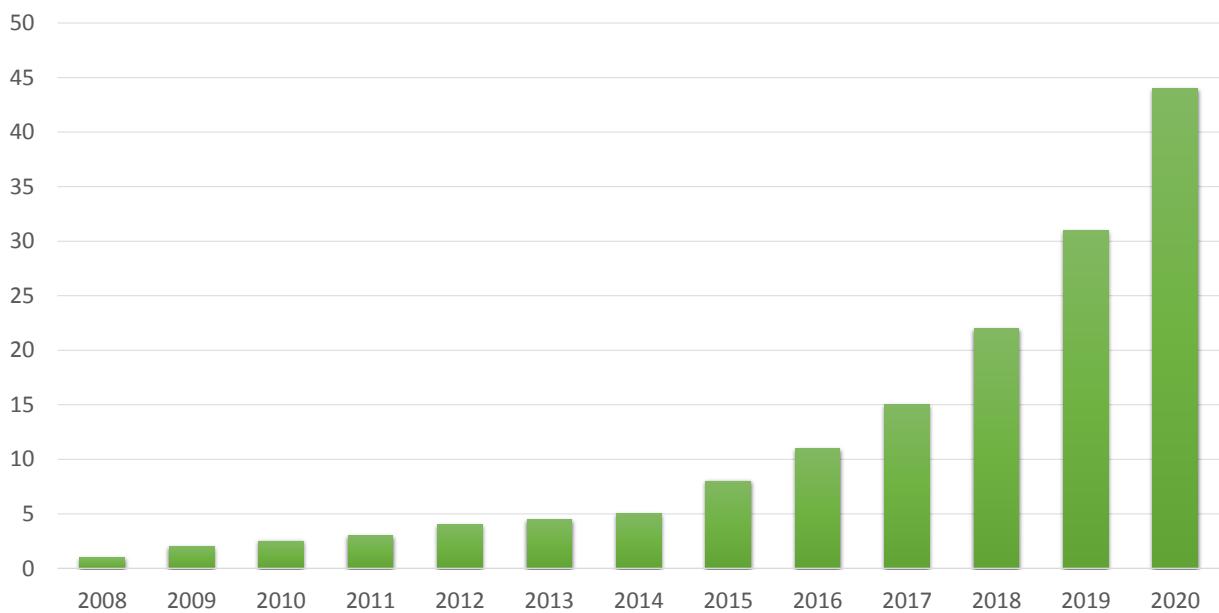


Figura 1.1: Estimativa do volume total de dados armazenados eletronicamente em ZB ao longo dos anos. Com uma taxa de crescimento composta anual de 40 por cento, a estimativa do volume de dados armazenados deve alcançar 45 ZB no ano de 2020 [3].

variedade contribuem, de forma independente, para o desenvolvimento de ferramentas e algoritmos para a gestão de dados. Em [6], os autores adicionam duas outras dimensões nesta perspectiva: variabilidade e valor. Estas propriedades estão resumidas na Tabela 1.1.

Com todo este montante de dados se tornando onipresente e barato, novas ferramentas de captura, descoberta e análise podem ajudar empresas e instituições governamentais a ganhar conhecimento a partir dos seus dados desestruturados, que respondem por mais de 90% do universo digital [8]. Estas ferramentas podem ser programadas para criar dados sobre dados de forma automatizada, assim como as rotinas de reconhecimento facial que ajudam a identificar fotos no Facebook. Dados sobre dados, ou metadados, crescem ao dobro da taxa do universo digital como um todo [8].

Como consequência deste cenário, onde mais e mais dados são armazenados e processados eletronicamente e vastas quantidades de informação estão se tornando facilmente disponíveis, o processo de tomada de decisão em empresas e organizações governamentais está consideravelmente mais rápido nos dias atuais. Tomadores de decisões estratégicas são expostos a um enorme fluxo de dados e informação e estão sob constante pressão para responder a situações excepcionais e para aproveitar oportunidades de negócio disponíveis em curtos espaços de tempo.

Neste ambiente, sistemas de BI unem diversas áreas do conhecimento para entregar informação e conhecimento para empresas e administrações, apoiando o processo de tomada de decisão em todos os níveis de gestão. Sistemas de BI são referenciados como um conjunto integrado de ferra-

Tabela 1.1: Os cinco V's das propriedades dos dados

Propriedade	Descrição
Volume	O volume mede a quantidade de dados disponíveis em uma organização, o que não significa necessariamente que a organização possua todos estes dados, desde que ela consiga ao menos acessá-los. Com o aumento do volume de dados, o valor dos dados registrados declinará proporcionalmente à idade, tipo, riqueza e quantidade, em meio a outros fatores [6].
Velocidade	É a medida da velocidade da criação, do fluxo e da aglutinação dos dados. Esta característica não está limitada à velocidade de novos dados, mas também à velocidade do fluxo de dados. Por exemplo, dados de dispositivos sensores estão sendo armazenados constantemente em um banco de dados, e este montante de dados não é desprezível. Portanto, sistemas tradicionais não possuem capacidade suficiente para analisar dados que estão constantemente em movimento [7].
Variedade	Variedade é a medida da riqueza da representação dos dados - texto, imagens, vídeo, áudio, etc. De um ponto de vista analítico, é o principal obstáculo na utilização efetiva de grandes volumes de dados. Formatos de dados incompatíveis, estrutura de dados não alinhadas e semântica de dados inconsistente são exemplos de desafios significativos [6].
Variabilidade	É uma grande tarefa criar correspondências, depurar e transformar dados originados de várias fontes. É necessário, também, conectar e correlacionar relações e hierarquias ou a espiral de dados pode sair do controle rapidamente [7].
Valor	O valor dos dados mede a utilidade dos dados na tomada de decisão. É consenso que o propósito da computação é o conhecimento, não os números. A ciência dos dados é exploratória e útil na construção do conhecimento dos dados, mas a ciência analítica engloba o poder preditivo dos grandes dados [6].

mentas e tecnologias que são utilizadas para coletar, integrar, analisar e disponibilizar dados [9]. Sistemas de BI diferem de sistemas de gestão da informação tradicionais por terem - antes de tudo - um escopo maior, análises multi-variáveis de dados semi estruturados que provém de diferentes fontes e sua apresentação multidimensional [10]. Um sistema de BI contribui para otimizar processos de negócios e recursos, maximizando lucros e melhorando a tomada de decisão proativa [10].

1.2 OBJETIVOS E CONTRIBUIÇÕES

Neste trabalho, objetiva-se incorporar estágios em sistemas existentes de BI do governo federal brasileiro, adicionando a estes capacidade de análise preditiva e melhorias de performance. O governo federal brasileiro possui um vasto número de sistemas de BI e, neste trabalho, dois destes sistemas serão utilizados. O primeiro sistema, mantido pela CGAUD, contém dados relativos à folha de pagamento dos servidores públicos federais. O segundo sistema, mantido pela SPU, contém dados relativos à arrecadação mensal de impostos daquele Órgão federal. Ambos os sistemas foram desenvolvidos com foco em detecção de fraudes e irregularidades, como a evasão fiscal e pagamentos não autorizados.

Métodos para detecção de irregularidades são classificados principalmente em duas categorias [11]. Uma é a detecção baseada no conhecimento, onde ocorrências fraudulentas são previamente definidas e categorizadas. Portanto, neste tipo de detecção, a irregularidade precisa ser conhecida e descrita *a priori* e o sistema normalmente não consegue lidar com tipos de irregularidades novos ou desconhecidos. Detecção de intrusão baseada no conhecimento em redes e sistemas de computadores são mostradas em [12].

Como alternativa, um esquema de detecção de fraudes baseado em comportamento assume que uma irregularidade possa ser detectada pela observação de ocorrências que são mais dissimilares que o normal [11]. Um comportamento válido e padrão pode ser extraído de informações prévias de referência, e este modelo pode ser comparado a um possível candidato fraudulento de modo a checar o grau de divergência entre eles. Em [13], os autores apresentam um sistema de detecção de fraudes em cartões de crédito cuja metodologia utiliza redes neurais treinadas com dados anteriores relacionados.

O sistema atual da CGAUD é baseado inteiramente em detecção de irregularidades por conhecimento. A CGAUD, subordinada ao MP, criou seu sistema de BI com o objetivo de detectar irregularidades na folha de pagamento dos servidores públicos federais. A proposta inicial do sistema de BI foi apresentada em [14] e diversos aperfeiçoamentos foram propostos em [15], [16] e [1*]. A versão mais recente do sistema de BI da CGAUD utiliza trilhas de auditoria construídas com indexação ontológica via mapas conceituais para detectar inconsistências [14, 15, 1*]. As trilhas de auditoria consistem em um conjunto de heurísticas baseado em uma legislação federal brasileira complexa, que dita o salário de cada servidor público federal de acordo com seu cargo e sua posição na administração pública.

Não obstante o complexo sistema legislativo, o quantitativo de dados gerados periodicamente referente à folha de pagamento dos servidores públicos federais é massivo: Por volta de 14GB de dados brutos por mês, e mais de 200 milhões de linhas na tabela da dados financeiros a cada ano [16]. Portanto, o custo de processamento para auditar esta quantidade de dados é

bastante elevado, dado que cada trilha de auditoria varre todo o banco de dados na busca por irregularidades.

De fato, enquanto a folha de pagamentos mensal dos servidores públicos federais brasileiros gira em torno de R\$12,5 bilhões, o atual sistema de BI da CGAUD é capaz de auditar aproximadamente R\$5 bilhões a cada mês [1*].

Neste cenário, nossa proposta é incorporar técnicas de modelagem de mistura finita com o objetivo de computar a distribuição de probabilidades das folhas de pagamento. É consenso que modelos de mistura constituem uma ferramenta probabilística versátil para representar a presença de subpopulações em conjuntos de observações. Desta forma, modelos de mistura facilitam uma descrição muito mais detalhada de sistemas complexos, ao passo que descrevem características diversas dos dados ao inferir todos os parâmetros de cada componente da mistura e explicando como este conjunto de fontes interagem para formar um modelo de mistura.

Evidências da versatilidade de modelos de mistura são demonstradas pela aplicação deste tópico em diversas áreas do conhecimento, como a astronomia [17], ecologia [18] e engenharia [19]. No contexto de sistemas de BI, modelos de mistura podem ser utilizados para representar funções de densidade de probabilidade arbitrariamente complexas [20]. Esta característica faz dos modelos de mistura uma escolha confiável para representar funções de verossimilhança complexas em cenários de aprendizado supervisionado [21], ou definições *a priori* em estimativas de parâmetros Bayesiana [22].

No sistema de BI da CGAUD, a hipótese que buscamos validar é a existência de uma relação direta entre a probabilidade de ocorrência das folhas de pagamento dos servidores públicos federais brasileiros e a detecção de fraudes atualmente existente nas trilhas de auditoria. Em outras palavras, contracheques improváveis - que mais divergem do padrão - possivelmente têm mais chances de possuírem algum tipo de irregularidade e de serem identificados pelas trilhas de auditoria.

Neste sentido, propomos uma abordagem estatística complementar, com um filtro gerativo baseado em MMG em um estágio de pré-processamento, com o objetivo de computar a probabilidade das folhas de pagamento e excluir as folhas mais prováveis das trilhas de auditoria subsequentes. Ao aprender um modelo de mistura que representa o comportamento mais provável das folhas de pagamento dos servidores públicos federais brasileiros, nós conseguimos executar uma seleção qualitativa no conjunto de todas as folhas de pagamento e entregar às trilhas de auditoria somente as folhas de pagamento que mais divergem da norma.

Esta nova abordagem aumentou a eficiência do sistema de BI, bem como a sua capacidade de processamento, com uma penalidade de perda de alguns falsos negativos neste estágio proposto.

No sistema de BI mantido pela SPU, propomos a adição de um módulo de análise preditiva com o objetivo de inferir o quantitativo de impostos a ser arrecadado por aquele Órgão federal. A

hipótese que buscamos validar é a possibilidade de aumentar a eficiência do algoritmo preditivo atualmente existente no sistema de BI da SPU - baseado em redes neurais artificiais - em termos de métricas de erro.

O modelo escolhido como núcleo da predição é baseado em RPG, uma família de processos estocásticos largamente utilizada na modelagem de dados interdependentes, primariamente devido a duas propriedades essenciais que ditam o comportamento da variável predita. Primeiro, um processo Gaussiano é completamente determinado por suas funções de média e de covariância, o que reduz a quantidade de parâmetros a serem especificados já que somente os primeiro e segundo momentos do processo são requeridos. Segundo, os valores preditos são função dos valores observados, onde todos os conjuntos de distribuições dimensionalmente finitas possuem uma distribuição Gaussiana multivariada.

Em um ambiente de BI, o fato de RPG retornar uma descrição estatística completa da variável predita pode adicionar confiança ao resultado final e ajudar na avaliação de sua própria performance. Ademais, a descrição estatística pode ser utilizada como gatilho para transformar um problema de regressão em um problema de classificação a depender do contexto. Quando lidamos com dados multidimensionais, RPG pode ser modelado de maneira independente em cada dimensão, o que adiciona flexibilidade para conjuntos de dados com diferentes graus de correlação entre suas dimensões.

Neste trabalho, nós utilizamos RPG para modelar a quantidade de imposto arrecadado mensalmente pela SPU. Considerando que a série temporal advinda da SPU possui uma estrutura multidimensional, ainda que oculta, foi desenvolvido neste trabalho um estágio de pré-processamento para reorganizar o conjunto original dos dados em uma estrutura bidimensional.

No sistema de BI da SPU, uma abordagem baseada em regressão preditiva utilizando processos Gaussianos infere a quantidade de imposto cuja probabilidade de ocorrência é a mais provável em um ponto de interesse. Ainda, como sistemas de BI voltados para a detecção de fraudes frequentemente requerem um estágio de classificação para etiquetar os dados em confiáveis ou possivelmente fraudulentos, nós demonstramos que RPG pode ser utilizado tanto no estágio preditivo quanto no estágio de classificação, com a utilização da descrição estatística da variável contínua predita como uma medida de gatilho para classificar os dados em regular ou possivelmente fraudulentos.

De acordo com [23], um problema de regressão pode ser visto como um problema de classificação onde o número de classes preditas tende ao infinito. De forma similar, pode ser dito que um problema de classificação pode ser resolvido com a aplicação de um conjunto de heurísticas em um ambiente de regressão para fatiar a variável contínua em um conjunto de classes com comprimento finito. Considerando também que RPG retorna não somente um valor predito, mas uma descrição estatística condicional completa para a variável estimada, o nosso foco recai na

regressão supervisionada.

RPG provê um ambiente completamente transparente, permitindo modelar uma relação de entrada-saída de um processo sem camadas ocultas ou nebulosas, restando possível adaptar características específicas de um sistema de detecção de fraude de modo a mantê-lo atualizado com um novo eventual comportamento fraudulento.

1.3 ORGANIZAÇÃO DESTE TRABALHO

No capítulo 3, expomos a fundamentação teórica que embasa este trabalho. Componentes chave em sistemas de BI e o estado da arte em aplicações no campo da detecção de fraudes são apresentados na Seção 3.1. A Seção 3.2 apresenta uma introdução a modelos de misturas finitas, com foco em MMG e no algoritmo de ME. Na Seção 3.3, um modelo preditivo genérico baseado em RPG é derivado, com a seleção da função de covariância e a otimização dos seus hiper-parâmetros.

No capítulo 4, descrevemos o atual sistema e dados de BI geridos pelo MP que receberam intervenções e melhorias ao longo deste trabalho. Na Seção 4.1, dissecamos o sistema de BI da CGAUD e exploramos a metodologia utilizada na auditoria da folha de pagamento de servidores públicos federais brasileiros. Na Seção 4.2, mostramos os dados de BI da SPU, que é constituído de uma série temporal histórica com a arrecadação de impostos efetuada por aquele Órgão.

No capítulo 5, desenvolvemos métodos para a aplicação de MMG no sistema de BI da CGAUD. Na Seção 5.1, analisamos as implicações de um módulo de análise estatística incorporado em um sistema de BI determinístico, *i.e.* o que o sistema tem a ganhar e quais são as possíveis armadilhas neste tipo de abordagem. Na Seção 5.2, discorremos sobre a otimização e os resultados experimentais desta aplicação específica.

No capítulo 6, propomos um módulo de análise preditiva com RPG em seu núcleo para os dados de BI da SPU. Na Seção 6.1, desenvolvemos um modelo preditivo unidimensional que captura as características intrínsecas dos dados de BI da SPU. Na Seção 6.2, propomos uma abordagem diferente para o módulo preditivo, transformando o conjunto de dados original em um conjunto de dados bidimensional. Na Seção 6.3, os resultados da otimização são apresentados e os resultados experimentais são comparados com outras técnicas preditivas por meio de diversas métricas de erro.

O Capítulo 7 desenha algumas conclusões e considerações sobre as realizações e resultados deste trabalho, com sugestões de desenvolvimentos futuros.

Chapter 2

INTRODUCTION

2.1 CONTEXT AND MOTIVATION

Knowledge is power. In corporations and governmental institutions, high-level management needs business intelligent information to efficiently manage business and institutional operations and support their process of decision making [1]. In this domain of expertise, BI has evolved as an important field of research. Furthermore, outside of the academic world, BI has been recognized as a strategic initiative and a key enabler for effectiveness and innovations in several practical applications in the business universe.

In this context, advances in technology has massively increased the volume of electronic data available, with about 2.5 EB of digital data being created each day in the world, and that number is doubling every 40 months approximately [2]. Fig. 2.1 shows the increase trend of data in volume each year. On the other hand, a great part of this new data lacks structure. Organize and analyze this rising volume of raw data and find meaningful and useful information in its content are key points in BI systems.

On this topic, Hal Varian, Chief Economist at Google and emeritus professor at the University of California, Berkeley, commented: “So what’s getting ubiquitous and cheap? Data. And what is complementary to data? Analysis. So my recommendation is to take lots of courses about how to manipulate and analyze data: databases, machine learning, econometrics, statistics, visualization, and so on.” [4].

In addition to the increasing volume of data, big data management also deals with data variety, velocity, variability and value. It was in [5] that data management was first shown as a three dimensional scheme, with volume, velocity and variety independently contributing to the development of tools and algorithms for handling and managing data. In [6], the authors add two other dimensions into this perspective: variability and value. These data properties are

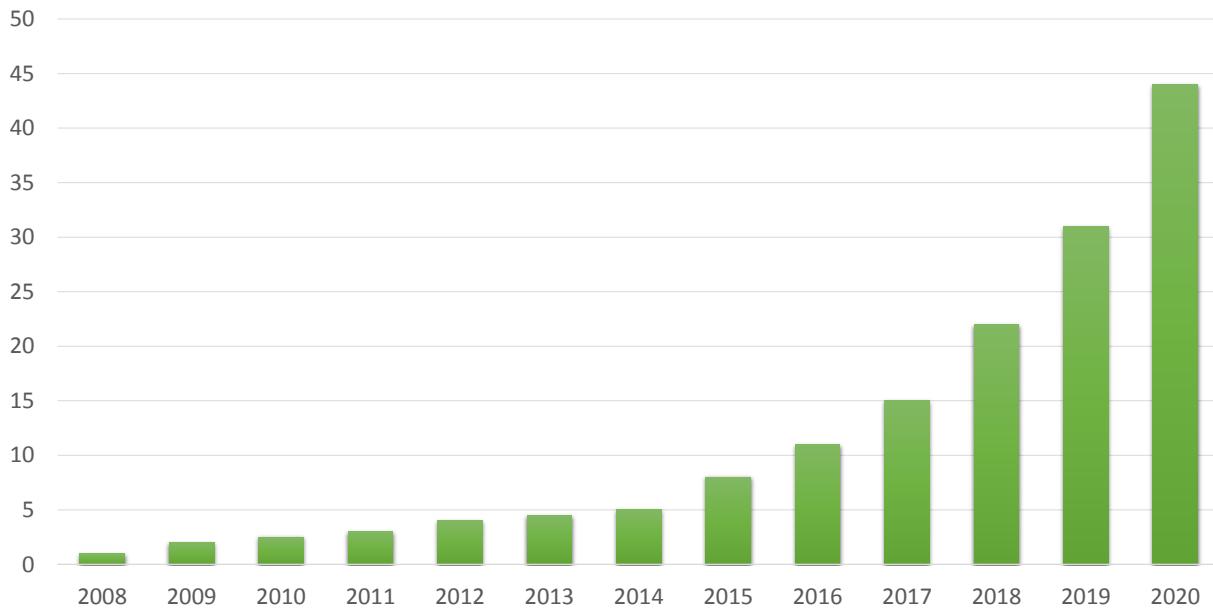


Figure 2.1: Estimation of the total volume of electronically stored data in ZB along the years. With a growing compound annual rate of 40 percent, stored data is estimated to reach nearly 45 ZB by 2020 [3].

summarized in Table 2.1.

With all this data getting cheap and ubiquitous, new capture, search, discovery, and analysis tools can help organizations and governmental institutions gain insights from their unstructured data, which accounts for more than 90% of the digital universe [8]. These tools can be programmed to create data about data in an automated way, much like facial recognition routines that help tag Facebook photos. Data about data, or metadata, is growing twice as fast as the digital universe as a whole [8].

As a consequence of this scenario, where more and more data are stored and processed electronically and vast amounts of information are becoming easily available and retrievable, the decision making cycles in enterprises and governmental organizations are considerably shorter in our days. Strategic decision makers are being exposed to huge inflows of data and information and are constantly under pressure to respond to exceptional situations and to take advantage of time-sensitive business opportunities.

In this environment, BI systems unite several areas of knowledge to deliver information and knowledge to business and administrations, supporting the decision making on all management levels. BI systems are referred to as an integrated set of tools and technologies that are used to collect, integrate, analyze and make data available [9]. They differ from traditional management information systems by - first of all - a wider subject range, multi-variant analyses of semi-structured data that come from different sources and their multi-dimensional presentation [10].

Table 2.1: The five V's of data properties

Property	Description
Volume	Data volume measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it. As data volume increases, the value of different data records will decrease in proportion to age, type, richness, and quantity among other factors [6].
Velocity	Data velocity measures the speed of data creation, streaming, and aggregation. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows. For example, the data from the sensor devices would be constantly moving to the database store and this amount would not be small enough. Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion [7].
Variety	Data variety is a measure of the richness of the data representation - text, images video, audio, etc. From an analytic perspective, it is mainly the biggest obstacle to effectively using large volumes of data. Incompatible data formats, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic sprawl [6].
Variability	It is quite an undertaking to link, match, cleanse and transform data across systems coming from various sources. It is also necessary to connect and correlate relationships, hierarchies and multiple data linkages or data can quickly spiral out of control [7].
Value	Data value measures the usefulness of data in making decisions. It has been noted that the purpose of computing is insight, not numbers. Data science is exploratory and useful in getting to know the data, but analytic science encompasses the predictive power of big data [6].

The BI systems contribute to optimizing business processes and resources, maximizing profits and improving proactive decision making [10].

2.2 OBJECTIVES AND CONTRIBUTIONS

In this work, we aim to incorporate stages into existing BI systems of the Brazilian federal government in order to add predictive analytics and performance enhancements. The Brazilian federal government possesses a wide number of BI systems and, in this work, two of those systems are used. The first system, maintained by CGAUD, contains data regarding the payroll of Brazilian federal staff. The second system, maintained by SPU, contains data regarding the monthly tax

collection of that federal department. Both systems were designed aimed at fraud and irregularities detection such as tax evasion and unauthorized payments.

Methods for irregularities detection are mainly classified in two categories [11]. One is knowledge-based detection, where fraudulent occurrences are previously defined and categorized. Thus, in this kind of detection, the irregularity must be known and described *a priori* and the system is usually unable to deal with new or unknown irregularities. Knowledge-based intrusion detection schemes in network and computer systems are shown in [12].

Alternatively, a behavior-based fraud detection scheme assumes that an irregularity can be detected by observing occurrences that are most dissimilar from the norm [11]. A valid and standardized behavior can be extracted from previous reference information, and this model can be compared to a fraudulent candidate in order to check for the degree of divergence between them. In [13], the authors present a credit card fraud detection method using neural networks trained with previous related data.

The current CGAUD BI system is entirely based on a knowledge-based approach for irregularity detection. CGAUD, subordinated to MP, created its own BI system with the objective of detect irregularities on the payrolls of the Brazilian federal staff. The initial BI solution was presented in [14] and several improvements were proposed in [15], [16] and [1*]. The most recent BI system of CGAUD uses audit trails built with ontological indexation via concept maps in order to detect inconsistencies [14, 15, 1*]. The audit trails consist on a set of heuristics based on a complex Brazilian federal legislation, which dictates the income of each public employees according to their position in the public administration organization.

In addition to a complex regulatory basis, the amount of data periodically generated regarding the payroll of federal employees is massive: Around 14GB of raw data per month, and more than 200 million rows in the financial data table each year [16]. Thus, the processing cost of auditing this amount of data is very high, since each audit trail has to go through all database performing relational statements on the search for irregularities.

In fact, whereas the monthly payroll of the Brazilian federal staff is around 12.5 billion *reais*, the current BI system of CGAUD is capable of auditing approximately 5 billion *reais* each month [1*].

In this scenario, we propose to incorporate *finite mixture models* techniques in order to compute the pdf of payrolls. It is known that mixture models constitute a versatile probabilistic tool for representing the presence of subpopulations within a set of observations. They thus facilitate a much more detailed description of complex systems, describing different features of the data by inferring all the parameters of each component of the mixture and by explaining how the set of sources interact together to form a mixture model.

Evidences of the versatility of mixture models is their application in diverse areas, such as

astronomy [17], ecology [18] and engineering [19]. In the context of BI systems, mixture models can be used to represent arbitrarily complex probability density functions [20]. This characteristic makes them a reliable choice for representing complex likelihood functions in supervised learning scenarios [21], or priors for Bayesian parameter estimation [22].

In the BI system of CGAUD, the hypothesis we seek to test is the existence of a direct relationship between the probability of occurrence of payrolls of the Brazilian federal employees and the fraud detection currently performed by the audit trails. In other words, unlikely payrolls - that most diverge from the norm - have more chances of being detected by the audit trails for some irregularity.

Under this perspective, we propose a complementary statistical approach, with a generative GMM filter in a pre-processing stage with the objective of compute payrolls with low probability of occurrence as being irregular and exclude them of the following audit trails. By learning a mixture model that represents the most probable behavior of the payrolls of the Brazilian federal staff, we are able to perform a selection on all payrolls and deliver to the audit trails only payrolls that diverge the most from the norm.

This new approach significantly increases the efficiency of the BI system and its processing capacity, with a penalty of losing a few false negatives at this proposed stage.

On the other hand, in the BI system maintained by SPU, we propose to add a predictive analytics module in order to forecast the amount of tax to be collected by that federal organization. The hypothesis we intent to validate is the possibility to improve the error rates of the predictive algorithm currently employed in the BI system of SPU, which is based on artificial neural networks.

The model chosen as the core predictor is based on GPR, a widely used family of stochastic process schemes for modeling dependent data primarily due two essential properties that dictate the behavior of the predicted variable [24]. First, a Gaussian process is completely determined by its mean and covariance functions, which reduces the amount of parameters to be 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 sets have a multivariate Gaussian distribution [25].

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 own performance. Additionally, the statistical description can be used as a trigger to transform a regression problem into a classification problem depending on the context. When dealing with multidimensional data, GPR can be independently modeled in each dimension, which adds flexibility for data sets with different degrees of correlation among its dimensions.

In this work, we use GPR for modeling the amount of tax collected monthly by SPU. Con-

sidering that the time series provided by SPU possess a latent multidimensional structure, we propose a pre-processing stage to reshape that original data set into a bidimensional structure.

In the BI system of SPU, a supervised regression method to predict the amount of tax collected at a given period by SPU is proposed. An approach based on predictive regression using Gaussian processes was developed to forecast the amount of tax that is most likely to occur at the point of interest. Furthermore, as BI systems aimed at fraud detection often requires a classification stage to label trusted and possibly fraudulent data, we show that GPR can be used both in the predictive and the classification stages using the statistical description of the continuous predicted variable as a trigger measure to classify regular or possibly fraudulent data.

According to [23], a regression problem can be seen as a classification problem where the number of predicted classes tends to infinity. Similarly, one can say that a classification problem can be solved by applying a set of heuristics into a regression environment to break the continuous variable into a finite-length set of classes. Considering also that GPR returns not only a predicted value, but a full conditional statistical description of the estimated variable, we focus on supervised regression.

GPR provides a complete transparent environment, allowing to model the input-output relationship of a process with no hidden or nebulous layers, making possible to adapt specific characteristics of a fraud detection system to keep in tune with an eventual new fraud behavior.

2.3 ORGANIZATION OF THIS WORK

In Chapter 3, we expose the theoretical foundation on which this work is based. Key components in BI systems and the state-of-the-art applications in the field of fraud detection are shown in Section 3.1. Section 3.2 presents a gentle introduction to finite mixture models, focusing on GMM and the EM algorithm. In Section 3.3, a generic predictive model based on GPR is derived, with covariance function selection and hyperparameters tuning.

In Chapter 4, we describe the current BI systems and data managed by MP that have received improvements and enhancements throughout this work. In Section 4.1, we dissect the BI system of CGAUD and explore the methodology used to audit payrolls of Brazilian federal employees. In Section 4.2, we show the BI data of SPU, which is consisted of a time series with historical federal tax collected.

In Chapter 5, we develop methods to apply GMM in the BI system of CGAUD. In Section 5.1, we analyze what are the implications of a statistical module on an entire deterministic BI system, *i.e.* what the system has to gain and what are the possible pitfalls in this approach. In Section 5.2, we go through the optimization and experimental results for this specific application.

In Chapter 6, we propose a predictive analytics module, with GPR at its core, for the BI data of SPU. In Section 6.1, we develop a unidimensional predictor model that captures the intrinsic characteristics of the BI data of SPU. In Section 6.2, we propose a different approach to the predictive module by transforming the original data set into a bidimensional data set. In Section 6.3, the optimization results are shown and the experimental results are compared with other predictive approaches by several error metric.

Chapter 7 draws some conclusions and thoughts about the accomplishments of this work, with future development suggestions.

Chapter 3

THEORETICAL FOUNDATION

3.1 BUSINESS INTELLIGENCE

The term BI has been used since late 1960s [26], where the author defined *Business* as a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera. The communication facility serving the conduct of a business, in the broad sense, may be referred to as an *intelligence system*. The notion of *intelligence* is also defined, in a more general sense, as “the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal” [26].

BI, as it is understood today, is said to have evolved from the decision support systems which began in the 1960s and developed throughout the 80s [27]. According to [28], the term BI was coined in 1989 as an umbrella term to describe “concepts and methods to improve business decision making by using fact-based support systems.” It was not until the late 1990s that this usage was widespread [27].

A BI system combines data gathering, data storage, and knowledge management with analytical tools to present complex and competitive information to planners and decision makers [29]. The objectives are to enable business managers and analysts of all levels to readily access any data in the organization and to conduct appropriate manipulation and analysis [30]. Implicit in this definition is the notion that a BI system can improve the timeliness and quality of the input to the decision making process [31].

In late 2000s, the authors in [32] argued that BI systems are composed of a set of three complementary data management technologies, namely data warehousing, OLAP and business analytics. In this definition, *business analytics* is the subset of BI systems related to knowledge discovery, which is predominantly aided by ML, statistics, prediction, and optimization techniques. More recently, big data and big data analytics have been used to describe the data sets and analytical

techniques in applications that are so large and complex that they require advanced and unique data storage, management, analysis, and visualization technologies [33].

3.1.1 Key Components

From a process point of view, BI systems can be divided into two primary activities: insert data into the system and extract information and knowledge out of the system [34]. The key components of a BI system framework is summarized in Table 3.1 [35].

Table 3.1: Key components in BI systems framework

Layer	Description
Data Source	Manages the external sources of data, operational databases and transaction systems
Data Integration	ETL tools, that are responsible for data transfer from data sources to data warehouses
Data Storage	Data warehouses, to provide some room for thematic storing of aggregated data
Data Analysis	Knowledge discovery tools, which determines patterns, generalizations, probabilities, regularities and heuristics in data resources
Data Presentation	Reporting and ad hoc inquiry tools, for different synthetic reports and customized graphical and multimedia interfaces

The traditional architecture of the key components in generic BI systems is shown in Fig. 3.1.

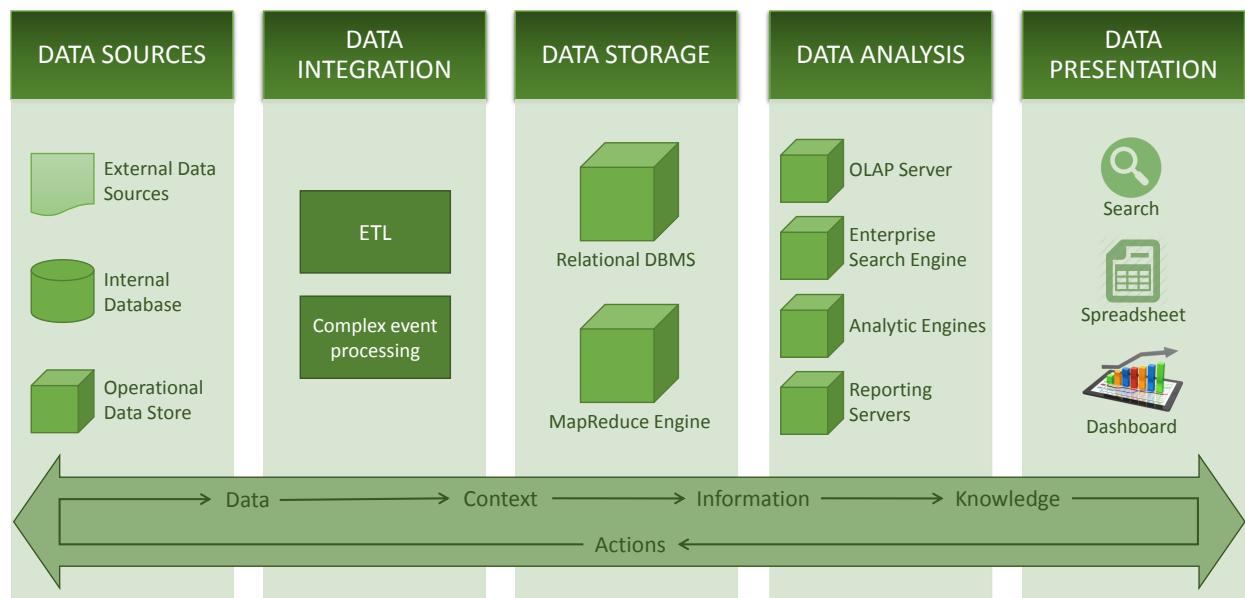


Figure 3.1: A traditional architecture and components of a generic BI system

Usually, the first conceptual step in the development of a BI system regards its data integration layer, which often translates into an ETL process, although some pre-processing stages can be performed when the data is coming from an external data source, such as the use of an ODS [36].

ETL refers to a collection of tools that plays 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 data analysis depends on being able to efficiently get high-quality data into the DW from operational databases and external data sources [37].

In the data storage layer, DW store current and historical data, which are later used for creating trending reports for senior management. Other basic elements in the data storage layer are the Data Marts. They are subsets of data stored at the DW, and are devoted to respond for a necessity to work with a specific population.

The architecture of the Data Mart and the DW is a main concept that will impact on the performance of the system [29]. Whether a bottom-up structure proposed by [38], or a top-down structure proposed by [39], or even if big data technology [29] will be incorporated to the system make a significant difference in the system responsiveness.

The data analysis layer contains the business analytics tools aimed at extracting knowledge from the stored data [29]. OLAP tools enable users to analyze multidimensional data interactively from multiple perspectives, whereas analytic engines, commonly based on ML techniques, seeks for patterns, classes, statistical parameters and other relevant characteristics on the data.

The presentation layer converts the raw knowledge information to different reports and customized interfaces for each different end user at any level of the organization [29]. Role based BI is a concept that suggests that it is not necessary to drown people with information, but rather delivery just the information they need, customized to their function. The architecture should support every level of end user, including external consumers to the organization [29].

3.1.2 Fraud Detection Applications

In the context of BI systems, fraud detection schemes is a continuously evolving topic. In 2012, global credit, debit and prepaid card fraud losses reach \$11.27 billion [40]. Of that, card issuers lost 63% and acquires lost the other 37% [40].

In a competitive environment, fraud can become a business critical problem if it is very prevalent and if the prevention procedures are not fail-safe [41]. Fraud detection, being part of the overall fraud control, has become one of the most established industrial and governmental data mining applications [41].

Fraud is a perpetually changing enterprise. When a new fraud detection scheme becomes

public domain, criminals are likely to use this information to evade themselves of this type of detection, limiting the public exchange of ideas regarding this topic [11].

In many applications, BI systems aimed at fraud detection deal with huge data sets. For example, business general-purpose credit card transactions in the United States reached 3.4 billion in 2012 [42]. Search for fraudulent transactions in such data sets makes data mining techniques relevant, requiring more than state-of-the-art statistical models [11].

The need for fast and efficient algorithms makes automated fraud detection techniques widely varied, but there are common features. Essentially, those methods compare observed or estimated data with expected values [11].

In addition, automated fraud detection methods can be divided in supervised and unsupervised. Supervised methods use samples of known to be either fraudulent and nonfraudulent data in order to construct a model that classifies new data into one of those two classes [11]. In this case, the objective is to obtain a model to maximize the differences between fraudulent and nonfraudulent data, which requires a high confidence about the records used as fraudulent and trustable. Also, the use of supervised methods can only be applied to detect types of fraud that have previously occurred or simulated.

On the other hand, unsupervised methods seek for samples that are most dissimilar from the norm [11]. In this case, the goal is to model the normal behavior of the monitored environment and to establish a quantifiable measure that segregates a possibly fraudulent event. Frequently, unsupervised methods are used to alert the fact that an observation is anomalous and requires a closer investigation.

Some of the most commonly used techniques for automated fraud detection applications are:

- Data preprocessing techniques for detection, validation, error correction, and filling up of missing or incorrect data.
- Calculation of various statistical parameters such as averages, quantiles, performance metrics, probability distributions, etc..
- Models and probability distributions of various business activities either in terms of various parameters or probability distributions.
- Computing user profiles.
- Time-series analysis of time-dependent data.
- Clustering and classification to find patterns and associations among groups of data.

- Matching algorithms to detect anomalies in the behavior of transactions or users as compared to previously known models and profiles. Techniques are also needed to eliminate false alarms, estimate risks, and predict future of current transactions or users.
- Data mining to classify, cluster, and segment the data and automatically find associations and rules in the data that may signify interesting patterns, including those related to fraud.
- Expert systems to encode expertise for detecting fraud in the form of heuristics.
- Pattern recognition to detect approximate classes, clusters, or patterns of suspicious behavior either unsupervised or to match given inputs.
- Machine learning techniques to automatically identify characteristics of fraud.
- Neural networks that can learn suspicious patterns from samples and used later to detect them.

Automated fraud detection approaches have been used in [43], where statistical analysis were used to detect medicaid¹ claim fraudulent requests; in [44], where an ANN is used for fraud detection in credit card operations; in [45], where an ANN based predictor was used in real world BI data for forecasting a time series and heuristics based on error metrics decides if the predicted data is possibly fraudulent or regular. In [46], supported vector machines and genetic algorithms are used to identify electricity theft.

3.2 FINITE MIXTURE MODELS

A convex combination of two or more pdf is a *mixture*. The approximation of any arbitrary distribution can be achieved by the combination of the properties of a set of individual pdf [47], making mixture models a powerful tool for modeling complex data. While within a parametric family, mixture models offer malleable approximations in non-parametric settings and, although based on standard distributions, mixture models pose highly complex computational challenges [48].

To accompany our model, let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_L)$ be an unlabeled random sample obtained in an iid manner. The pdf of a mixture model is defined as

$$p(\mathbf{x}|\Theta) = \sum_{k=1}^K \alpha_k p_k(\mathbf{x}|\theta_k) \quad k = 1, \dots, K, \quad (3.1)$$

¹Medicaid is a social health care program for families and individuals with low income and limited resources in the United States. Please refer to <http://www.medicaid.gov> for further details.

where $\mathbf{x} = (x_1, \dots, x_d)^T \in \mathbb{R}^d$ is a set of d observed random samples, $K \in \mathbb{Z}^+$ is the number of components (sources) in the mixture, $p_k(x|\theta_k)$ is the pdf of the k^{th} component and $\Theta = (\alpha_1, \dots, \alpha_K, \theta_1, \dots, \theta_K) \in \Omega$ is the set of parameters of the mixture, with Ω being the parameter space of all possible combinations of values for all the different parameters of the mixture [49]. The collection α_k is the mixing proportion (or weighting factor) of the k^{th} component, representing the probability that a randomly selected $\mathbf{x}_i \in \mathcal{X}$ was generated by the k^{th} component.

In the particular case of a Gaussian mixture model, (3.1) can be written as

$$p(\mathbf{x}|\Theta) = \sum_{k=1}^K \alpha_k p_k(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (3.2)$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^d$ is the mean vector and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$ is the covariance matrix, both of them originated by the k^{th} Gaussian component. Each of those component density is a Gaussian function of the form

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\}. \quad (3.3)$$

Note that the Gaussian mixture model is completely parametrized by its mean vectors, covariance matrices and mixture weights from all component densities [50].

Given that (3.1) and (3.2) represent a convex combination of K distributions [47], it can be stated that

$$\alpha_k \geq 0, \text{ for } k \in \{1, \dots, K\}, \text{ and}$$

$$\sum_{k=1}^K \alpha_k = 1. \quad (3.4)$$

In addition, since each $p_k(\mathbf{x}|\theta_k)$ defines a pdf, $p(\mathbf{x}|\Theta)$ will also be a pdf [47].

One straightforward interpretation of mixture models is that (3.1) describes a complete stochastic model [51], thus giving us a recipe to generate new data points. Another point of view, in the mixture model context, is that any observed data sample is generated from a combination of K distinct random processes, each one modeled by the density $p_k(x|\theta_k)$, with α_k defining the proportion of a particular random process in the overall observations.

3.2.1 Estimation of Parametric Mixture Models

Once defined the mixture model and the particular case of a GMM, next a numerical approach that will allow the estimate of the parameters set $\Theta = (\theta_1, \theta_2, \dots, \theta_k)$ is presented.

Let $\mathbf{X} \in \mathbb{R}^{N \times L}$ be a set of N unlabeled observations, where \mathbf{x}_{ik} is the value of the i^{th} observation for the k^{th} component. Since the observed set \mathbf{X} is iid, the joint pdf for \mathbf{X} can be written as [52]

$$p(\mathbf{X}|\Theta) = \prod_{i=1}^N p(\mathbf{x}_i|\theta_1, \dots, \theta_k). \quad (3.5)$$

The likelihood function of the data, also assuming that \mathbf{x}_i are independently distributed, is defined as

$$p(\mathbf{X}|\Theta) = \mathcal{L}(\Theta|\mathbf{X}) = \prod_{i=1}^N \sum_{k=1}^K \alpha_k p_k(\mathbf{x}_i|\theta_k). \quad (3.6)$$

The likelihood can be thought of as a function of the parameters Θ where the observed data \mathbf{X} is fixed. In the maximum likelihood problem, our goal is to find the Θ that maximizes $\mathcal{L}(\Theta|\mathbf{X})$, thus determining which parameters values are more likely for the observed values [53]:

$$\Theta^* = \arg \max_{\Theta \in \Omega} \mathcal{L}(\Theta|\mathbf{X}). \quad (3.7)$$

In general cases, it is often preferable to maximize $\log(\mathcal{L}(\Theta|\mathbf{X}))$ instead, since it is analytically easier [53]. However, in many scenarios an analytical solution is not possible to develop. One alternative is to maximize the likelihood in an EM approach.

3.2.2 Expectation Maximization Algorithm

The EM algorithm is an iterative method for estimating the maximum likelihood of a stochastic model where exists a dependency upon latent, or unobserved, data [54].

Throughout the remainder of this subsection, the EM algorithm is used to obtain an accurate approximation of the maximum likelihood of a mixture model which has *incomplete* data associated with it. This consideration is taken into account when optimizing the likelihood function is analytically intractable, but the likelihood function can be simplified by assuming the existence of additional but missing values [53].

Therefore, let \mathcal{X} be a random incomplete observed data set, \mathcal{Y} be a random unobserved data set and $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ be a *complete* data set.

To establish the notation, let $p(\mathbf{z}|\Theta) = p(\mathbf{x}, \mathbf{y}|\Theta)$ be the joint pdf of the random variables \mathcal{X} and \mathcal{Y} , $g(\mathbf{x}|\Theta)$ be the marginal pdf of \mathcal{X} and $k(\mathbf{y}|\mathbf{x}, \Theta)$ be the conditional probability of \mathcal{Y} given $\mathcal{X} = \mathbf{x}$.

The EM algorithm aims to maximize the incomplete data log-likelihood [54],

$$\log[\mathcal{L}(\Theta|\mathcal{X})] = \log[g(\mathbf{x}|\Theta)] \quad \text{for } \Theta \in \Omega,$$

by using $p(\mathbf{x}, \mathbf{y}|\Theta)$ and $g(\mathbf{x}|\Theta)$. From Bayes' rule, $p(\mathbf{z}|\Theta)$ can be represented as

$$p(\mathbf{z}|\Theta) = p(\mathbf{x}, \mathbf{y}|\Theta) = k(\mathbf{y}|\mathbf{x}, \Theta) \cdot g(\mathbf{x}|\Theta), \quad (3.8)$$

for $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$.

The E-step of the EM algorithm seeks to find the expected value of the complete data log-likelihood, defined as

$$\log[\mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y})] = \log[p(\mathbf{x}, \mathbf{y}|\Theta)]. \quad (3.9)$$

In (3.9), the observed samples \mathcal{X} and some *a priori* parameter estimate $\Theta^p \in \Omega$ are given as inputs. In addition, an auxiliary function \mathcal{Q} is defined such as

$$\mathcal{Q}(\Theta|\Theta^p) = \mathbb{E}[\log[p(\mathbf{x}, \mathbf{y}|\Theta)|\mathbf{x}, \Theta^p]], \quad (3.10)$$

where $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$, $\Theta^p \in \Omega$ and $\mathbb{E}[\cdot]$ denotes the expectation operator. The key thing in (3.10) is that \mathcal{X} and Θ^p are constants, Θ is a regular variable which we want to optimize and \mathcal{Y} is a random variable governed by the distribution $k(\mathbf{y}|\mathbf{x}, \Theta)$.

The M-step of the EM algorithm intents to maximize (3.10) by selecting a new set of parameters $\Theta^* \in \Omega$ such that

$$\Theta^* \in \arg \max_{\Theta \in \Omega} \mathcal{Q}(\Theta|\Theta^p). \quad (3.11)$$

The EM algorithm presented in [54] can abstractly be summarized as follows:

1. E-Step: Calculate $\mathcal{Q}(\Theta|\Theta^p)$.
2. M-Step: Pick $\Theta^* \in \arg \max_{\Theta \in \Omega} \mathcal{Q}(\Theta|\Theta^p)$.
3. $\Theta^p \leftarrow \Theta^*$.
4. Iterate (1)-(3) until some convergence criterion is met.

At each iteration, the EM algorithm increases the log-likelihood converging to a local maximum [55]. More properties on convergence of the EM algorithm can be found at [54] and [56].

3.3 GAUSSIAN PROCESS FOR REGRESSION

Gaussian processes belong to the family of stochastic processes schemes that can be used for modeling dependent data observed over time and/or space [25]. Our main interest relies on 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 order to make predictions based on a finite data set, a function h needs to link the known sets of the training data with all the other possible sets of input-output values. The characteristics of this underlying function h can be defined in a wide variety of ways [57], and that is where Gaussian processes are applied. Stochastic processes, as the Gaussian process, dictate the properties of the underlying function as well as probability distributions govern the properties of a random variable [25].

Two properties make Gaussian processes 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 [58].

3.3.1 Multivariate Gaussian Distribution

Consider a multivariate random variable $\mathbf{x} \in \mathbb{R}^n$. If \mathbf{x} has a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{S}_{++}^n$, where \mathbb{S}_{++}^n is the space of symmetric positive definite $n \times n$ matrices², the pdf of \mathbf{x} has the form [58]:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right], \quad (3.12)$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

To denote a random variable to be Gaussian distributed we write:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (3.13)$$

Now, consider that the random vector $\mathbf{x} \in \mathbb{R}^n$, with $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is partitioned into two sets \mathbf{x}_A and \mathbf{x}_B , such that:

$$\begin{aligned} \mathbf{x}_A &= [x_1, x_2, \dots, x_r]^T \in \mathbb{R}^r, \\ \mathbf{x}_B &= [x_{r+1}, x_{r+2}, \dots, x_n]^T \in \mathbb{R}^{n-r}. \end{aligned} \quad (3.14)$$

²In some cases, $\boldsymbol{\Sigma}$ can be positive semidefinite but not positive definite, such as the case where $\boldsymbol{\Sigma}$ is not full rank. In those cases, $\boldsymbol{\Sigma}^{-1}$ does not exist, and the definition given by (3.12) does not apply.

Similarly, the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ can also be partitioned into two sets, resulting in:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB} \end{bmatrix}, \quad (3.15)$$

where $\boldsymbol{\Sigma}_{AA} = \text{var}(\mathbf{x}_A)$, $\boldsymbol{\Sigma}_{BB} = \text{var}(\mathbf{x}_B)$ and $\boldsymbol{\Sigma}_{AB} = (\boldsymbol{\Sigma}_{AB})^T = \text{cov}(\mathbf{x}_A, \mathbf{x}_B)$.

In Subsections 3.3.1, 3.3.1 and 3.3.1 some useful properties for (3.15) are proven.

Marginalization

Given the marginal density function of \mathbf{x}_A and \mathbf{x}_B ,

$$\begin{aligned} f(\mathbf{x}_A) &= \int_{\mathbf{x}_B} f(\mathbf{x}) d\mathbf{x}_B, \\ f(\mathbf{x}_B) &= \int_{\mathbf{x}_A} f(\mathbf{x}) d\mathbf{x}_A, \end{aligned} \quad (3.16)$$

it can be stated that the densities in (3.16) are Gaussian distributed [58]. Therefore, it is possible to write:

$$\begin{aligned} \mathbf{x}_A &\sim \mathcal{N}(\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_{AA}), \\ \mathbf{x}_B &\sim \mathcal{N}(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_{BB}). \end{aligned} \quad (3.17)$$

Conditioning

Given the conditional density function of $\mathbf{x}_A|\mathbf{x}_B$ and $\mathbf{x}_B|\mathbf{x}_A$,

$$\begin{aligned} f(\mathbf{x}_A|\mathbf{x}_B) &= \frac{f(\mathbf{x})}{f(\mathbf{x}_B)}, \\ f(\mathbf{x}_B|\mathbf{x}_A) &= \frac{f(\mathbf{x})}{f(\mathbf{x}_A)}; \end{aligned} \quad (3.18)$$

their conditional densities in (3.18) are also Gaussian distributed [58], which leads to:

$$\begin{aligned} \mathbf{x}_A|\mathbf{x}_B &\sim \mathcal{N}(\boldsymbol{\mu}_{A|B}, \boldsymbol{\Sigma}_{A|B}), \\ \mathbf{x}_B|\mathbf{x}_A &\sim \mathcal{N}(\boldsymbol{\mu}_{B|A}, \boldsymbol{\Sigma}_{B|A}); \end{aligned} \quad (3.19)$$

where

$$\begin{aligned} \boldsymbol{\mu}_{A|B} &= \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}^{-1}(\mathbf{x}_B - \boldsymbol{\mu}_B), \\ \boldsymbol{\mu}_{B|A} &= \boldsymbol{\mu}_B + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}(\mathbf{x}_A - \boldsymbol{\mu}_A), \\ \boldsymbol{\Sigma}_{A|B} &= \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{BA}, \\ \boldsymbol{\Sigma}_{B|A} &= \boldsymbol{\Sigma}_{BB} - \boldsymbol{\Sigma}_{BA}\boldsymbol{\Sigma}_{AA}^{-1}\boldsymbol{\Sigma}_{AB}. \end{aligned} \quad (3.20)$$

Summation

The sum of independent multivariate Gaussian random variables with the same dimensionality results in another multivariate Gaussian random variable [58]. As an example, let $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$. The sum of \mathbf{y} and \mathbf{z} , provided they are independent, can be stated as:

$$\mathbf{y} + \mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\mu}', \boldsymbol{\Sigma} + \boldsymbol{\Sigma}'). \quad (3.21)$$

3.3.2 Gaussian Processes

Multivariate Gaussian distributions are useful for modeling finite collections of real-valued random variables due to their analytical properties showed in Subsection 3.3.1. *Gaussian processes* extend this scenario, evolving from distributions over random vectors to distributions over random functions.

A stochastic process is a collection of random variables, *e.g.* $\{h(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, defined on a certain probability space and indexed by elements from some set [59]. Just as a random variable assigns a real number to every outcome of a random experiment, a stochastic process assigns a sample function to every outcome of a random experiment [59].

A Gaussian process is a stochastic process where any finite subcollection of random variables has a multivariate Gaussian distribution. In other words, a collection of random variables $\{h(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ is a Gaussian process with mean function $m(\cdot)$ and covariance function $k(\cdot, \cdot)$ if, for any finite set of elements $\{x_1, x_2, \dots, x_n \in \mathcal{X}\}$, the associated finite set of random variables $\{h(x_1), h(x_2), \dots, h(x_n)\}$ have a distribution of the form:

$$h(\mathbf{x}) = \begin{bmatrix} h(x_1) \\ h(x_2) \\ \vdots \\ h(x_n) \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} m(x_1) \\ m(x_2) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & \cdots & k(x_2, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix} \right). \quad (3.22)$$

The notation for defining $h(\mathbf{x})$ as a Gaussian process is

$$h(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \quad (3.23)$$

for any \mathbf{x} and $\mathbf{x}' \in \mathcal{X}$. The mean and covariance functions are given, respectively, by:

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[\mathbf{x}], \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(\mathbf{x} - m(\mathbf{x}))(\mathbf{x}' - m(\mathbf{x}'))]; \end{aligned} \quad (3.24)$$

also for any \mathbf{x} and $\mathbf{x}' \in \mathcal{X}$.

Intuitively, a sample function $h(\mathbf{x})$ drawn from a Gaussian process can be seen as an extremely high dimensional vector obtained from an extremely high dimensional multivariate Gaussian, where each dimension of the multivariate Gaussian corresponds to an element x_k from the index \mathcal{X} , and the corresponding component of the random vector represents the value of $h(x_k)$ [25].

3.3.3 Regression Model and Inference

Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, $\mathbf{x} \in \mathbb{R}^n$ and $y \in \mathbb{R}$, be a training set of iid samples from some unknown distribution. In its simplest form, GPR models the output nonlinearly by [24]:

$$y_i = h(\mathbf{x}_i) + \nu_i; \quad i = 1, \dots, m \quad (3.25)$$

where $h(\mathbf{x}) \in \mathbb{R}^m$. An additive iid noise variable $\nu \in \mathbb{R}^m$, with $\mathcal{N}(0, \sigma^2)$, is used for noise modeling. Other noise models can be seen in [60]. Assume a prior distribution over function $h(\cdot)$ being a Gaussian process with zero mean:

$$h(\cdot) \sim \mathcal{GP}(0, k(\cdot, \cdot)), \quad (3.26)$$

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

$$\mathbf{X} = \begin{bmatrix} (\mathbf{x}_1)^T \\ (\mathbf{x}_2)^T \\ \vdots \\ (\mathbf{x}_m)^T \end{bmatrix} \in \mathbb{R}^{m \times n}; \quad \hat{\mathbf{X}} = \begin{bmatrix} (\hat{\mathbf{x}}_1)^T \\ (\hat{\mathbf{x}}_2)^T \\ \vdots \\ (\hat{\mathbf{x}}_{\hat{m}})^T \end{bmatrix} \in \mathbb{R}^{\hat{m} \times n},$$

and

$$\begin{aligned} \mathbf{h} &= \begin{bmatrix} h(\mathbf{x}_1) \\ h(\mathbf{x}_2) \\ \vdots \\ h(\mathbf{x}_m) \end{bmatrix}, \quad \boldsymbol{\nu} = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_m \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m; \\ \hat{\mathbf{h}} &= \begin{bmatrix} h(\hat{\mathbf{x}}_1) \\ h(\hat{\mathbf{x}}_2) \\ \vdots \\ h(\hat{\mathbf{x}}_{\hat{m}}) \end{bmatrix}, \quad \hat{\boldsymbol{\nu}} = \begin{bmatrix} \hat{\nu}_1 \\ \hat{\nu}_2 \\ \vdots \\ \hat{\nu}_{\hat{m}} \end{bmatrix}, \quad \hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_{\hat{m}} \end{bmatrix} \in \mathbb{R}^{\hat{m}}. \end{aligned}$$

Given the training data S , the prior distribution $h(\cdot)$ and the testing inputs $\hat{\mathbf{X}}$, the use of standard tools of Bayesian statistics such as the Bayes' rule, marginalization and conditioning allow the computation of the posterior predictive distribution over the testing outputs \hat{y} [25].

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

$$\begin{bmatrix} \mathbf{h} \\ \hat{\mathbf{h}} \end{bmatrix} \Big| \mathbf{X}, \hat{\mathbf{X}} \sim \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) & \mathbf{K}(\mathbf{X}, \hat{\mathbf{X}}) \\ \mathbf{K}(\hat{\mathbf{X}}, \mathbf{X}) & \mathbf{K}(\hat{\mathbf{X}}, \hat{\mathbf{X}}) \end{bmatrix} \right), \quad (3.27)$$

where $\mathbf{K}(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{m \times m}$, $\mathbf{K}(\mathbf{X}, \hat{\mathbf{X}}) \in \mathbb{R}^{m \times \hat{m}}$, $\mathbf{K}(\hat{\mathbf{X}}, \mathbf{X}) \in \mathbb{R}^{\hat{m} \times m}$ and $\mathbf{K}(\hat{\mathbf{X}}, \hat{\mathbf{X}}) \in \mathbb{R}^{\hat{m} \times \hat{m}}$.

Considering the assumed iid noise model,

$$\begin{bmatrix} \boldsymbol{\nu} \\ \hat{\boldsymbol{\nu}} \end{bmatrix} \sim \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} \sigma^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma^2 \mathbf{I} \end{bmatrix} \right), \quad (3.28)$$

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

$$\begin{bmatrix} \mathbf{y} \\ \hat{\mathbf{y}} \end{bmatrix} \Big| \mathbf{X}, \hat{\mathbf{X}} = \begin{bmatrix} \mathbf{h} \\ \hat{\mathbf{h}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\nu} \\ \hat{\boldsymbol{\nu}} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}^{[1]}, \boldsymbol{\Sigma}^{[1]}), \quad (3.29)$$

where

$$\begin{aligned} \boldsymbol{\mu}^{[1]} &= \mathbf{0}, \\ \boldsymbol{\Sigma}^{[1]} &= \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I} & \mathbf{K}(\mathbf{X}, \hat{\mathbf{X}}) \\ \mathbf{K}(\hat{\mathbf{X}}, \mathbf{X}) & \mathbf{K}(\hat{\mathbf{X}}, \hat{\mathbf{X}}) + \sigma^2 \mathbf{I} \end{bmatrix}. \end{aligned}$$

Deriving the conditional distribution of $\hat{\mathbf{y}}$ results in the predictive equations of GPR:

$$\hat{\mathbf{y}} | \mathbf{y}, \mathbf{X}, \hat{\mathbf{X}} \sim \mathcal{N}(\boldsymbol{\mu}^{[2]}, \boldsymbol{\Sigma}^{[2]}), \quad (3.30)$$

where

$$\begin{aligned} \boldsymbol{\mu}^{[2]} &= \mathbf{K}(\hat{\mathbf{X}}, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}]^{-1} \mathbf{y}, \\ \boldsymbol{\Sigma}^{[2]} &= \mathbf{K}(\hat{\mathbf{X}}, \hat{\mathbf{X}}) + \sigma^2 \mathbf{I} - \mathbf{K}(\hat{\mathbf{X}}, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}]^{-1} \mathbf{K}(\mathbf{X}, \hat{\mathbf{X}}). \end{aligned}$$

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

3.3.4 Covariance Functions and its Hyperparameters

In the previous section, it was assumed that the covariance function $k(\cdot, \cdot)$ is known, which is not usually the case. In fact, the power of the Gaussian process to express a rich distribution on functions rests solely on the shoulders of the covariance function [61], 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 [24].

A wide variety of families of covariance functions exists, including squared exponential, polynomial, etc. See [25] 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 [25].

The covariance function must be positive semi-definite, given that it represents the covariance matrix of a multivariate Gaussian distribution [24]. It is possible to build composite covariance functions by adding simpler covariance functions, weighted by a positive hyperparameter, or by multiplying them, as adding and multiplying positive definite matrices results in a positive definite matrix [24].

One of the most commonly used covariance function in GPR is the squared exponential kernel given by (3.31), which reflects the prior assumption that the latent function to be learned is smooth [62]

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \cdot \exp\left(-\frac{(\mathbf{x} - \mathbf{x}')}{2\theta^2}\right). \quad (3.31)$$

In a nutshell, the hyperparameter σ controls the overall variance of the kernel function and the hyperparameter θ controls the distance from which two points will be uncorrelated, both of them presented in (3.31). These free parameters allow a flexible customization of the problem at hand [62], and maybe selected by inspection or automatically tuned by ML using the training data set. Also, the kernel is isotropic and stationary.

The covariance function in GPR plays the same roll as the kernel function in other approaches such as SVM and KRR [63]. Typically, these kernel methods use cross-validation techniques to adjust its hyperparameters [24], which are highly computational demanding and essentially consists of splitting the training set into k disjoint sets and evaluate the probability of the hyperparameters [25].

On the other hand, GPR can infer the hyperparameters from samples of the training set using the Bayesian framework [24]. The marginal likelihood of the hyperparameters of the kernel given the training data set can be defined as:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X})p(\mathbf{f}|\mathbf{X})d\mathbf{f}. \quad (3.32)$$

Recalling that \mathbf{X} is dependent of the hyperparameter's set, [64] proposes to maximize the

marginal likelihood in (3.32) in order to obtain the optimal setting of the hyperparameters. Although setting the hyperparameters by maximum likelihood is not a purely Bayesian solution, it is fairly standard in the community and it allows using Bayesian solutions in time sensitive applications [24]. More detailed information regarding practical considerations about this topic will be presented in Subsection 6.3.1 and can be found in [65].

Chapter 4

BUSINESS INTELLIGENCE SYSTEMS AND DATA

The BI systems and data used in this work were developed in partnership with MP, a Brazilian federal Ministry whose legal attributions [66] are summarized in Table 4.1.

Table 4.1: Legal attributions of MP

#	Description
I	Participate in the elaboration of the national strategic planning
II	Evaluate the economic and social impact of federal programs and politics
III	Conduct studies and research to follow the social and economic conjuncture and manage the cartographic national system
IV	Elaborate and evaluate the laws initiated by the executive power related to art. 165 of the Brazilian Federal Constitution
V	Conduct viability studies on new sources of revenue for the government planning
VI	Coordinate the partnerships with the private sector
VII	Formulate guidelines, coordinate negotiations and evaluate foreign finance of public projects with multilateral organizations and governmental agencies
VIII	Manage and coordinate the systems of federal budget, planning, civilian employees, patrimony, information technology and general services, as well as the government actions for public administration improvement
IX	Define and coordinate the criteria of corporate governance of federal companies
X	Manage the federal patrimony

It can be noticed that, among the legal attributions of MP listed in Table 4.1, the development

and maintenance of key information systems in the areas of civilian federal staff and patrimony management.

SIAPE is a national system that manages the monthly payroll of Brazilian federal employees [67], and SPU is responsible for managing the Brazilian federal patrimony. In Section 4.1, the BI system fed by the SIAPE database is explained, and in Section 4.2 the BI system maintained by SPU is shown.

4.1 PAYROLLS OF FEDERAL EMPLOYEES

SIAPE includes information of approximately two and half million workers among active, retired and pensioners; 14GB of raw data are generated each month, with 212 fields of personal, functional and financial data [16]. In the 2012 fiscal year, the size of the database of SIAPE ended up with more than 27 million rows for the public workers table and about 200 million rows in the financial data table [16]. An example of data contained in SIAPE is shown in Fig. 4.1.

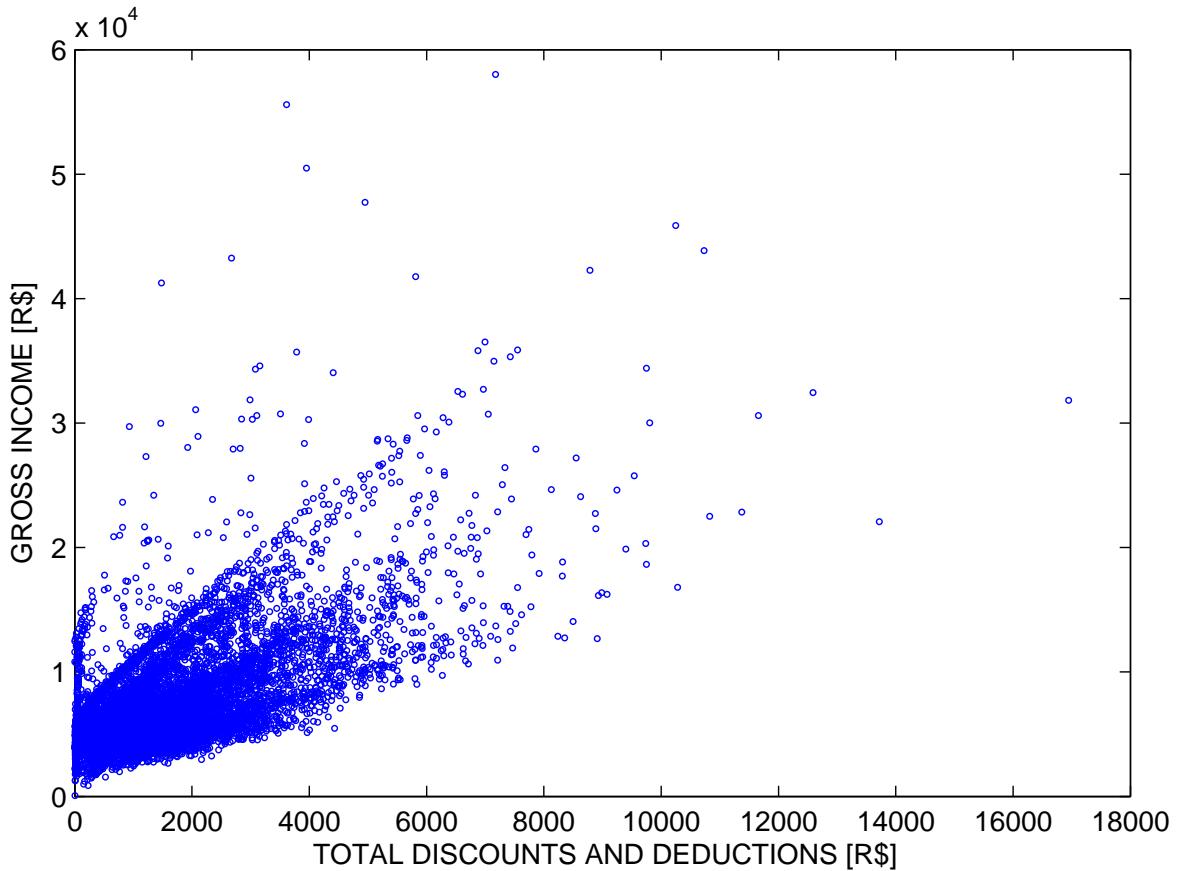


Figure 4.1: Scatter plot of 10,000 samples of payroll data, with gross income in one dimension (ordinate) and total discounts and deductions in the other dimension (abscissa). Both dimensions are plotted in *Reais*, the Brazilian currency.

Furthermore, there is a massive legal basis from which the payroll of the Brazilian federal staff are generated. The Federal Constitution of Brazil, laws, decrees and executive orders created more than 2,200 different rubrics [1*], the basic element of the payroll, consisting of a positive or negative value according to the characteristics of the position of the employee in the public administration organization.

According to the Brazilian legislation, CGAUD is the responsible department for auditing the rubrics of every payroll aimed at fraud detection such as incompatibility of benefits, inconsistencies and irregularities. Before the initial BI solution proposed in [14], CGAUD performed the audit process in a manual fashion.

After the implementation of the BI System proposed in [14], with several improvements proposed in [15], [16] and [1*], the current state-of-the-art BI system for auditing SIAPE is based on an ontology indexation process via concept maps in order to detect irregularities on the payrolls, big data technologies such as Hadoop and Hbase for increasing the performance of the processing stage and a reimbursement tracking system for monitoring the payroll of federal employees who have to reimburse the Brazilian Treasury. Fig 4.2 shows the architecture of the current BI system of CGAUD. Please refer to [1*] for further details on the existing BI architecture.

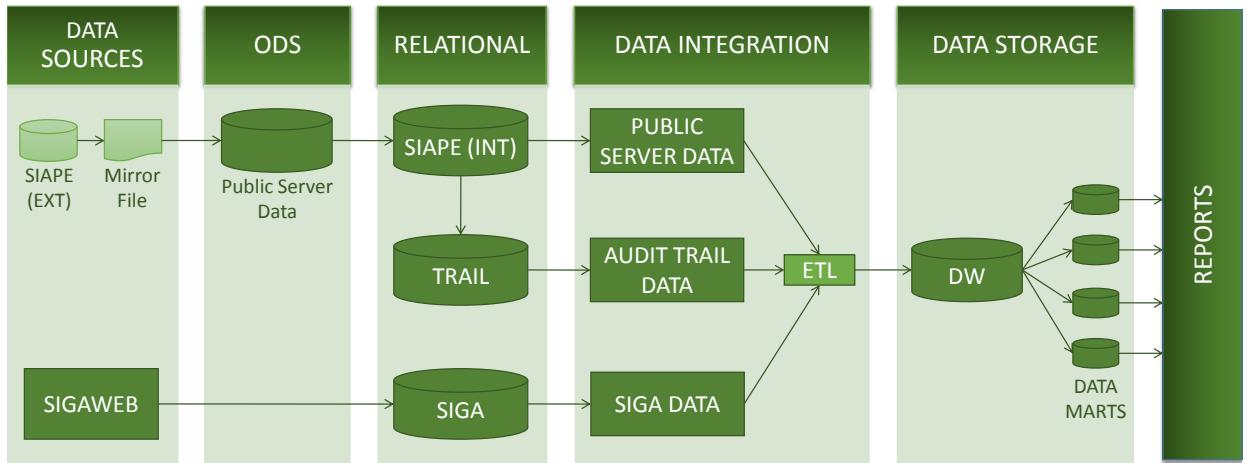


Figure 4.2: Architecture of the current state-of-the-art BI system of CGAUD

Despite the fact that the audit process is made before the payroll is actually paid to the employee, the existing audit process is fully based on audit trails, *i.e.* a deterministic analysis of the complete data structure where the information is presumably encoded in the *hypothesis*. Ontological audit trails mapping summarizes a set of hypothetic rules based on Brazilian legislation, such as incompatibility of rubrics, and the real world data validates or refutes those hypothesis. Fig. 4.3 shows an example of audit trail concept map. Please refer to [14] for more details on the construction of audit trails.

Hence, the current audit process has no predictive component and no pre-processing of the

huge amount of monthly incoming data, having to check every row of a specific rubric in order to detect any irregularity.

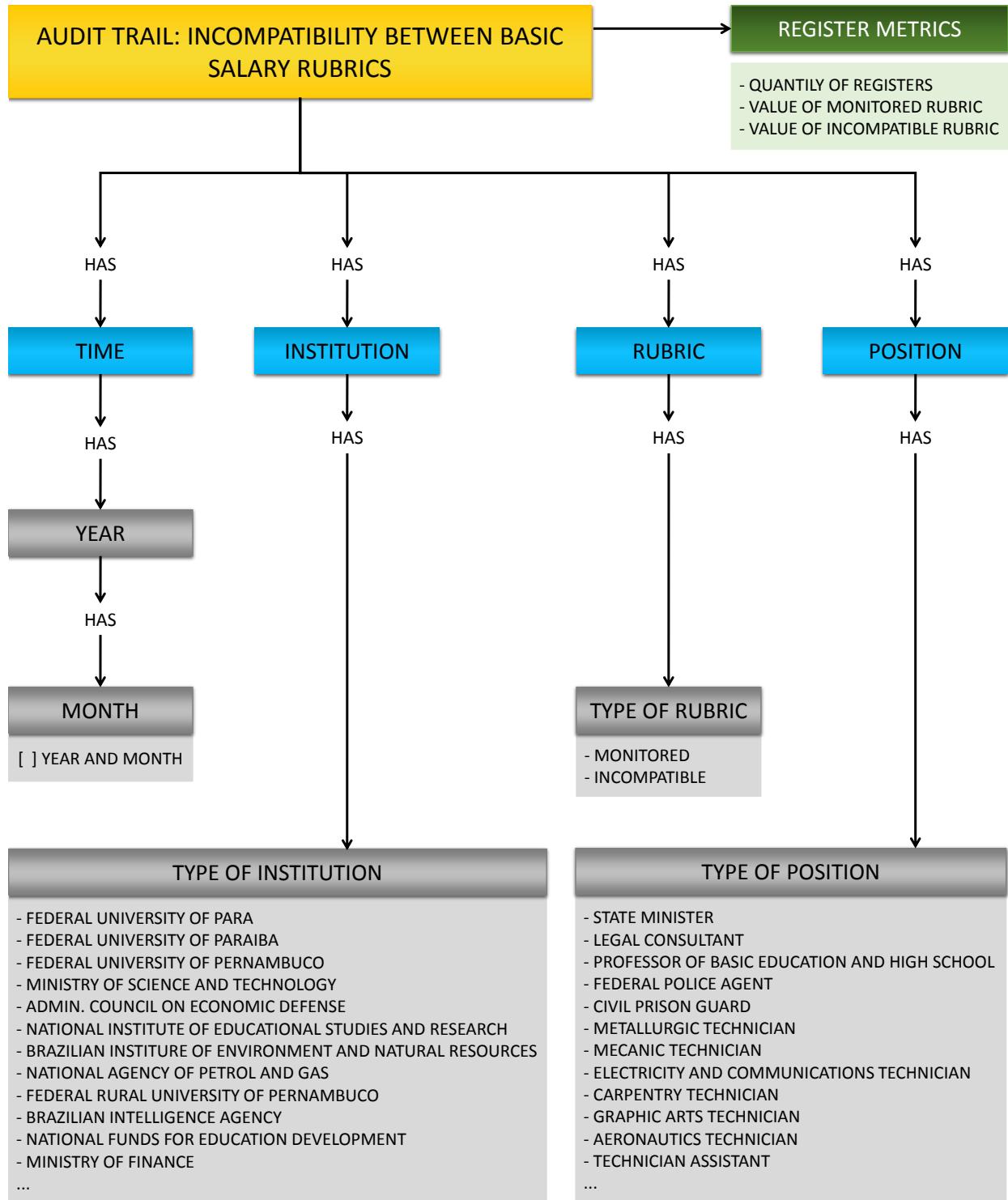


Figure 4.3: Example of a concept map for an audit trail. Adapted from [14].

4.2 FEDERAL TAX COLLECTION

SPU is legally responsible for managing, supervising and grant permission to use federal real estate properties in Brazil. The monthly revenue managed by this branch of the federal government comes mainly from taxes and other associated fees collected by its Department of Patrimony Revenue Management¹ [68].

Similarly to the data regarding the payrolls of the Brazilian federal staff, the tax collected by SPU is based on a massive amount of federal legislation spread out among the Constitution of Brazil, laws, decrees and executive orders. A BI system designed for SPU was first implemented by [69], where a predictive analytics module based on artificial neural network forecasted the monthly amount of tax to be collected.

Fig 4.2 shows the architecture of the current BI system of SPU. Please refer to [69] for more details on the existing BI architecture.

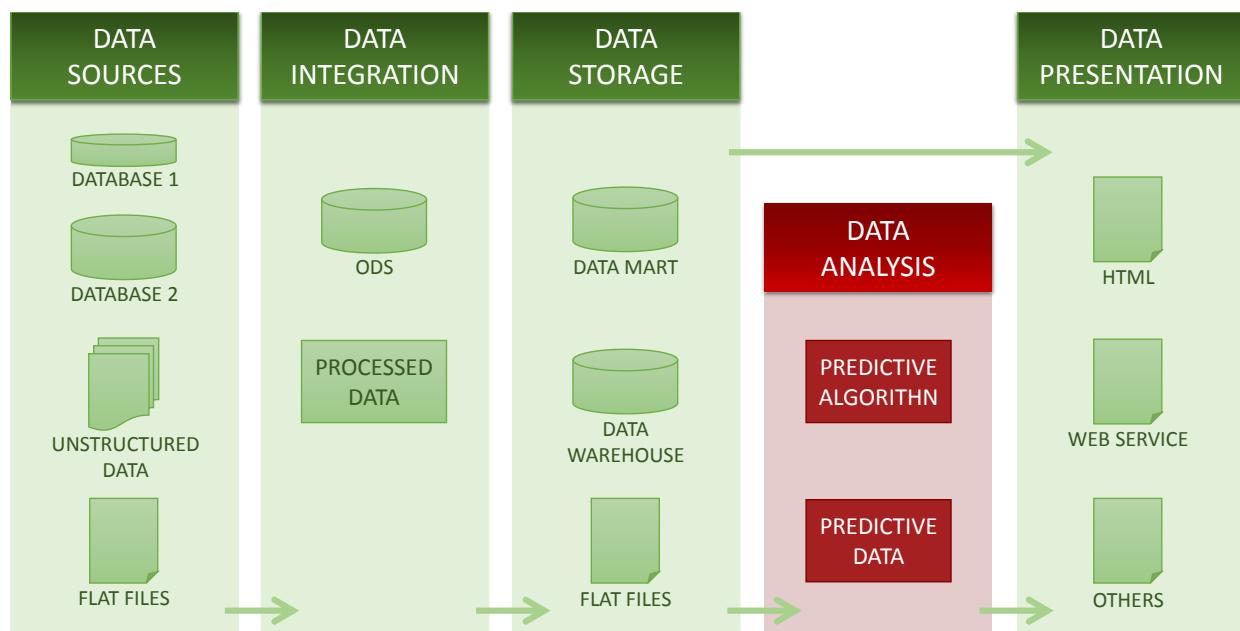


Figure 4.4: Architecture of the current state-of-the-art BI system of SPU

Since the BI system of SPU already possesses a predictive analytics model embedded in its current architecture, from this point forward our focus will rely only on techniques inside that particular module, in order to obtain improved prediction results.

In addition, the same input data of SPU used in [45] will be used in this work for comparison purposes. The data regards the monthly tax collection of SPU, ranging from years 2005 to 2010. The amount collected, expressed in *reais* (R\$), is treated as a random variable indexed by the m^{th} month, where m ranges from 1 to 72. Thus, $m = 1, \dots, 12$ is related to the first year's collection

¹In portuguese, *Departamento de Gestão de Receitas Patrimoniais*.

(2005); $m = 13, \dots, 24$ is related to the second year's collection (2006), and so forth.

In alignment with the strategy used in [45], also for comparison purposes, it was used only the first 60 months of the data (ranging from 2005 to 2009) to train the predictive algorithm. The data regarding the year 2010 was exclusively used to evaluate the performance of the proposed predictor by error measurement. Therefore, the first five years of data will be referred as the training data set, and the sixth year of data will be referred as the target data set. Figure 4.5 shows a bar plot of the the data model used in this work.

The gray scale bars, representing the years between 2005 and 2009, were chosen as the training set, and the red bars, representing the year 2010, were chosen as the target set.

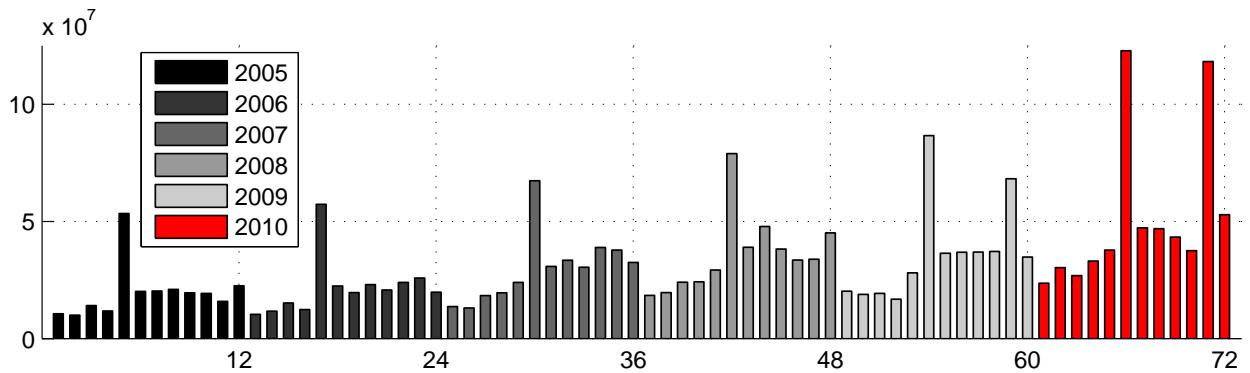


Figure 4.5: Monthly tax collected by SPU, in *reais* (R\$), indexed by the m^{th} month.

Chapter 5

STATISTICAL AUDIT

This Chapter is dedicated to present a new approach to the existing BI system in CGAUD, which audits the monthly payroll of Brazilian federal employees. In Section 5.1, we put into perspective the arguments for incorporating a statistical analysis approach into a complete deterministic BI system in its state-of-the-art. The statistical technique proposed is based on GMM, which models the probabilistic behavior of payrolls and act as a filter to improve the computational efficiency of the system as well as its responsiveness. In Section 5.2, we go through the algorithm optimization process and overall experimental results.

5.1 STATYSTICAL ANALYSIS ON A DETERMINISTIC BI SYSTEM

Considering the amount of data to be analyzed by the current BI system of CGAUD, and the rising trend of the number of audit trails which also causes the processing requirements of the system to rise proportionally, this work proposes a complementary statistical approach to the system based on GMM described in Section 3.2. Focused on the *data*, the system aims to model a pdf for each category of the Brazilian federal staff with common payroll characteristics (professors, police officers, judges, etc.), hence defining a regular behavior for the payrolls of each category.

After the definition of a standard payroll behavior for a given category of employees, the next goal is to classify each individual payroll as regular or possibly inconsistent. In other words, the principle of the proposed system can be stated as the higher the probability of a random payroll, the less likely that payroll is to be inconsistent. One way to validate this hypothesis is to use the current BI system as a qualitative measure, where the removal of the most probable payrolls from the audit trails should not drastically impact the result of the original audit trails.

The data used in this work consists of 101,400 payroll entries of the federal professors category,

since this is one of the categories with more employees of the Brazilian federal staff [70]. The chosen month of application of the proposed technique was June/03, since federal employees in Brazil are monthly payed and June is one month of the year with a high variance among all the other months given that the first part of the 13rd salary is payed on that month for a substantial part of the governmental staff [71].

Each payroll is arranged in a two dimensional structure, where instead of dealing with more than 2,200 different rubrics, the whole information of the rubrics is condensed into gross income in one dimension and total deductions in the other dimension. Generalizing positive rubrics (incomes) in one dimension and negative rubrics (deductions) in another dimension resulted in the scatter diagram of the data sets shown in Fig. 5.1.

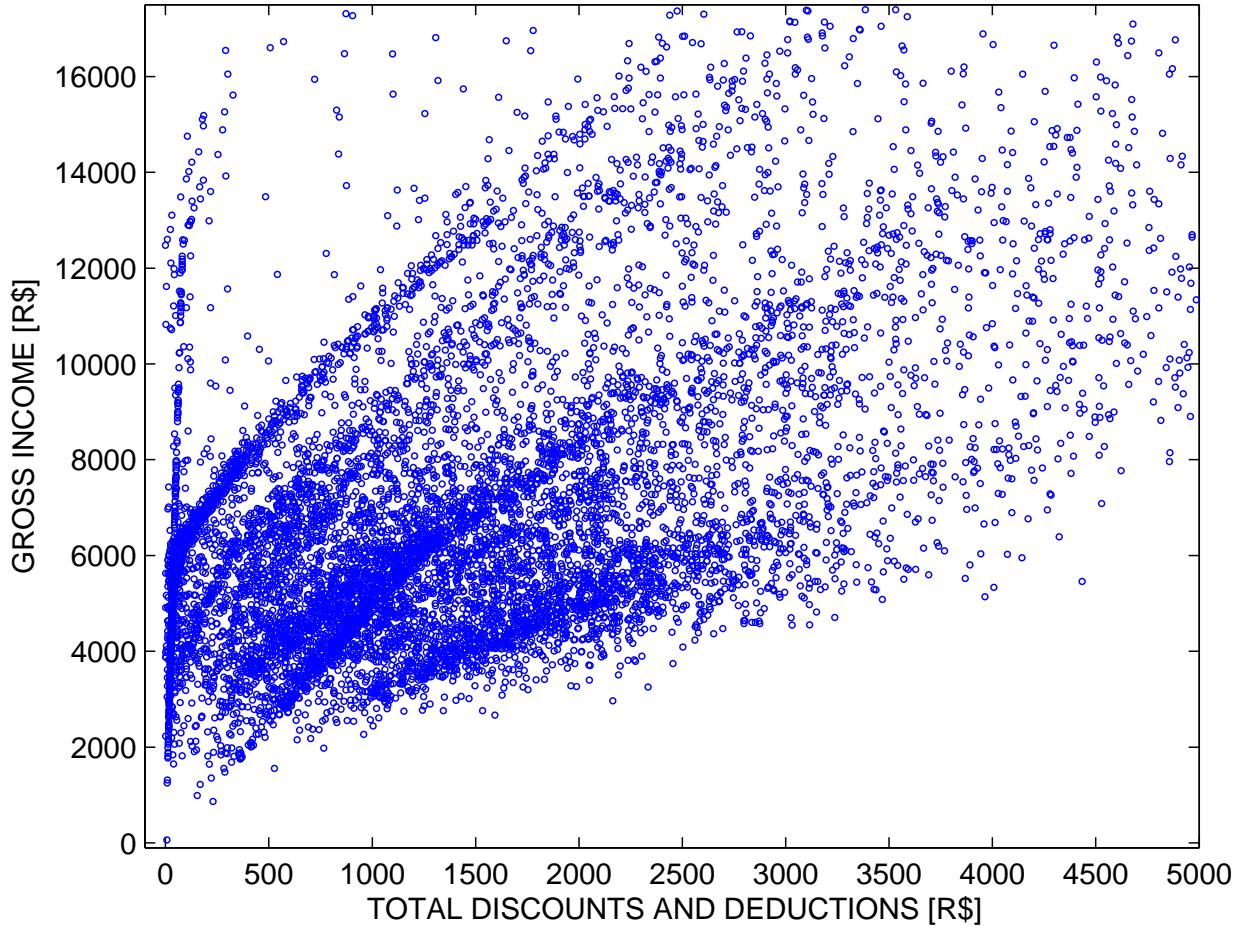


Figure 5.1: Scatter plot of 10,000 samples of payroll data, showed in Fig. 4.1, with a zoom around the origin for a better visualization of the correlation profile.

Fig. 5.1 shows a 10,000 points sample of the data. As expected, it can be noted in the cropped scatter plot a positive correlation between the amount of gross income and the total deductions and discounts from the payroll.

5.1.1 Statistical Audit Module

The proposed statistical audit module, based on a generative GMM pdf, was incorporated into the original BI system described in Section 4.1 between SIAPE and TRAIL databases, inside the relational stage (see Fig. 4.2). A block diagram in Fig. 5.2 shows the new audit module in the BI architecture.

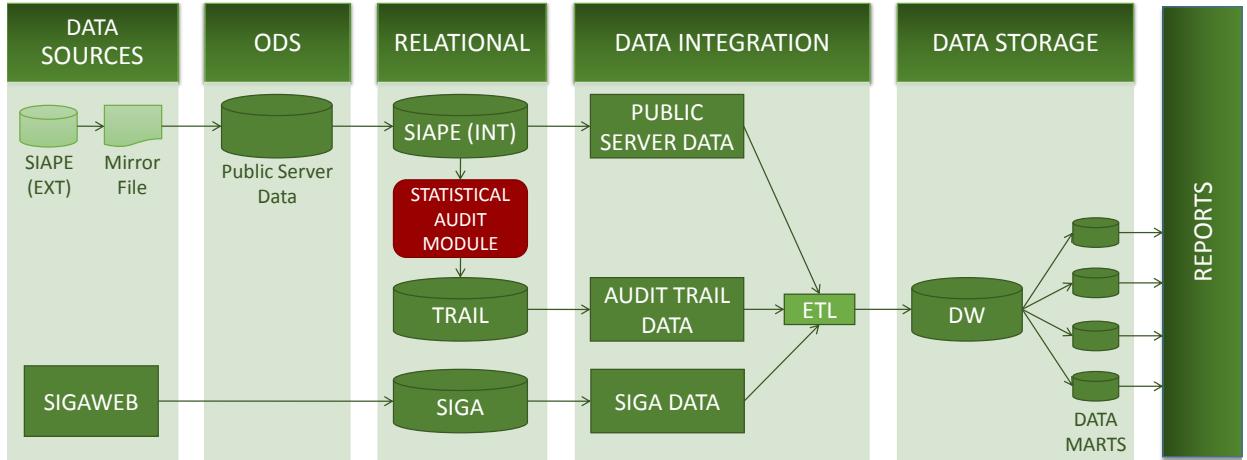


Figure 5.2: Block architecture of the proposed statistical audit module solution in the original BI architecture shown in Fig. 4.2.

The statistical audit module acts as a filter for the audit trail database, removing high probability payrolls contained in the SIAPE database and feeding the TRAIL database only with payrolls that are most dissimilar from the normal payroll behavior modeled by a GMM pdf.

Given that the TRAIL database is generated through computationally cost relational statements between each audit trail and the hole SIAPE database, and the GMM pdf inside the statistical audit module is generated by a one time optimization process via EM algorithm, a positive trade off between cost and effort could be established. Although any probabilistic method has a certain degree of uncertainty associated with it, which in this specific case translates into losing some false negatives (*i.e.* high probability payrolls that have some kind of irregularity) in the statistical audit module, the gain obtained in terms of computational efficiency and velocity of execution enables the hole BI system to audit a more significant portion of the overall payroll of Brazilian public employees.

5.1.2 GMM for Statistical Auditing

The approach chosen to model the data is a finite GMM, which gives a complete statistical description of the latent underlying system that generated the data. GMM is a parametric model, completely defined by its mixing weights, mean vectors and covariance matrices. Therefore, in the

context of this work, GMM can be seen as a generative model capable of defining the probability of a random payroll to occur.

In order to obtain the pdf of our GMM, *i.e.* learn the parameters $\Theta = \{\alpha_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$, we apply the EM algorithm for a GMM. The derivation of closed form solutions for the equations presented in Subsection 3.2.2 for a mixture of multivariate Gaussians require techniques which are beyond the scope of this project. For a detailed derivation of those, please refer to [53].

Using the framework previously defined in subsection 3.2.2, a closed form solution for the auxiliary function \mathcal{Q} , the conditional expectation of the complete data, can be written as:

$$\mathcal{Q}(\Theta|\Theta^p) = \sum_{i=1}^N \sum_{k=1}^K \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k)}{p(\mathbf{x}_i|\Theta)} \log(\alpha_k) + \sum_{i=1}^N \sum_{k=1}^K \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k)}{p(\mathbf{x}_i|\Theta)} \log(p_k(\mathbf{x}_i|\theta_k)) \quad (5.1)$$

The expression for \mathcal{Q} derived in (5.1) appears in the E-Step of the EM algorithm and may be maximized for a particular pdf p_k .

Assuming p_k being a multivariate Gaussian distribution in the form of (3.3), and that

$$\Theta^p = (\alpha_1^p, \dots, \alpha_K^p, \theta_1^p, \dots, \theta_K^p) \in \Omega$$

is our prior set of parameters, the goal is to implement the M-Step of the EM algorithm to obtain updated maximizers denoted by $\Theta^* = (\alpha_1^*, \dots, \alpha_K^*, \theta_1^*, \dots, \theta_K^*) \in \Omega$.

This can be achieved by maximizing \mathcal{Q} with respect to α_k and $\theta_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, leading to the *updated parameter equations* of the M-Step f the EM algorithm for α_k^* , $\boldsymbol{\mu}_k^*$ and $\boldsymbol{\Sigma}_k^*$ as being, respectively:

$$\alpha_k^* = \frac{1}{K} \sum_{i=1}^N \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k^p)}{p(\mathbf{x}_i|\Theta^p)}; \quad (5.2)$$

$$\boldsymbol{\mu}_k^* = \frac{\sum_{i=1}^N \mathbf{x}_i \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k^p)}{p(\mathbf{x}_i|\Theta^p)}}{\sum_{i=1}^N \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k^p)}{p(\mathbf{x}_i|\Theta^p)}}; \quad (5.3)$$

$$\boldsymbol{\Sigma}_k^* = \frac{\sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu}_k^*)(\mathbf{x}_i - \boldsymbol{\mu}_k^*)^T \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k^p)}{p(\mathbf{x}_i|\Theta^p)}}{\sum_{i=1}^N \frac{\alpha_k^p p_k(\mathbf{x}_i|\theta_k^p)}{p(\mathbf{x}_i|\Theta^p)}}. \quad (5.4)$$

Again, please refer to [53] for a detailed derivation of (5.2), (5.3) and (5.4).

Initialization and Convergence Issues for EM

A crucial point of the EM algorithm is the initial set of parameters Θ^p of the model. A standard way to obtain Θ^p is to choose random α_k values uniformly from $[0, 1]$ and estimate the individual source parameters with a M-Step [72].

In order to deal with the effects of random initialization and a possible convergence to a local maximum, all estimations can be repeated a number of times and the solution with the highest likelihood is selected [72].

5.2 OPTIMIZATION AND EXPERIMENTAL RESULTS

One key aspect of modeling the payroll data set as a bidimensional GMM is the number of source components K in (3.1). Whereas the number of sources can be linked directly to the number of clusters of a classification algorithm, in many cases extending the finite mixture model such as $K \rightarrow \infty$ produces densities whose generalization is highly competitive with other commonly used methods [73].

Recalling that our classification proposal is not based on the number of classes, or source components, but instead is based exclusively on the pdf generated by the mixture model, where payrolls that have a probability level above a certain threshold are classified as less likely to be irregular. In this particular case, not limiting the number of classes *a priori* removes an extra parameter of the stochastic model to be estimated.

Hence, whereas the number of clusters increases with the number of sources in a classical mixture model, the underlying pdf of the mixture tends to stabilize as shown in Fig. 5.3. Due to computational constraints, it is not possible to extend the number of classes to infinity, but in our particular case the pdf showed to be stable with a number of sources $K \geq 30$. It is important to state that, in Fig. 5.3, our main interest lies at the areas inside the red and brown contours. These are the areas with highest probability values and thus these are the areas we look for stability.

Another interesting feature noticed in Fig. 5.3 is the decay rate of the log-likelihood function. As the number of sources increases, the resulting likelihood function tends to increase as well [72]. Taking that assumption to the limit, when the number of sources is equal to the number of observed data points, the likelihood of each point being generated by its own data source is equal to 1. Nevertheless, it can be observed in our system that, as the number of source components increase, the rate of decay of the log-likelihood function decreases, leading to the conclusion that adding more source components to the mixture does not add much more significant information about the system.

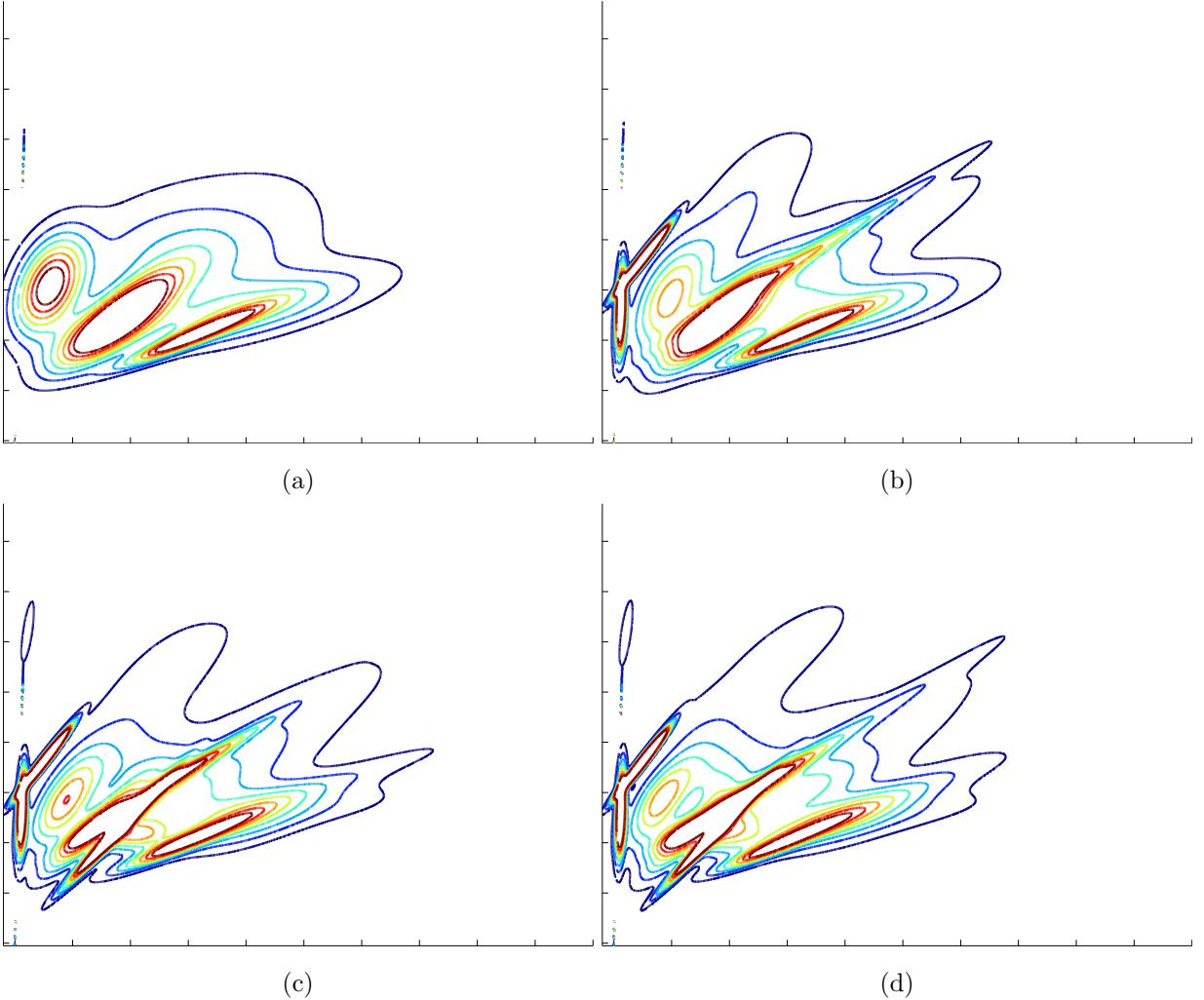


Figure 5.3: Contour plot of the estimated pdf of the dataset presented in Fig. 5.1 with (a) 8 sources (log-likelihood: -174317); (b) 16 sources (log-likelihood: -173282); (c) 24 sources (log-likelihood: -173019) and (d) 32 sources (log-likelihood: -172937). The axis in all subfigures are the same as in Fig.5.1.

With the number of sources $K = 30$ defined in (3.1), the EM algorithm was applied to the set of data of SIAPE regarding the federal professors staff, originally a data set with 101,400 payroll entries.

In order to avoid a possible convergence to a local maximum, all estimations are made 15 times [72] with different random set of initial parameters $\Theta^p = (\alpha_k^p, \mu_k^p, \Sigma_k^p) \in \Omega$ in (5.2), (5.3) and (5.4). The initial set of parameters are obtained by randomly choosing K observations from $\mathcal{X} = (\mathbf{x}_1, \dots, \mathbf{x}_L)$ in (3.1) as initial component means. The mixing weights are uniform. The covariance matrices for all components are diagonal, where the element j on the diagonal is the variance of $\mathcal{X}(:, j)$.

The convergence criteria adopted for the EM algorithm is the termination tolerance on the

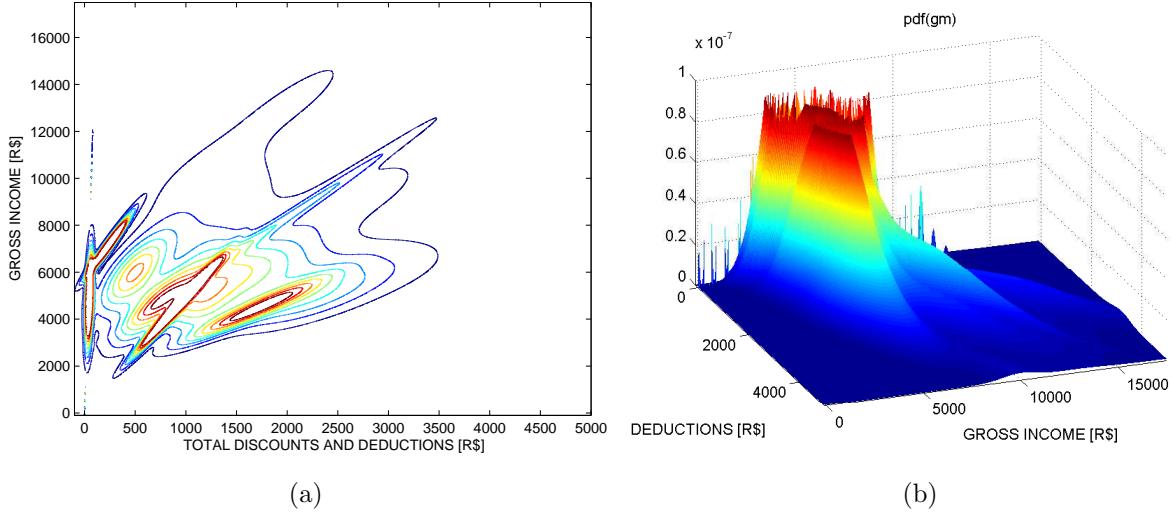


Figure 5.4: (a) Contour plot and (b) surface plot of the resulting pdf of the proposed GMM.

log-likelihood function in (3.8), where the algorithm stops when the new guesses of parameters Θ^* produce only minimal increments of the log-likelihood function given in (3.8), *e.g.* increments smaller than 10^{-5} . Thus, the convergence criteria is met when only negligible improvements of the solution can be achieved by performing new iterations.

The resulting pdf of the GMM, optimized with the EM algorithm according to the settings previously described, is shown in Fig. 5.4.

It can be seen in Fig. 5.4, through the equi-probability contour lines, that the learned GMM pdf possess high values of probability where the set of observed data samples are more dense. Since the observed data set regards the proportion of gross income and total discounts and deductions from payrolls, the hypothesis we are seeking to test is that employees which have a proportion of gross income *versus* discounts that are far away from the normal behavior of the GMM pdf are more likely to have irregularities in their payroll.

To confirm that hypothesis, a pre-processing stage was created on the current BI system model described in Section 4.1. As previously discussed, this stage consists of a statistical filter, where high probability payrolls according to the GMM pdf are presumed regular and are not processed by the deterministic BI system based on audit trails. Please refer to Fig. 5.2.

Tables 5.1 and 5.2 show the statistical filter results by comparison with unfiltered data. The audit trails chosen to populate the tables are the ones that contain the most significant number of occurrences.

The information in the Table 5.1 is organized as follows: In the first line of the table, it can be seen that if we filter, *i.e* remove, the 5% most probable payrolls off our observed data set, this implicates in a loss of 1.26% of the total occurrences of the current audit process in trail

Table 5.1: Fraud occurrences detected by audit trails, grouped by trail ID # and divided according to their probability of occurrence.

Probability of Occurrence	% Reference	Audit Trail #13		Audit Trail #18		Audit Trail #31	
		Abs	%	Abs	%	Abs	%
0 ~ 5 %	5%	329	1.26%	226	0.77 %	28	0.72 %
5 ~ 10 %	5%	847	3.25%	410	1.40 %	65	1.67 %
10 ~ 20 %	10%	1806	6.93%	1152	3.93 %	177	4.54 %
20 ~ 40 %	20%	4704	18.05%	5308	18.09 %	857	21.99 %
40 ~ 60 %	20%	4074	15.63%	7147	24.36 %	1145	29.38 %
60 ~ 80 %	20%	5785	22.20%	7245	24.69 %	983	25.22 %
80 ~ 100 %	20%	8512	32.67%	7853	26.76 %	642	16.47 %
Total	100%	26057	100%	29341	100 %	3897	100 %

#13 (false negatives). Analogously, filtering 20% of the most probable payrolls off our observed data set causes a loss in the trail #13 of 11.44% in false negatives, given that we lose 1.26% of the 5% most probable payrolls plus 3.25% of payrolls with probabilities between 5% and 10% and 6.93% of payrolls with probabilities between 10% and 20%. The ABS columns contain the number of occurrences per trail #; the probability of occurrence column contains the intervals which a random payroll can be classified; the % reference column contains the range of the intervals in the column probability of occurrence.

Table 5.2: Total fraud occurrences detected by audit trails, divided according to their probability of occurrence.

Probability of Occurrence	% Reference	% Average	% Cumulative
0 ~ 5 %	5%	0.92 %	0.92 %
5 ~ 10 %	5%	2.11 %	3.02 %
10 ~ 20 %	10%	5.13 %	8.16 %
20 ~ 40 %	20%	19.38 %	27.53 %
40 ~ 60 %	20%	23.12 %	50.66 %
60 ~ 80 %	20%	24.04 %	74.70 %
80 ~ 100 %	20%	25.30 %	100.00 %

The information in the Table 5.2 is organized as follows: In the first line of the table, it can be seen that if we filter, *i.e* remove, the 5% most probable payrolls off our observed data set, this implicates in an cumulative loss of 0.92% of the total occurrences of the current audit process

(false negatives). Analogously, filtering 20% of the most probable payrolls off our observed data set causes an cumulative loss in the current audit trails of 8.16%. The probability of occurrence column contains the intervals which a random payroll can be classified; the % reference column contains the range of the intervals in the column probability of occurrence; the % average column contains the % of the filtered observed payroll population at that level of probability.

Given that the GMM filter proposed in this work is unique for all the audit trails, so the output of the GMM filter feeds all the audit trails in the current BI system, the gain in terms of efficiency is considerable since, with the use of the filter, it is possible to reduce the processing requirements of the system by 20% with an average audit loss of about 8.16%. In other words, it can be stated that if we submit 80% of the less probable payrolls to the audit trails, we will be able to detect 91.84% of the irregularities.

In addition, the underlying rules that dictate the behavior of payrolls are highly related to federal Brazilian legislation. Taking that into consideration, the proposed GMM of a certain month of the year should not present severe changes if the legislation regarding public workers remains unchanged, reducing the need to recompute the GMM for every month.

Chapter 6

PREDICTIVE ANALYTICS

In this Chapter, we address the data analysis module in the SPU BI system highlighted in Fig. 4.4. The BI system maintained by SPU performs a predictive analytics on the monthly tax collection, among other tasks, and its current state-of-the-art algorithm uses artificial neural networks to find hidden patterns in the time series. In Section 6.1, we propose to model the data regarding the monthly tax collection with the use of GPR in an unidimensional fashion. Section 6.2, we perform a transformation in the original data set, taking advantage of the bidimensional structure of the data and enabling the use of an adapted multidimensional GPR model. In Section 6.3, we discuss the optimization of the hyperparameters of the GPR model, present the results of the predictive module and propose a classification stage based on the statistical description natively produced by GPR.

6.1 UNIDIMENSIONAL PREDICTOR MODEL

In practice, a Gaussian process can be fully defined by just its second moment, or covariance function, if the mean function can be set or assumed to be zero. The implications of that approach takes place in Subsection 6.1.1, where the data normalization and a unidimensional model for the mean and covariance functions are discussed. The prediction results using this unidimensional model is presented in Subsection 6.1.2.

6.1.1 Mean and Covariance Function Modeling

Considering the training SPU data set in Fig. 4.5, 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 in (3.24) can be set to zero and the focus of the GPR modeling 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 modeled by the following linear covariance function:

$$k_1(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'. \quad (6.1)$$

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

$$k_{2,1}(\mathbf{x}, \mathbf{x}') = \sigma_1^2 \exp\left(-\frac{2 \sin^2[\frac{\pi}{\theta_2}(\mathbf{x} - \mathbf{x}')] }{\theta_1^2}\right)$$

is modified by the squared exponential covariance function

$$k_{2,2}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{(\mathbf{x} - \mathbf{x}')}{2\theta_3^2}\right),$$

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

$$k_2(\mathbf{x}, \mathbf{x}') = k_{2,1} \cdot k_{2,2} = \sigma_1^2 \exp\left(-\frac{2 \sin^2[\frac{\pi}{\theta_2}(\mathbf{x} - \mathbf{x}')] }{\theta_1^2} - \frac{(\mathbf{x} - \mathbf{x}')}{2\theta_3^2}\right). \quad (6.2)$$

Finally, the sum of the characteristic components in (6.1) and (6.2) leads to the proposed noiseless covariance function:

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') = \sigma_1^2 \exp\left(-\frac{2 \sin^2[\frac{\pi}{\theta_2}(\mathbf{x} - \mathbf{x}')] }{\theta_1^2} - \frac{(\mathbf{x} - \mathbf{x}')}{2\theta_3^2}\right) + \mathbf{x}^T \mathbf{x}'. \quad (6.3)$$

In (6.3), the hyperparameter σ_1 gives the magnitude, or scaling factor, of the covariance function. The θ_1 and θ_3 give the relative length scale of periodic and squared exponential functions, respectively, and can be interpreted as a "forgetting factor". The smaller the values of $\theta_{1,3}$, the more uncorrelated two given observations x and x' are. The θ_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 θ_2 time indexes.

To complete the modeling profile, the measured noise is assumed to be additive white Gaussian with variance σ_n^2 , which leads to the final noisy covariance function:

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') + \sigma_n^2 \mathbf{I}. \\ k(\mathbf{x}, \mathbf{x}') &= \sigma_1^2 \exp\left(-\frac{2 \sin^2[\frac{\pi}{\theta_2}(\mathbf{x} - \mathbf{x}')] }{\theta_1^2} - \frac{(\mathbf{x} - \mathbf{x}')}{2\theta_3^2}\right) + \mathbf{x}^T \mathbf{x}' + \sigma_n^2 \mathbf{I}. \end{aligned} \quad (6.4)$$

As an example of the individual contributions of each component of the covariance function to the final prediction, Fig. 6.1 shows the decomposed product function $k_2(\mathbf{x}, \mathbf{x}')$ of (6.2) in terms of the periodic and the squared exponential components. The input observed data is the normalized SPU data set in Fig. 4.5.

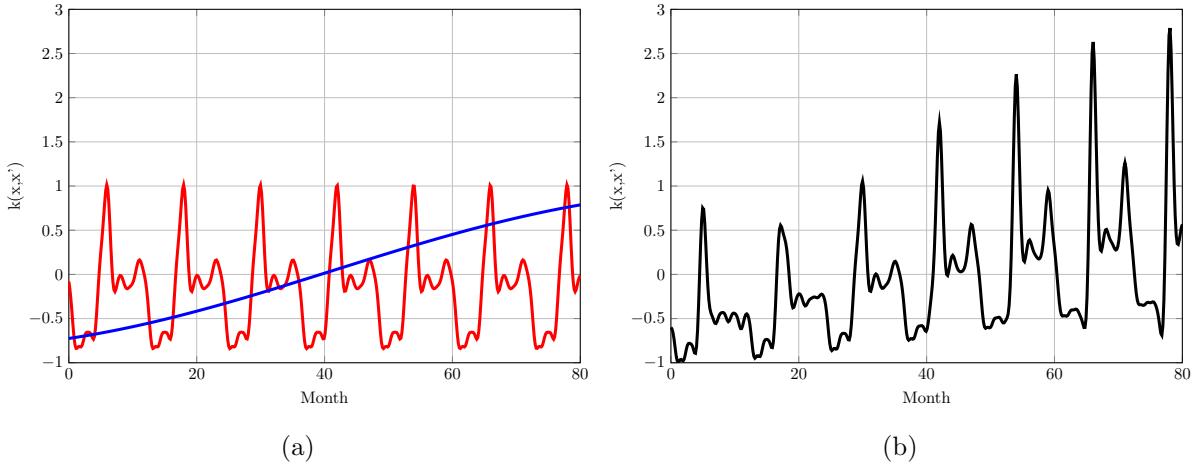


Figure 6.1: Normalized plot of the posterior inference of the Gaussian process, indexed by a continuous time interval $\mathcal{X} = [0, 80]$, obtained using the covariance function (a) $k_{2,1}(\mathbf{x}, \mathbf{x}')$ in red (the periodic component) and $k_{2,2}(\mathbf{x}, \mathbf{x}')$ in blue (the squared exponential component); (b) $k_2(\mathbf{x}, \mathbf{x}')$ in black (the product of both components).

The plots of Fig. 6.1 were obtained with the hyperparameters

$\sigma_1^2 = 1$; $\theta_1 = 0.3$; $\theta_2 = 12$; $\theta_3 = 60$ and $\sigma_n^2 = 0.1$.

The magnitude σ_1^2 was set to 1 not to distort the resulting function regarding the training set; the θ_1 was set to 0.3 month due to the poor month-to-month correlation that the data presents; the θ_2 was set to 12 months due the periodicity of the data; the θ_3 was set to 60 months to ensure all data points are taken into account in the final prediction results and, at least, the σ_n^2 was set to 0.1 to add some white Gaussian noise on the observation set. At this point, it is important to remember that the initial choice of hyperparameters have only taken into consideration the characteristics of the original data set. Later, on Subsection 6.3.1, we present a optimization method for tuning them.

6.1.2 Unidimensional Prediction Results

With the covariance function defined in (6.4) and a set of training points given by the first 60 months of the normalized SPU data of Fig. 4.5, it is possible to formulate a GPR with time as input.

The GPR's characteristic of returning a probability distribution over a function enables the evaluation of the uncertainty level of a given result. For each point of interest, the Gaussian process can provide the expected value and the variance of the random variable, as shown in Fig. 6.2.

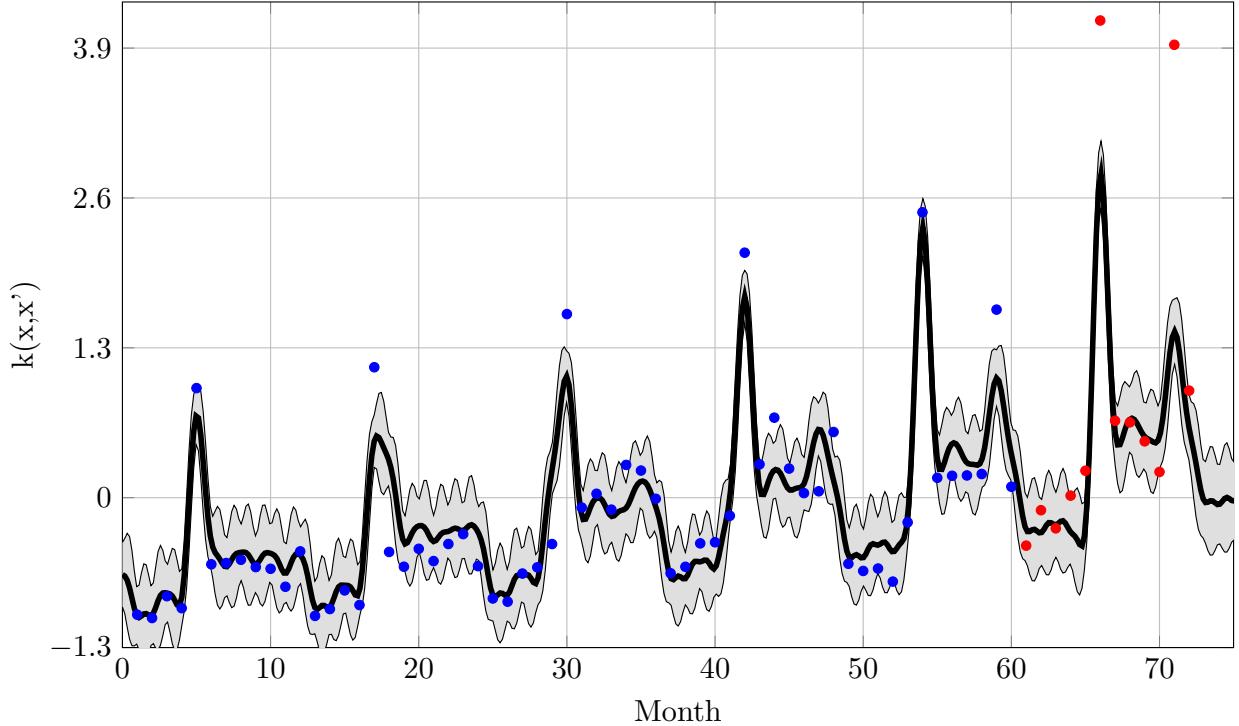


Figure 6.2: Prediction results from conditioning the posterior Gaussian jointly distribution at a continuous time interval $\mathcal{X} = [0, 75]$. The blue dots are the training data, the red dots are the target data, the black tick line is the expected value at a time index and the gray band represents the 95% confidence interval (two standard deviations above and below the expected value).

It is noticeable that, for the twelve month prediction using the proposed model, two predicted months fell off the confidence band that delimitates the 95% certainty interval - June and November. These two months have a high contribution on the overall prediction error on this initial approach.

6.2 BIDIMENSIONAL DATASET RESHAPE

In this section, we propose a pre-processing stage based on the cross-correlation profile of the original data set. This profile is used to separate highly correlated months into one dimension and poor correlated months into a different dimension, leading to a two dimensional structure. Subsection 6.2.1 shows an analysis of the time cross-correlation results and implications on the proposed model, and Subsection 6.2.2 shows the proposed reshaped data set.

6.2.1 Time Cross-Correlation

Although the uptrend and the periodic seasonal characteristics are prominent in our data set, some important features of the data are not visible at first sight. Considering that the covariance function used to define the GPR is based on a measure of distance, where closer pairs of observation points tend to have a strong correlation and distant pairs of points tend to have a weak correlation, a measure of month-to-month correlation in SPU data can reveal the accuracy of that approach.

The cross-correlation between two any infinite length sequences is given by [74]:

$$\mathbf{R}_{\mathbf{x}\mathbf{y}}(m) = \mathbb{E}[\mathbf{x}_n \mathbf{y}_{n-m}^*] \quad (6.5)$$

In practice, sequences \mathbf{x} and \mathbf{y} are likely to have a finite length, therefore the true cross correlation stated in (6.5) needs to be estimated since only partial information about the random process is available. Thus, the estimated cross-correlation, with no normalization, can be calculated by [74]:

$$\hat{\mathbf{R}}_{\mathbf{x}\mathbf{y}}(m) = \begin{cases} \sum_{n=0}^{N-m-1} \mathbf{x}_{n+m} \mathbf{y}_n^* & \text{if } m \geq 0 \\ \hat{\mathbf{R}}_{\mathbf{y}^*\mathbf{x}}(-m) & \text{if } m < 0 \end{cases} \quad (6.6)$$

Fig. 6.3 shows a plot of the absolute cross-correlation of the entire SPU data as sequence \mathbf{x}_n , and the last year's target data as sequence \mathbf{y}_n . The smaller sequence was zero-padded to give both sequences the same length. The resulting cross-correlation was also normalized to return 1.0 exactly where the lag m matches the last year's target data month-by-month.

The cross-correlation between the target data and the rest of the sequence exhibited a couple of interesting features about the data. First, it can be noted that the first two years are poorly correlated with the last year. Second, there are some clear peaks on the cross-correlation function where the lag m is a multiple of 12.

Some important conclusions arise from those features. First one is that there is not much information about the last year on the first two years of data, and the amount of information rises as it gets closer to the target. This complies with the distance based correlation function previously proposed.

Also, the peaks pattern shows that the month-to-month correlation is poor, since we only get high correlation values when comparing January of 2010 with January of 2009, 2008, 2007; February of 2010 with February of 2009, 2008, 2007 and so forth. Although some secondary order

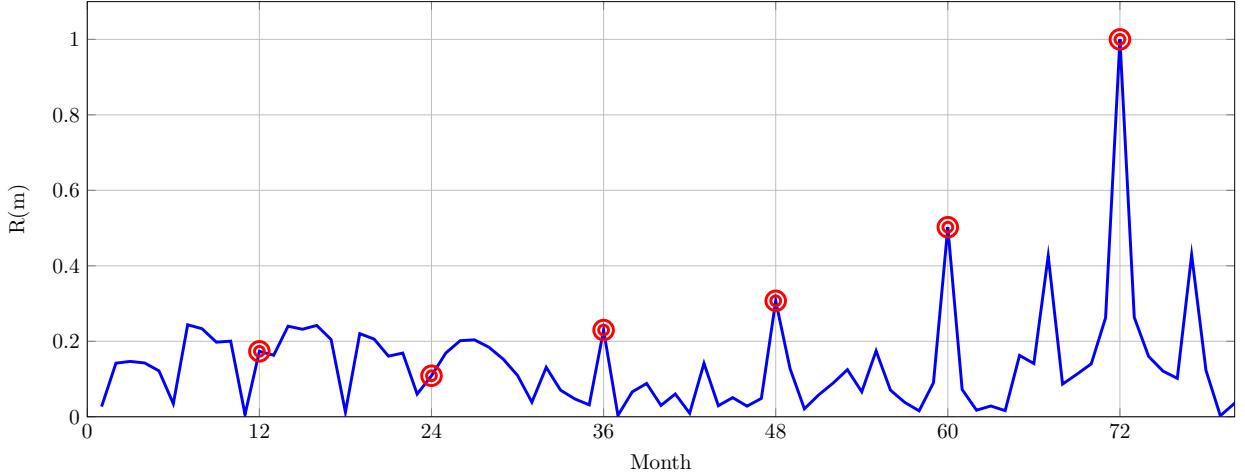


Figure 6.3: Estimated absolute normalized cross-correlation between the target data and the hole SPU data set. The sequence was trimmed due the zero-padding, and the red circles highlights where the lag m is a multiple of 12 months.

correlation peaks can be noted, their correlation are smaller than the noisy first two years, leading to the assumption that they do not provide much information.

6.2.2 Dataset Reshape

With the objective of incorporating the knowledge obtained from the time cross-correlation showed in the previous subsection, some changes were made in the overall modeling proposed. An exponential profile shows a good approximation for modeling the cross-correlation peaks, although the vicinity of the peaks demonstrates a very low correlation with the target data.

In spite the fact that an exponential profile is the main characteristic of the squared exponential covariance function, for it to be a good approximation the exponential profile is required to be present at all times. In this case, the cross-correlation profile shows that the tax collected 12 months before the prediction is more correlated than the tax collected on the previous month of the prediction.

In order to take advantage of the squared exponential covariance function in translating the peaks correlation profile and, at the same time, to carry the characteristics of the original data, this section proposes to convert the original one dimensional SPU data into a two dimensional array, with the first dimension indexed by month $\mathbf{M} = 1, 2, \dots, 12$ and the second dimension indexed by year $\mathbf{Y} = 1, 2, \dots, 6$. This leads to a reshape of the 1D data of Fig. 4.5 into a 2D data array presented at Fig. 6.4.

With this new array as the input of our Gaussian process, we can now separate the mean and the covariance function in a two dimensional structure, with different hyperparameters for it in

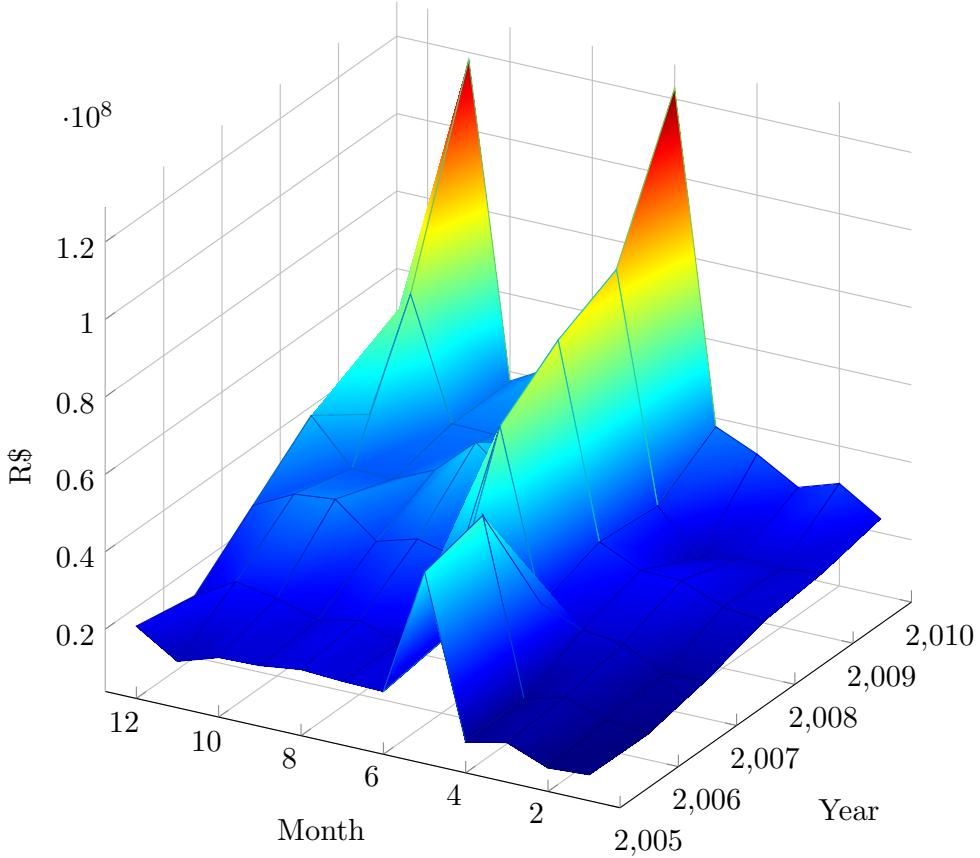


Figure 6.4: Plot of the SPU data set converted in a 2D array.

each dimension. Considering the cross-correlation profile of our data shown in Subsection 6.2.1, we will assume that only the amount of tax collected on January of 2005, 2006, 2007, 2008 and 2009 will influence the predictive amount of tax collected in January of 2010, and analogously to the other months. In other words, the information used by the predictor will be obtained exclusively from the highlights of Fig. 6.3. Therefore, from this point forward, the selected approach is to apply the final covariance function showed in (6.4) exclusively in the monthly dimension.

6.3 OPTIMIZATION AND PREDICTION RESULTS

This section describes the technique used to optimize the hyperparameters of the proposed covariance function and the resulting prediction using the optimum settings. In addition, we describe preliminary proposals for a classification stage aimed at future studies. In Subsection 6.3.1, the knowledge of the cross-correlation profile is applied into the covariance function model and the hyperparameters evaluation. In Subsection 6.3.2, the bidimensional resulting prediction is shown and in Subsection 6.3.3 a series of performance measurements and error comparisons are made with the previously obtained results, including comparisons with a similar approach using

Neural Networks proposed in the literature and a usual financial estimating technique. In Subsection 6.3.4, a classification stage based on the statistical description of GPR is discussed, labeling the data into regular or possibly fraudulent.

6.3.1 Hyperparameters Tuning

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 inferencing 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 [25]. Therefore, the marginal likelihood maximization was chosen to optimize the hyperparameter's set.

The marginal likelihood of the training data is the integral of the likelihood times the prior:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X})p(\mathbf{f}|\mathbf{X})d\mathbf{f}. \quad (6.7)$$

Recalling that \mathbf{X} is dependent of the hyperparameter's set Θ , [25] shows that the log marginal likelihood can be stated as:

$$\log p(\mathbf{y}|\mathbf{X}, \Theta) = -\frac{1}{2}\mathbf{y}^T \mathbf{K}_y^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K}_y| - \frac{n}{2} \log 2\pi. \quad (6.8)$$

In (6.8), $\mathbf{K}_y = \mathbf{K}_f + \sigma_n^2 \mathbf{I}$ is the covariance matrix of the noisy targets \mathbf{y} and \mathbf{K}_f is the covariance matrix of the noise-free latent \mathbf{f} . To infer the hyperparameters by maximizing the marginal likelihood in (6.7), [25] shows a numerically stable algorithm that seeks the partial derivatives of the logarithmic marginal likelihood in (6.8) with respect to the hyperparameters.

The methodology above described was used to determine the optimum set of hyperparameters $\hat{\Theta}$. However, [25] states two problems regarding this approach. The first one is that the likelihood distribution is multimodal, *i.e.* is dependent of the initial conditions of Θ . Also, the inversion of the matrix \mathbf{K}_y is computationally demanding.

In addition, our case presents another important restriction. Our final covariance function in (6.4) possess an hyperparameter θ_2 , one of the periodic covariance function's hyperparameters, that dictates the overall period of that function. As seen in Subsection 6.2.1, the optimum

periodicity of the covariance function should be within a finite set of multiples of 12, leading to $\hat{\theta}_2 = \{12, 24, 36, 48, 60\}$.

Imposing that restriction, the proposed algorithm for hyperparameter's optimization follows the sequence below:

- Define the initial values of the hyperparameter's set Θ ;
- Evaluate the marginal likelihood of the periodic component among the finite set of θ_2 , keeping the other hyperparameters fixed at their initial values;
- Choose the periodic hyerparameter with the maximum marginal likelihood;
- Evaluate the marginal likelihood of the resting hyperparameters, keeping the periodic hyerparameter fixed;
- Choose the final set of hyperparameters with the maximum marginal likelihood.

The initial hyperparameter's set is $\Theta = \{1; 12; 60\}$. The initial magnitude $\sigma_1^2 = 0.7$ and initial noise variance $\sigma_n^2 = 0.1$ were also treated as hyperparameters and, therefore, optimized together with the set Θ . As already discussed, the technique used to optimize the hyperparameters is the algorithm described in [25], whose optimization results are shown in Table 6.1.

Table 6.1: Optimized set of hyperparameters Θ , σ_1^2 and σ_n^2 after 100 iterations, using the marginal likelihood with the kernel in (6.4).

Predicted Month	θ_1	θ_2	θ_3	σ_n^2	σ_1^2
01. January	0.9907	12	1019	0.2653	0.6935
02. February	0.9360	12	1361	0.2670	0.6552
03. March	0.9151	12	71.12	0.3952	0.6406
04. April	0.8792	12	46.87	0.3662	0.6154
05. May	1.0012	12	23.02	1.5523	0.7008
06. June	1.0000	24	6465	0.5056	0.7000
07. July	0.8919	12	90.50	0.4273	0.4273
08. August	0.7594	12	48.60	0.5075	0.5315
09. September	0.8613	12	88.59	0.3749	0.6029
10. October	0.8994	12	39.55	0.4587	0.6296
11. November	1.0000	24	1252	0.3934	0.7000
12. December	0.8705	12	77.79	0.4636	0.6094

6.3.2 Bidimensional Prediction Results

Fig. 6.5 shows a plot of the predicted values using the optimized hyperparameters in Table 6.1, where it can be seen that the uncertainty of May's prediction is quite higher, presenting an optimized $\sigma_n^2 = 1.5523$, mainly because the tax collection profile changed drastically in the training data. This behavior contradicts the linear increasing trend that were 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.

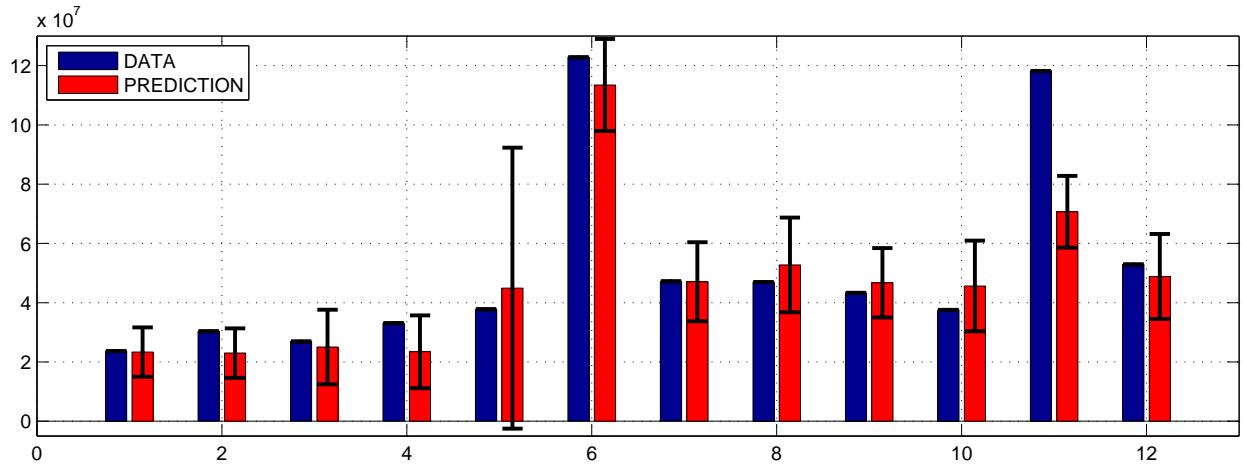


Figure 6.5: Plot of the Gaussian process prediction in blue, target SPU data in red. The error bars corresponds to a confidence interval of two standard deviations with respect to the predictive mean (around 95% of confidence).

6.3.3 Prediction Comparison and Error Metrics

The resulting prediction obtained in Subsection 6.3.2 will be evaluated by comparison with other predictive techniques and analyzed by different error metrics between the target data and the predictive data. The comparative evaluation will be made month-by-month with two other predictive approaches, one using an artificial neural network and another using an economical indicator. Also, an yearly comparison will be made with the projected tax collection, a revenue estimation made by the Brazilian federal government and published by SPU.

The approach proposed by [45] addressed the same problem, where an artificial neural network is used to predict the SPU tax collection for the year of 2010. On the other hand, a pure financial

approach consists of projecting the annual tax collection of SPU by readjusting the previous year's collection by an economic indicator. In this case, the chosen indicator to measure the inflation of the period is the National Index of Consumer's Prices (IPCA), consolidated by the Brazilian Institute of Geography and Statistics (IBGE). In 2009, the twelve month accumulated index was 4,31% [75].

The error metrics used in this subsection aim to evaluate the goodness of fit between the predicted and the testing data set for all the predictive approaches, using the mean squared error (MSE), the normalized mean squared error (NMSE), the root mean squared error (RMSE), the normalized root mean squared error (NRMSE), the mean absolute error (MAE), the mean absolute relative error (MARE), the coefficient of correlation (r), the coefficient of determination (d) and the coefficient of efficiency (e). The descriptive formulas of each metric are described in Appendix A.

All the predictive approaches, including the one proposed in this work, have their prediction error calculated with respect to the target data and the results are summarized in Table 6.2.

Table 6.2: Performance comparison by several error metrics

Error Metric	Optimum Value	Gaussian Process	Art. Neural Network	Inflation
MSE	0	22275×10^{10}	23777×10^{10}	35059×10^{10}
NMSE	0	0.20100	0.21455	0.31636
RMSE	0	14924×10^3	15419×10^3	18724×10^3
NRMSE	0	0.44833	0.46320	0.56246
MAE	0	87190×10^2	13207×10^3	13660×10^3
MARE	0	0.14830	0.31021	0.23222
r	1	0.90613	0.94585	0.96230
d	1	0.82107	0.89463	0.92603
e	1	0.78072	0.7659	0.67730

It is important to notice that the overall error in the Gaussian process prediction showed in Table 6.2 is mainly concentrated in November. Removing this month from the error measurements would lead to $\text{MSE} = 37782 \times 10^9$, $\text{NMSE} = 0.05127$, $\text{RMSE} = 61467 \times 10^2$, $\text{NRMSE} = 0.22644$, $\text{MAE} = 51924 \times 10^2$, $\text{MARE} = 0.12524$, $r = 0.97220$, $d = 0.94517$ and $e = 0.94359$.

Fig. 6.6 shows a comparative plot among the target data and all the predictive approaches side by side.

Finally, the Brazilian government revenue estimation, published by SPU on its annual report [68], projects an amount of tax collection by SPU in 2010 of R\$ 444,085,000.00, whereas the

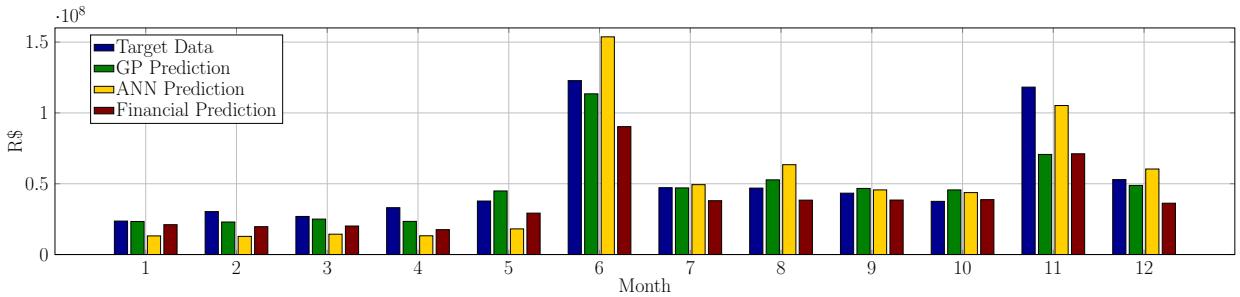


Figure 6.6: Monthly plot of target data and predictive results, in *Reais* (R\$), indexed by the m^{th} month.

total amount collected that year was R\$ 635,094,000.00 - a gross difference of 38.48% between the estimated and the executed amount of tax collection. The GPR approach presented in this work, in a yearly basis, projected a total tax collection amount of R\$ 620,703,197.42, resulting in a gross difference of 2.27% between the projected and executed amounts.

6.3.4 Classification Stage Proposals

The statistical description of the estimated variable, natively given by Gaussian processes in the regression stage, can be used to build heuristics to classify a predicted dataset into regular or possibly fraudulent. Here, we propose two different heuristics that are suitable to fraud detection scenarios. However, given the limited information publicly available from SPU regarding the dataset used in this work, the evaluation of the proposed schemes is incomplete and deserve to be better investigated in future studies.

The resulting regression obtained through GPR, presented in Fig. 6.5, shows the variance of the estimated variable as a measure of confidence by translating it into error bars. Since this confidence can be as large or as small as we desire it to be, it is possible to optimize a classification stage based on this information and, hence, build a trigger where high error bars means high probability of fraud and vice versa. In our case, without any doubt this system would classify May (month number 5) as a possibly fraudulent one. Despite the high uncertainty level of the prediction of this month, the prediction showed to be accurate when compared to the target data.

Another classification approach using the variance information can be build simply by confronting the predicted confidence interval with the real data, when it becomes available. In our case, this system would classify November (month number 11) as a possibly fraudulent one. SPU's annual report [68] states that an extraordinary revenue of R\$ 73,759,533.99 happened in 2010, but it is not possible to precise in which month it happened. In november, the difference between the predicted value and the actual revenue was R\$ 55,015,235.13.

Whereas the first proposed system returns the classified data in advance, together with the

predicted values in the regression stage, the second system needs the real revenue data in order to classify it. On the other hand, the second approach seeks for samples that are most dissimilar from the norm, whereas the first approach needs to be optimized in order to learn the norm and distinguish anomalous behaviors.

As previously mentioned, it is not possible to evaluate the performance of these classification stage proposals due to the limited information regarding our dataset, but the preliminary results using the statistical description of the estimated variable showed in this section encourages further studies on this topic.

Chapter 7

CONCLUSIONS

Business Intelligence is one of the most challenging and active fields of research nowadays. The multidisciplinary aspect of BI, not rarely embracing knowledge from exact and social sciences, makes its overall development not trivial. Business executives, end users, customers, CIO's and engineers are a few examples of what a BI solution in an organization must address. Often, a BI system must be broken into smaller portions to allow an expert suitable approach for each one of its parts.

Governmental BI systems aimed at fraud and irregularities detection are a continuously evolving topic due to the changing nature of fraudulent behavior. In addition, the exchange of ideas regarding fraud detection is limited in the public domain, as publishing information about fraud detection schemes ends up helping the circumvention of that particular scheme. It is not by chance that the technology and development process of major BI systems focused on fraud detection in financial institutions, banks, credit card issuers, governmental organizations, etc., are not made public.

In this work, we proposed to incorporate stages into existing BI systems maintained by MP, an agency of Brazilian federal government, to add predictive analytics capability, improve its performance, error rates and overall system responsiveness. The predominant goal in the BI systems addressed in this work is fraud detection. The BI system maintained by CGAUD runs audit trails on payrolls of the Brazilian public employees, whereas the BI system maintained by SPU seeks to predict the amount of tax to be collected by that branch of the government.

In the BI system of CGAUD, we proposed a statistical filter based on GMM applied in a BI environment. In its early versions, the BI system used a purely deterministic approach based on audit trails via concept maps to detect irregularities and inconsistencies in the payroll of the Brazilian public employees. With the insertion of the proposed statistical filter as a pre-processing stage of the BI system, it was possible to obtain a gain of efficiency in the overall system [2*].

The statistical filter developed in this work models a generative underlying pdf that governs the observed data set as a mixture of Gaussians. When applied to the real world data, the filter successfully reduced in 20% the amount of data to be analyzed by the audit trails, with a penalty of losing 8.16% in false negatives. In addition, the statistical filter is unique for all the audit trails, which extends its efficiency gain since each audit trail can use the one time filtered data as input. Finally, considering that Brazilian legislation set the rules for the payroll of federal public staff, the generative underlying pdf behavior should not present expressive changes if the legislation remains unmodified, which enables the GMM pdf of a certain month to be used in the subsequent years used without having to be recomputed [2*].

Considering that, nowadays, the BI system of CGAUD is capable of auditing approximately 5 billion *reais* each month, where the total payroll of the Brazilian public employees is around 12.5 billion *reais* [1*], an increase in the processing capacity of the BI system through a statistical pre-processing filter can lead to a more comprehensive auditory in the overall payroll, even with the penalty of false negatives that are intrinsic to any probabilistic model [2*].

Future developments in this particular area include predictive serial analytics, moving from a spacial analysis repeated every month to a predictive time series analysis, enabling the system to feedback itself and learn to track irregularities over time.

Regarding the BI system of SPU, we presented a GPR application, aimed to model the intrinsic characteristics of a specific financial series. A unidimensional model for the GPR's covariance function was proposed, and a pre-processing stage reshaped the original data set based on its cross-correlation profile. That approach empowered the use of a unidimensional GPR in a bidimensional environment by isolating high correlated months in one dimension and poor correlated months in another dimension.

Although Neural Networks are known for their flexibilities and reliable results when used for regression of time series, GPR are a transparent environment, with a parametric covariance function and no hidden layers, which can be an advantage when evaluating different components of a time series. 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 [3*].

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 [3*].

The regression results outperformed some classical predictive approaches such as ANN and economical indicator by several error metrics. In a yearly basis, the difference between the estimated and the real tax collection for 2010 using the approach proposed in this work was of 2.27%, whereas that difference reached 38.48% with the Brazilian government own estimation method [3*].

The approach explored in this work showed to be particularly useful for a small number of training samples, since the covariance function chosen to model the series results in a strong relationship for closer training points and a weak relationship for distant points. On the other hand, adding more training years before 2005 should not make a substantial difference in the prediction result using this method.

PUBLICATIONS FROM THIS WORK

Conference Publications

- [1*] A. M. R. Serrano, P. H. B. Rodrigues, R. C. Huacarpuma, J.P.C.L. da Costa, E. P. de Freitas, V. L. de Assis, A. A. Fernandes, R. T. de Sousa Jr., M. A. M. Marinho and B. H. A. Pilon, "Improved Business Intelligence Solution with Reimbursement Tracking System for the Brazilian Ministry of Planning, Budget and Management," in *6th International Conference on Knowledge Management and Information Sharing (KMIS)*, 2014.
- [2*] B. H. A. Pilon, J. P. C. L. da Costa, J. J. Murillo-Fuentes and R. T. de Sousa Jr., "Statistical Audit via Gaussian Mixture Models in Business Intelligence Systems," in *11th Brazilian Symposium on Information Systems (SBSI)*, 2015.

Publications to be Submitted

- [3*] B. H. A. Pilon, J. J. Murillo-Fuentes, J. P. C. L. da Costa and A. M. R. Serrano, "Gaussian Process for Regression in Business Intelligence: A Fraud Detection Application," to be submitted to *7th International Conference Conference on Knowledge Management and Information Sharing. (KMIS)*, 2015.

REFERENCES

- [1] H. Cheng, Y.-C. Lu, and C. Sheu, “An ontology-based business intelligence application in a financial knowledge management system,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 3614–3622, 2009.
- [2] A. McAfee and E. Brynjolfsson, “Big data: the management revolution.” *Harvard business review*, no. 90, pp. 60–6, 2012.
- [3] C. Hagen, K. Khan, M. Ciobo, J. Miller, D. Wall, H. Evans, and A. Yadav, “Big data and the creative destruction of today’s business models,” *AT Kearney Inc.*, 2013.
- [4] S. J. Dubner. (2008, 02) Hal Varian answers your questions. [Online]. Available: <http://freakonomics.com/2008/02/25/hal-varian-answers-your-questions/>
- [5] D. Laney, “3d data management: Controlling data volume, velocity and variety,” *META Group Research Note*, vol. 6, 2001.
- [6] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, “Big data: Issues and challenges moving forward,” in *System Sciences (HICSS), 2013 46th Hawaii International Conference on*. IEEE, 2013, pp. 995–1004.
- [7] A. Katal, M. Wazid, and R. Goudar, “Big data: Issues, challenges, tools and good practices,” in *Contemporary Computing (IC3), 2013 Sixth International Conference on*. IEEE, 2013, pp. 404–409.
- [8] J. Gantz and D. Reinsel, “Extracting value from chaos,” *IDC iview*, no. 1142, pp. 9–10, 2011.
- [9] J. Reinschmidt and A. Françoise, “Business intelligence certification guide,” *IBM International Technical Support Organisation*, 2000.
- [10] C. M. Olszak and E. Ziembka, “Business intelligence systems in the holistic infrastructure development supporting decision-making in organisations,” *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 1, no. 1, pp. 47–57, 2006.

- [11] R. J. Bolton and D. J. Hand, “Statistical fraud detection: A review,” *Statistical Science*, pp. 235–249, 2002.
- [12] D. Anderson, T. Frivold, and A. Valdes, *Next-generation intrusion detection expert system (NIDES): A summary*. SRI International, Computer Science Laboratory, 1995.
- [13] S. Ghosh and D. L. Reilly, “Credit card fraud detection with a neural-network,” in *System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on*, vol. 3. IEEE, 1994, pp. 621–630.
- [14] S. R. Campos, A. A. Fernandes, R. T. de Sousa Jr, E. P. De Freitas, J. P. C. L. da Costa, A. M. R. Serrano, D. D. C. Rodrigues, and C. T. Rodrigues, “Ontologic audit trails mapping for detection of irregularities in payrolls.” in *International Conference on Next Generation Web Services Practices (NWeSP)*, 2012, pp. 339–344.
- [15] A. A. Fernandes, L. C. Amaro, J. P. C. L. Da Costa, A. M. R. Serrano, V. A. Martins, and R. T. de Sousa, “Construction of ontologies by using concept maps: A study case of business intelligence for the federal property department,” in *Business Intelligence and Financial Engineering (BIFE), 2012 Fifth International Conference on*. IEEE, 2012, pp. 84–88.
- [16] R. C. Huacarpuma, D. d. C. Rodrigues, A. M. R. Serrano, J. P. C. L. da Costa, R. T. de Sousa Jr, M. Holanda, and A. P. F. Araujo, “Big data: A case study on data from the Brazilian ministry of planning, budgeting and management,” *IADIS Applied Computing 2013 (AC) Conference*, 2013.
- [17] K. Lee, L. Guillemot, Y. Yue, M. Kramer, and D. Champion, “Application of the Gaussian mixture model in pulsar astronomy-pulsar classification and candidates ranking for the fermi 2fgl catalogue,” *Monthly Notices of the Royal Astronomical Society*, vol. 424, no. 4, pp. 2832–2840, 2012.
- [18] D. Lu, E. Moran, and M. Batistella, “Linear mixture model applied to Amazonian vegetation classification,” *Remote sensing of environment*, vol. 87, no. 4, pp. 456–469, 2003.
- [19] M. K. Sönmez, L. Heck, M. Weintraub, E. Shriberg, M. Kemal, S. Larry, H. Mitchel, and W. E. Shriberg, “A lognormal tied mixture model of pitch for prosody-based speaker recognition,” *SRI International*, 1997.
- [20] M. A. Figueiredo and A. K. Jain, “Unsupervised learning of finite mixture models,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 3, pp. 381–396, 2002.
- [21] T. Hastie and R. Tibshirani, “Discriminant analysis by Gaussian mixtures,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 155–176, 1996.

- [22] S. Dalal and W. Hall, “Approximating priors by mixtures of natural conjugate priors,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 278–286, 1983.
- [23] C. K. Williams and D. Barber, “Bayesian classification with Gaussian processes,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 12, pp. 1342–1351, 1998.
- [24] F. Pérez-Cruz, S. Van Vaerenbergh, J. J. Murillo-Fuentes, M. Lázaro-Gredilla, and I. Santamaría, “Gaussian processes for nonlinear signal processing,” *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 40–50, 2013.
- [25] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press, 2006, ISBN 0-262-18253-X.
- [26] H. P. Luhn, “A business intelligence system,” *IBM Journal of Research and Development*, vol. 2, no. 4, pp. 314–319, 1958.
- [27] C. Elena, “Business intelligence,” *Journal of Knowledge Management, Economics and Information Technology*, vol. 1, no. 2, 2011.
- [28] M. Gibson, D. Arnott, I. Jagielska, and A. Melbourne, “Evaluating the intangible benefits of business intelligence: Review & research agenda,” in *Proceedings of the 2004 IFIP International Conference on Decision Support Systems (DSS2004): Decision Support in an Uncertain and Complex World*. Citeseer, 2004, pp. 295–305.
- [29] S. Negash, “Business intelligence,” *The Communications of the Association for Information Systems*, vol. 13, no. 1, p. 54, 2004.
- [30] K. Nazari and M. Emami, “Conceptual and theoretical foundations of business intelligence,” *amran Nazari et al./Elixir Inter. Busi. Mgmt*, vol. 46, pp. 8195–8202, 2012.
- [31] G. Gangadharan and S. N. Swami, “Business intelligence systems: design and implementation strategies,” in *Information Technology Interfaces, 2004. 26th International Conference on*. IEEE, 2004, pp. 139–144.
- [32] T. H. Davenport, “Competing on analytics.” *harvard business review*, no. 84, pp. 98–107, 2006.
- [33] H. Chen, R. H. Chiang, and V. C. Storey, “Business intelligence and analytics: From big data to big impact.” *MIS quarterly*, vol. 36, no. 4, pp. 1165–1188, 2012.
- [34] H. J. Watson and B. H. Wixom, “The current state of business intelligence,” *Computer*, vol. 40, no. 9, pp. 96–99, 2007.

- [35] T. H. Davenport and J. G. Harris, *Competing on analytics: the new science of winning*. Harvard Business Press, 2007.
- [36] M. Golfarelli, S. Rizzi, and I. Cella, “Beyond data warehousing: what’s next in business intelligence?” in *Proceedings of the 7th ACM international workshop on Data warehousing and OLAP*. ACM, 2004, pp. 1–6.
- [37] S. Chaudhuri, U. Dayal, and V. Narasayya, “An overview of business intelligence technology,” *Communications of the ACM*, vol. 54, no. 8, pp. 88–98, 2011.
- [38] R. Kimball, *The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses*. John Wiley & Sons, 1998.
- [39] W. H. Inmon, *Building the data warehouse*. John wiley & sons, 2005.
- [40] D. Robertson, “Global card fraud losses reach \$11.27 billion in 2012,” *Nilson Report, The*, no. 1023, p. 6, August 2013.
- [41] C. Phua, V. Lee, K. Smith, and R. Gayler, “A comprehensive survey of data mining-based fraud detection research,” *arXiv preprint arXiv:1009.6119*, 2010.
- [42] F. R. Bank, “The 2013 Federal Reserve payments study: Recent and long-term payment trends in the United States 2003–2012,” 2013.
- [43] L. Copeland, D. Edberg, A. K. Panorska, and J. Wendel, “Applying business intelligence concepts to medicaid claim fraud detection,” *Journal of Information Systems Applied Research*, vol. 5, no. 1, p. 51, 2012.
- [44] J. R. Dorronsoro, F. Ginel, C. Sánchez, and C. Cruz, “Neural fraud detection in credit card operations,” *Neural Networks, IEEE Transactions on*, vol. 8, no. 4, pp. 827–834, 1997.
- [45] A. M. R. Serrano, J. P. C. L. da Costa, C. H. Cardonha, A. A. Fernandes, and R. T. de Sousa Jr., “Neural network predictor for fraud detection: A study case for the federal patrimony department,” in *Proceeding of the Seventh International Conference on Forensic Computer Science (ICoFCS) 2012*. Brasília, Brazil: ABEAT, pp. 61–66.
- [46] J. Nagi, K. Yap, S. Tiong, S. Ahmed, and A. Mohammad, “Detection of abnormalities and electricity theft using genetic support vector machines,” in *TENCON 2008-2008 IEEE Region 10 Conference*. IEEE, 2008, pp. 1–6.
- [47] G. McLachlan and D. Peel, *Finite mixture models*. John Wiley & Sons, 2004.
- [48] J.-M. Marin, K. Mengersen, and C. P. Robert, “Bayesian modelling and inference on mixtures of distributions,” *Handbook of statistics*, vol. 25, pp. 459–507, 2005.

- [49] M. Aitkin and D. B. Rubin, “Estimation and hypothesis testing in finite mixture models,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 67–75, 1985.
- [50] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.
- [51] G. J. McLachlan and K. E. Basford, “Mixture models. inference and applications to clustering,” *Statistics: Textbooks and Monographs*, New York: Dekker, 1988, vol. 1, 1988.
- [52] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using gaussian mixture speaker models,” *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, 1995.
- [53] J. A. Bilmes *et al.*, “A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models,” *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.
- [54] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.
- [55] R. A. Redner and H. F. Walker, “Mixture densities, maximum likelihood and the EM algorithm,” *Society of Industrial and Applied Mathematics Review*, vol. 26, no. 2, pp. 195–239, 1984.
- [56] C. J. Wu, “On the convergence properties of the EM algorithm,” *The Annals of statistics*, pp. 95–103, 1983.
- [57] J. Bernardo, J. Berger, A. Dawid, A. Smith *et al.*, “Regression and classification using Gaussian process priors,” *Bayesian statistics*, vol. 6, p. 475, 1998.
- [58] R. A. Davis, “Gaussian process,” in *Encyclopedia of Environmetrics, Section on Stochastic Modeling and Environmental Change*, D. Brillinger, Ed. NY: Willey, 2001.
- [59] E. Cinlar, *Introduction to stochastic processes*. Courier Dover Publications, 2013.
- [60] R. Murray-Smith and A. Girard, “Gaussian process priors with ARMA noise models,” in *Irish Signals and Systems Conference*. Maynooth, 2001, pp. 147–152.
- [61] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms.” in *NIPS*, 2012, pp. 2960–2968.
- [62] M. Blum and M. Riedmiller, “Optimization of Gaussian process hyperparameters using Rprop,” in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2013.

- [63] F. Pérez-Cruz and O. Bousquet, “Kernel methods and their potential use in signal processing,” *Signal Processing Magazine*, vol. 21, no. 3, pp. 57–65, 2004.
- [64] C. K. Williams and C. E. Rasmussen, “Gaussian processes for regression,” 1996.
- [65] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
- [66] “Brasil, Approves the Organizational Structure of Ministry of Planning, Budget and Management,” Decree number 8189 of January 21, 2014.
- [67] SIAPE. (2015) Sistema Integrado de Administração de Recursos Humanos. [Online]. Available: <https://www.serpro.gov.br/conteudo-solucoes/produtos/administracao-federal/>
- [68] Secretaria de Patrimonio da União (SPU). (2011) Relatório de gestão 2010. Brasília, DF, Brazil. [Online]. Available: http://www.planejamento.gov.br/secretarias/upload/Arquivos/processo_contas/2010/SPU/1_SPU2010_Relatorio_de_Gestao.pdf
- [69] A. M. R. Serrano, “Study and implementation of a predictive analytics module for the business intelligence system of the Brazilian Ministry of Planning, Budget and Management,” *Universitat Politècnica de Catalunya*, 2012.
- [70] MPOG. (2013) Boletim Estatístico de Pessoal. [Online]. Available: http://www.planejamento.gov.br/secretarias/upload/Arquivos/servidor/publicacoes/boletim_estatistico_pessoal/2013/Bol207_Jul2013_2.pdf
- [71] “Brasil, Legal Regime of the Federal Public Employee,” Law number 8112 of December 11, 1990.
- [72] G. J. McLachlan, R. Bean, and D. Peel, “A mixture model-based approach to the clustering of microarray expression data,” *Bioinformatics*, vol. 18, no. 3, pp. 413–422, 2002.
- [73] C. E. Rasmussen, “The infinite Gaussian mixture model.” in *NIPS*, vol. 12, 1999, pp. 554–560.
- [74] S. J. Orfanidis, *Optimum signal processing: an introduction*. New York, NY: McGraw-Hill, 2007, ISBN 0-979-37131-7.
- [75] IBGE. (2013) Historical series of IPCA. [Online]. Available: www.ibge.gov.br/home/estatistica/indicadores/precos/inpc_ipca

Appendix A

ERROR METRIC FORMULAS

Being $\mathbf{t} \in \mathbb{R}^n$ a target vector with the desired values and $\mathbf{y} \in \mathbb{R}^n$ an output vector of a regression model, the goodness of fit between \mathbf{t} and \mathbf{y} will be given in terms of:

1. Mean Squared Error (MSE):

$$\frac{1}{n} \sum_{i=1}^n (t_i - y_i)^2$$

2. Normalized Mean Squared Error (NMSE):

$$\frac{\frac{1}{n} \sum_{i=1}^n (t_i - y_i)^2}{\text{Var}[\mathbf{t}]}$$

3. Root Mean Squared Error (RMSE):

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (t_i - y_i)^2}$$

4. Normalized Root Mean Squared Error (NRMSE):

$$\sqrt{\frac{\frac{1}{n} \sum_{i=1}^n (t_i - y_i)^2}{\text{Var}[\mathbf{t}]}}$$

5. Mean Absolute Error (MAE):

$$\frac{1}{n} \sum_{i=1}^n |t_i - y_i|$$

6. Mean Absolute Relative Error (MARE):

$$\frac{1}{n} \sum_{i=1}^n \left| \frac{t_i - y_i}{t_i} \right|$$

7. Coefficient of Correlation (r):

$$\frac{\sum_{i=1}^n (t_i - \bar{t})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (t_i - \bar{t})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

8. Coefficient of Determination (d):

$$\left(\frac{\sum_{i=1}^n (t_i - \bar{t})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (t_i - \bar{t})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2$$

9. Coefficient of Efficiency (e):

$$1 - \frac{\sum_{i=1}^n (t_i - y_i)^2}{\sum_{i=1}^n (t_i - \bar{t})^2}$$