PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Wilfredo Mamani Ticona

Estudo de Métodos de Mineração de Dados Aplicados à Gestão Fazendária de Municípios

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientador: Prof[a]. Marley Maria Bernardes Rebuzzi Vellasco
Co-orientador: Prof[a]. Karla Tereza Figueiredo Leite

Rio de Janeiro
Abril de 2013

Wilfredo Mamani Ticona

Estudo de Métodos de Mineração de Dados Aplicados à Gestão Fazendária de Municípios

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Profa. Marley Maria Bernardes Rebuzzi Vellasco
Orientadora
Departamento de Engenharia Elétrica – PUC-rio

Profa. Karla Tereza Figueiredo Leite
Co-orientadora
Centro Universitário Estadual da Zona Oeste

Prof. André Vargas Abs da Cruz
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Fabiano Saldanha Gomes de Oliveira
Centro Universitário Estadual da Zona Oeste

Prof. José Karam Filho
LNCC

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico

Rio de Janeiro, 19 de abril de 2013.

Wilfredo Mamani Ticona

Graduou-se em Engenharia de Sistemas pela Universidade Andina em 1998. Cursou mestrado em Engenharia de Sistemas na UNI (Universidad Nacional de Ingieniería) em 2005. Com foco em Sistemas de Gestão Integrada.

Ficha Catalográfica

Ticona, Wilfredo Mamani

Estudo de métodos de mineração de dados aplicados à gestão fazendária de municípios / Wilfredo Mamani Ticona ; orientador: Marley Maria Bernardes Rebuzzi Vellasco ; co-orientador: Karla Tereza Figueiredo Leite. – 2013.

128 f. : il. (color.) ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2013.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Previsão de tributos. 3. Redes neurais. 4. Fuzzy C-Means. 5. Irregularidades. 6. Gestão fazendária. I. Vellasco, Marley Maria Bernardes Rebuzzi. II. Leite, Karla Tereza Figueiredo. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Dedicado:
A meus pais Bartolomé e Simona

Agradecimentos

Agradeço as minhas orientadoras professoras Marley e Karla, pela parceria e pelas centenas de horas designadas para a orientação deste trabalho. Pela enorme paciência, confiança e pela extrema cordialidade.

Ao Capes e à PUC-Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Aos professores que participaram da Comissão examinadora, pelas suas contribuições e pela revisão precisa.

Aos professores do departamento, pelos conhecimentos fornecidos na Pós-graduação.

Ao pessoal técnico e administrativo, pelo suporte e toda ajuda prestada no decorrer do curso.

Ao pessoal do ICA (Projeto de Municípios Eficientes), em especial ao Bernardo, Rafael e Antônio pela parceria no desenvolvimento do projeto vinculado à presente dissertação.

Agradeço aos meus pais e meus irmãos pelo exemplo de bondade, carinho e honestidade. Pelo amparo nas horas difíceis.

Agradeço, ainda, aos meus amigos e colegas da PUC, Joseph Ballon, Eliomar Araújo, Roger Resmini, Ivan Silva, Edwin Maldonado, Javier Ortega e Samuel Gustavo.

Agradeço a Deus pela vida.

Resumo

Ticona, Wilfredo Mamani; Vellasco, Marley M. B. Rebuzzi; Leite, Karla. T. Figueiredo. **Estudo de Métodos de Mineração de Dados Aplicados à Gestão Fazendária de Municípios.** Rio de Janeiro, 2013. 128p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Os impostos arrecadados pelas prefeituras são revertidos para o bem comum, para investimentos (tais como infraestrutura) e custeio de bens e serviços públicos, como saúde, segurança e educação. A previsão de valores futuros a serem arrecadados é uma das tarefas que as prefeituras têm como desafio. Essa é uma tarefa importante, pois as informações obtidas das previsões são valiosas para dar apoio à decisão com relação ao planejamento estratégico da prefeitura. Sendo assim, a investigação de modelos de previsão de impostos municipais, através de técnicas inteligentes, é de grande importância para a administração municipal. Deste modo, um dos objetivos desta dissertação foi desenvolver dois modelos para previsão de impostos utilizando redes neurais. Um modelo considerando variáveis endógenas e outro considerando variáveis endógenas e exógenas. Outro grande desafio para as prefeituras são as irregularidades no pagamento de tributos (erro ou fraude), que também prejudica o planejamento estratégico. A fiscalização mensal de todos os contribuintes é uma tarefa impossível de se realizar devido à desproporção entre o número de contribuintes e o reduzido número de agentes fiscais. Assim, a investigação de métodos baseados em técnicas inteligentes para indicar os possíveis suspeitos de irregularidade, é importante para o desempenho das atividades do agente fiscal. Deste modo, outro objetivo desta dissertação foi desenvolver um modelo visando identificar possíveis suspeitos de irregularidades no pagamento do ISSQN (Imposto Sobre Serviços de Qualquer Natureza). Os modelos de previsão foram avaliados, com três estudos de caso usando dados do município de Araruama. Para o modelo de previsão utilizando variáveis endógenas utilizou-se dois estudos de caso: o primeiro caso para a previsão de Receitas da Dívida Ativa e o segundo caso para a previsão de Receitas Tributárias, e um terceiro estudo caso para o modelo de previsão do ISSQN, utilizando variáveis endógenas e exógenas. Essas previsões obtiveram resultados, que se julgam promissores, a despeito dos dados utilizados nos estudos de caso. Com relação à irregularidade, apesar de não ter sido possível avaliar os resultados obtidos, entende-se que a ferramenta poderá ser utilizada como indicador para novas diligências.

Palavras-chave

Previsão de Tributos; Redes Neurais; Fuzzy C-Means; Irregularidades; Gestão Fazendária.

Abstract

Ticona, Wilfredo Mamani; Vellasco, Marley M. B. Rebuzzi (Advisor); Leite, Karla. T. Figueiredo (Co-advisor). **Study of Data Mining Methods Applied to the Financial Management of Municipalities.** Rio de Janeiro, 2013. 128p. MSc. Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Taxes collected by city halls are reverted towards common welfare; investments (such as infrastructure), and funding of public goods, as services on health, safety and education. The prediction of tax revenues is one of the tasks that have as challenges the city hall. This is an important task; because the information obtained from these predictions are important to support the city halls with relation the strategic planning. Thus, the investigation of prediction models designed for tax revenues through intelligent techniques is of great importance for public administration. One of the goals of this dissertation was to develop two models to prediction tax revenue using neural networks. The first model was designed considering endogenous variables only. The latter, considered both endogenous and exogenous variables. Another major challenge for city hall are irregularities in the taxes payment (error or fraud), which also affect the strategic planning. A monthly of all taxpayers is an impossible task to accomplish, due to the disproportion between the number of taxpayers and the reduced number of tax agents. Thus, research of methods based on intelligent techniques that indicate possible irregularities, is of great importance for tax agents. This way, another objective of this dissertation was to develop a model to identify possible suspects irregularities in the payment of the ISSQN (tax services of any nature). Prediction models were evaluated with three case studies using data from the city hall of Araruama. For the prediction model using endogenous variable, two case studies we used: (i) active debt revenues prediction, (ii) tax revenues prediction and (iii) ISSQN prediction, the latter using both endogenous and exogenous variables. In spite of the data used in the case studies, the results obtained from modeling are promising. Regarding tax irregularities, even though is not possible to evaluate the obtained results, the developed tool may be used as an indicator for future applications.

Keywords

Tributes Forecasting; Neural Networks; Fuzzy C-Means; Fraud Detection; Financial Management.

Sumário

Lista de figuras

Lista de tabelas

Lista de siglas

ADCT	Ato das Disposições Constitucionais Transitórias
AMERJ	Associação Estadual de Municípios
BCB	Banco centra do Brasil
CAGED	Cadastro geral de empregados e desempregados
CNAE	Classificação nacional de atividades econômicas
CNPJ	Cadastro nacional da pessoa jurídica
CPF	Cadastro de pessoa física
DA	Dívida Ativa
EBRJ	Exportação de bens de Rio de janeiro
EQM	Erro Quadrático Médio
IBGE	Instituto Brasileiro de Geografia e Estatística
ICMS	Imposto sobre operações relativas à circulação de mercadorias e prestação de serviços
IPCA-E	Índice nacional de preços ao consumidor amplo especial
IPTU	Imposto sobre a propriedade predial e territorial urbana
IPVA	Imposto sobre a propriedade de veículos automotores
ISSQN	Imposto sobre serviços de qualquer natureza
ITBI	Imposto sobre a transmissão de bens imóveis
LC	Lei Complementar
LRF	Lei de Responsabilidade Fiscal
LSE	Método do Estimador por Mínimos Quadrados
MAPE	Mean absolute percentage error
MLP	Multi-Layer Perceptron
PCAM	Método de Análise de Componentes Principais Modificado
RNA	Redes Neurais Artificiais
SELIC	Sistema Especial de Liquidação e Custódia
SEPLAG	Secretaria de Planejamento e Gestão
SFMA	Secretaria de Fazenda do Município de Araruama
SIE	Método da Efetividade de uma Entrada Singular
SMF	Secretaria Municipal da Fazenda
TANSING	Hyperbolic Tangent Sigmoid
TCE	Tribunal de contas do estado do Rio de Janeiro

1
Introdução

A atividade fazendária evoluiu para uma ampla estrutura de administração dos recursos financeiros e do patrimônio dos municípios. A tradução dos tributos pagos pelos cidadãos em serviços de qualidade para a comunidade exige uma atuação responsável da Secretaria Municipal da Fazenda (SMF). Atualmente, para muitos municípios no território brasileiro, a SMF é o órgão da prefeitura encarregado da administração financeira, patrimonial, contábil e de material, além da arrecadação de tributos e rendas e do pagamento dos compromissos da municipalidade. Compete à SMF efetuar o lançamento de impostos e taxas devidos pelos contribuintes, receber e controlar a arrecadação procedida pela rede bancária, controlar saldos bancários, a dívida pública e efetuar os pagamentos dos compromissos do município. É função da SMF, ainda, prestar orientação fiscal ao contribuinte e realizar diligências fiscais com o objetivo de assegurar o cumprimento da legislação tributária municipal. Os serviços prestados estão relacionados aos tributos municipais: ISSQN, IPTU e ITBI. Esses impostos e taxas arrecadados pelas prefeituras são revertidos para o bem comum, para investimentos em infraestrutura e custeio de bens e serviços públicos, como saúde, segurança e educação (Araújo et al., 2007). Assim, para se realizar uma boa gestão são necessárias ferramentas e funcionalidades visando o apoio à decisão desse segmento administrativo.

1.1.
Motivação

Com o aumento da população, da quantidade de impostos e de valores arrecadados tornam-se necessárias ferramentas que apoie às decisões dos secretários e subsecretários de fazenda, além de inspetores e auditores fiscais.

No planejamento estratégico da prefeitura a aplicação dos recursos públicos está entre as tarefas mais importantes na gestão fazendária. Um dos insumos para a realização do planejamento estratégico é a previsão de valores futuros a serem arrecadados. Essa é uma tarefa importante, pois as informações obtidas da previsão de arrecadação são valiosas para o fluxo de caixa futuro do

município. Os tributos provenientes de contribuinte relativo à pessoa física ou jurídica podem estar em maior ou menor grau, atrelados às características socioeconômicas, que devem ser avaliadas e identificadas de forma a contribuir para uma melhor previsão de arrecadação. Sendo assim, a investigação de modelos de previsão de arrecadação fiscal municipal (p. ex., tributo ISSQN) é de grande importância para todos os níveis da administração pública (municipal, estadual ou federal).

Outro grande desafio para as prefeituras é a irregularidades no pagamento de tributos (erro ou fraude), que também prejudica o planejamento orçamentário dos governantes. A fiscalização mensal de todos os contribuintes é uma tarefa impossível de se realizar devido à desproporção entre o número de contribuintes e ao reduzido número de agentes fiscais. Os dados obtidos sobre os pagamentos de tributos são, normalmente, bastante ruidosos, indicando a necessidade de se estudar modelos específicos de filtragem de dados, de forma a garantir um dado mais confiável.

Assim, a investigação de métodos baseados em técnicas inteligentes para prever arrecadações futuras e indicar os possíveis suspeitos de irregularidade de ISSQN é fundamental para o desempenho das atividades da secretaria de fazenda municipal.

1.2. Objetivo

O objetivo principal desta dissertação foi desenvolver três modelos de apoio à Gestão Fazendária Municipal, baseadas em técnicas inteligentes, visando melhorar a acurácia das previsões de arrecadação e na identificação de Irregularidades no pagamento do ISSQN.

O primeiro modelo prevê Receitas Tributárias e Receitas de Divida Ativa, baseado em Redes Neurais Artificiais (RNA) a partir de variáveis endógenas. O segundo modelo prevê o ISSQN, utilizando RNA e empregando variáveis endógenas (variáveis explicativas construídas a partir da variável que se deseja prever), e exógenas (variáveis explicativas construídas a partir de outras variáveis distintas da variável que se deseja prever). O terceiro modelo identifica irregularidades no pagamento do ISSQN dos contribuintes, baseada no algoritmo de clusterização Fuzzy C-Means (FCM), agrupando contribuintes cujo comportamento é semelhante de forma que se possa analisar as mudanças de comportamento dos contribuintes intra e inter grupos.

O que se pretende com os modelos desenvolvidos é:

- Melhorar a qualidade (diminuição do erro MAPE) das previsões de Receitas da Divida Ativa, Receitas Tributárias e ISSQN.
- Identificar contribuintes com diversos tipos de Irregularidades no pagamento do ISSQN.

A qualidade da previsão de Receitas e a identificação de Irregularidades no pagamento do ISSQN são fundamentais para dar apoio à decisão com relação ao planejamento estratégico da prefeitura na aplicação dos recursos públicos. Para os estudos de caso foram utilizados dados do Município de Araruama.

1.3. Descrição do Trabalho

Este trabalho consiste em uma breve descrição sobre Impostos Municipais, Dívida Ativa e Sonegação do ISSQN. Além disso, também foram incluídos os principais conceitos sobre Redes Neurais Artificiais (RNA) e métodos de seleção de variáveis, além de uma breve descrição do modelo Fuzzy C-Means, utilizado no modelo proposto para avaliação de irregularidades.

Os métodos de seleção de variáveis, como serão visto adiante, têm como foco reduzir as entradas da rede neural, no intuito de selecionar as variáveis mais relevantes do ponto de vista do objetivo final. Pelo fato de existirem vários métodos de seleção de variáveis, após alguns testes realizados, foram escolhidos os métodos que mais se adequaram à previsão de impostos.

Em métodos de previsão, a inclusão de variáveis exógenas relevantes no modelo visa aumentar a acurácia da previsão. Assim, foram investigadas possíveis variáveis exógenas socioeconômicas que poderiam contribuir para melhorar (diminuição do erro MAPE) a previsão da variável que se deseja prever (ISSQN). Após alguns testes com o software Econométrico Eview (Stock e Watson, 2007), foram escolhidas as variáveis exógenas que apresentaram correlações significativas. Posteriormente é realizado a seleção de entradas utilizando os métodos PCAM, LSE, SIE e ReliefF. É importante salientar que os métodos de seleção de variáveis apresentam apenas, uma noção (a eficiência depende da adequabilidade do método aos dados) de quais variáveis têm mais influência na previsão (Fayal, 2005).

Os resultados obtidos demostraram mais uma vez a grande capacidade das redes neurais na previsão de receitas e a melhoria substancial de qualidade que a utilização de variáveis exógenas traz para a previsão de receitas.

Espera-se que o modelo proposto para previsão de receitas usando variáveis endógenas e exógenas socioeconômicas para Gestão Municipal, relacionada com os menores erros de previsão encontrados, possa também ser estendida para a previsão de outras receitas como IPTU, Receitas Tributárias, Receitas de Divida Ativa e para diversos municípios.

Em relação à identificação de irregularidades no pagamento do ISSQN, foi realizado o cruzamento de bases de dados do Município e do CAGED. O cruzamento é um importante instrumento contra possíveis fraudes já que ampliam a busca de indícios de irregularidades. Para agrupar os contribuintes foi utilizado o algoritmo *Fuzzy C-means* (Bezdek, 1981). Os resultados obtidos facilitam aos fiscalizadores na identificação de irregularidades (erro ou fraude).

1.4. Organização da Dissertação

Esta dissertação inclui cinco capítulos adicionais, com conteúdos resumidos na sequência:

O capítulo 2 é incluído uma revisão bibliográfica, cujo objetivo é apresentar a teoria e conceitos importantes sobre Gestão Fazendária, ISSQN, Dívida Ativa, Receitas Tributárias e sonegação do ISSQN. Além disso, também foram incluídos trabalhos relevantes já publicados sobre previsão de receitas.

No capítulo 3, descrevem-se os conceitos necessários para a compreensão de métodos e modelos, tais como Redes Neurais Artificiais e Fuzzy C-Means, além dos métodos de seleção de variáveis SEI, LSE, PCAM e ReliefF, os quais foram utilizados nas soluções propostas nessa dissertação.

O quarto capítulo apresenta em detalhe, os três modelos desenvolvidos nessa dissertação, indicando a metodologia para sua aplicação. O primeiro modelo é para a previsão *multistep* (até 12 passos à frente) de Receitas da Dívida Ativa, Receitas Tributárias e ISSQN, empregando variáveis endógenas. O segundo modelo prevê o ISSQN até 12 passos à frente *(multistep),* empregando variáveis endógenas e exógenas. O terceiro modelo se propõe a indicar contribuintes suspeitos de Irregularidades no pagamento do ISSQN.

No quinto capítulo, destacam-se os resultados alcançados com quatro estudos de casos obtidos a partir da aplicação dos três modelos propostos. Os estudos de caso foram criados a partir dos dados do município de Araruama e variáveis exógenas.

No capítulo 6, são apresentadas as conclusões geras sobre os modelos e metodologias desenvolvidas, a partir da modelagem por RNA, com a utilização de variáveis endógenas e exógenas socioeconômicas, e a proposta para identificação de irregularidades de ISSQN desenvolvido, visando facilitar a identificação de irregularidades (erro ou fraude) e, finalmente, os trabalhos futuros que possam complementar ou melhorar os modelos propostos.

2
Gestão fazendária

Atualmente, para muitos municípios no território brasileiro, a Secretaria Municipal da Fazenda (SMF) é o órgão da prefeitura encarregado da administração financeira, patrimonial, contábil e de material, além da arrecadação de tributos, rendas e do pagamento dos compromissos da municipalidade. Assim, para se realizar uma boa gestão e cumprir as imposições da Lei de Responsabilidade Fiscal (LRF), são necessárias ferramentas e funcionalidades visando o apoio à decisão desse segmento administrativo.

O ponto de partida da Lei de Responsabilidade Fiscal (LRF) é o planejamento. O Capítulo III, Seção I, da LRF trata especificamente da previsão e da arrecadação de tributos (Khair, 2010; Caldart, 2006), um dos pontos que é tratado nesse trabalho.

A seguir serão brevemente descritos aspectos relacionados à arrecadação de tributos e impostos.

Impostos Municipais — Os impostos municipais representam fontes de recursos próprios para os cofres municipais, mas para o ingresso dessa receita é necessário que haja a instituição de tais impostos por lei. Tal lei deve seguir a competência reservada aos municípios pela Constituição Federal, princípio da reserva legal.

A instituição de Tributos sob a forma de legislação municipal coloca à disposição dos municípios um leque de Tributos (ISSQN, ITBI, IPTU, Taxas e Contribuição de Melhoria). Assim, quando a estrutura Fazendária municipal está organizada e o potencial econômico do município for promissor, a arrecadação própria é uma consequência natural.

Por outro lado, a sociedade em geral, por pagar impostos ao município, exige em contra partida que o município coloque a sua disposição serviços de educação, saúde, transporte e outros (Paiva, 2005).

2.1.
Receita tributária municipal

Segundo Bremaeker (2011) a receita tributária municipal é constituída pelas seguintes espécies tributárias:

- Impostos
  - Imposto predial e territorial urbano (IPTU)
  - Imposto sobre serviços de qualquer natureza (ISSQN)
  - Imposto sobre a transmissão de bens imóveis inter vivos (ITBI – inter vivos)
- Taxas
  - Taxas pela utilização de serviços públicos
  - Taxas pelo exercício do poder de polícia
- Contribuição de melhoria

A competência tributária municipal se iniciou com a Constituição de 1934, vindo a ser reforçada com o passar dos tempos.

Na Tabela 1 é apresentada a distribuição da receita orçamentária municipal e da receita tributária municipal, segundo as grandes regiões do Brasil do ano 2010.

Tabela 1. Distribuição da receita orçamentária municipal e da receita tributária municipal, segundo as grandes regiões

Brasil e Grandes Regiões	Receita Orçamentária Municipal	Receita Tributária Municipal	%*
BRASIL	319.800.226.643,00	56.522.396.985,00	17,67
Norte	20.091.675.994,00	2.058.651.608,00	10,25
Nordeste	66.890.041.064,00	6.426.026.892,00	9,61
Sudeste	163.549.444.417,00	37.719.303.405,00	23,06
Sul	49.900.024.401,00	7.678.644.614,00	15,39
Centro-oeste	19.369.040.767,00	2.639.770.466,00	13,63

* (coluna 3 / coluna 2)*100.

A receita orçamentária é a receita total recebida pelo município e receita tributária é aquela proveniente de tributos e impostos diretamente arrecadada pelos municípios.

2.1.1.
Imposto Sobre Serviços de Qualquer Natureza (ISSQN)

O Imposto Sobre Serviços (ISS), também chamado de Imposto Sobre Serviços de Qualquer Natureza (SSQN). O ISS é tributo de competência municipal, cobrança que onera o consumo, sendo de grande importância no quadro de impostos discriminados na Federação brasileira, por sua significação econômica (Cardoso, 2012).

De acordo com Albuquerque (1998) dado que o ISSQN é auto-lançável, ele é mais susceptível à evasão fiscal. É o próprio contribuinte quem declara o valor de sua receita bruta, calcula o valor do imposto e o paga. Para o Município só a ação da fiscalização poderá dizer se houve ou não evasão tributária.

Segundo Martins (2010) o ISSQN é de competência dos Municípios (e do Distrito Federal, CF, art. 147), a possibilidade de se instituir este imposto encontra-se na Constituição Federal, no art. 156, III, cuja redação é:

> "Art. 156. Compete aos Municípios instituir impostos sobre:
> (...)
> III – serviços de qualquer natureza, não compreendidos no art. 155, II, definidos em lei complementar.
> § 3° Em relação ao imposto previsto no inciso III, cabe à lei complementar:
> I – fixar as alíquotas máximas;
> II – excluir da sua incidência exportações de serviços para o exterior."

O ISSQN, como qualquer tributo, tem como pressupostos essenciais os seguintes: elemento material (fato gerador da obrigação tributária), elemento pessoal (contribuinte) e elemento quantitativo (base de cálculo e alíquota fiscal).

2.1.1.1.
Fato gerador

Dispõe o Código Tributário Nacional que o fato gerador da obrigação tributária "é a situação definida em lei como necessária e suficiente à sua ocorrência" (art. 114). Situação quer dizer o conjunto de fatos que são definidos em lei dando ensejo à exigência do tributo.

Somente a lei é que poderá estabelecer o fato gerador da obrigação tributária principal (art. 97, III, do CTN). A situação de fato estabelecida na lei será necessária e suficiente, ou seja, é bastante para a ocorrência do fato gerador da obrigação tributária.

Segundo Jèze (1937) o fato gerador é "o fato ou conjunto de fatos que permitem aos entes do fisco exercer sua competência legal de criar um crédito de tal importância, a título de tal imposto, contra tal contribuinte".

2.1.1.2. Aspecto pessoal

O aspecto pessoal é composto pelo sujeito ativo (o credor da obrigação tributária: União, Estado, Distrito Federal ou Município) e o sujeito Passivo (o devedor da obrigação tributária).

Sujeito ativo — O sujeito ativo do ISSQN, nos moldes do art. 156, inciso III, da Constituição Federal de 1988 e art. 1 da Lei Complementar (LC) 116/2003, é o Município (ou o Distrito Federal) onde o serviço for prestado, independentemente do local da sede do prestador (art. 3 da Lei Complementar 116/2003). Em que pese este dispositivo afirmar que o sujeito ativo será aquele onde estiver localizado o estabelecimento do prestador de serviços (ou seu domicílio caso não haja estabelecimento), fato é que o art. 4 da LC 116/2003 estabelece: "Considera-se estabelecimento prestador o local onde o contribuinte desenvolva a atividade de prestar serviços, de modo permanente ou temporário, e que configure unidade econômica ou profissional, sendo irrelevante para caracterizá-lo as denominações sede, filial, agência, posto de atendimento, sucursal, escritório de representação ou contato ou quaisquer outras que venham a ser utilizadas" (Melo, 2005).

Sujeito Passivo — Em linhas gerais, sujeito passivo é a pessoa, física ou jurídica, privada ou pública, da qual se exige o cumprimento da obrigação (Carvalho, 2010).

O contribuinte do ISSQN sempre será o prestador do serviço, uma vez que, conforme disposto no art. 121, inciso I, do Código Tributário Nacional, o contribuinte é a pessoa que guarda relação direta com o fato que é alvo da incidência do tributo em questão.

Ademais, tal condição está expressamente prevista na Lei Complementar nº 116/03, "Art 5º Contribuinte é o prestador do serviço" (Melo, 2005).

2.1.1.3. Base de cálculo

A base de cálculo do ISSQN é o preço do serviço, conforme previsto no art. 7° da Lei Complementar n° 116/2003. O preço do serviço constitui-se no número de unidades monetárias que se paga para adquirir um bem incorpóreo

(imaterial): serviço. O preço do serviço é o valor da contraprestação relativa ao fornecimento de trabalho (Martins, 2010).

2.1.1.4. Alíquota do ISSQN

Alíquota é o percentual que irá incidir sobre a base de cálculo para apurar-se o montante devido do tributo.

Cabe à lei complementar fixar alíquotas máximas e mínimas. A Lei Complementar n° 116/2003, no entanto, limitou-se a fixar a alíquota máxima (5%), nada definindo sobre a alíquota mínima. Enquanto uma alíquota mínima não for determinada por lei complementar, vale o disposto no artigo 88 do Ato das Disposições Constitucionais Transitórias (ADCT), que fixa o mínimo de 2% (Cardoso, 2012).

2.1.1.5. Local de incidência tributária

O artigo 3° da Lei Complementar n° 116/2003 determina que "o serviço considera-se prestado e o imposto devido no local do estabelecimento prestador ou, na falta do estabelecimento, no local do domicílio do prestador".

2.2. Dívida Ativa

Nos âmbitos legais a Dívida ativa é definida, como créditos do ente público, estes podem ser de origem tributária ou não, sendo que a arrecadação destes é importante para "o equilíbrio autossustentável" (Nogueira e Sato, 2008).

O Crédito Tributário é o vínculo jurídico, de natureza obrigacional, por força do qual o Estado (sujeito ativo) pode exigir do particular, o contribuinte ou responsável (sujeito passivo), o pagamento do tributo ou da penalidade pecuniária (objeto da relação obrigacional) (Machado, 2002).

Um aspecto importante é quando existem tributos devidos não pagos no vencimento, porém reconhecidos por parte do contribuinte. Esses valores não pagos no vencimento devem ser inscritos em dívida ativa, normalmente no ano subsequente ao ano fiscal do vencimento do tributo ou imposto devido. Nesse caso, quando o tributo está escriturado corretamente, ou seja, o contribuinte

declara corretamente, mas não paga, não se configura crime. A próxima seção faz uma breve explanação sobre dívida ativa.

Classificação da dívida ativa — A Dívida Ativa constitui-se em um conjunto de direitos ou créditos de várias naturezas, em favor da Fazenda Pública, com prazos estabelecidos na legislação pertinente, vencidos e não pagos pelos devedores, por meio de órgão ou unidade específica instituída para fins de cobrança na forma de lei (Filho *et al*., 2004).

Em Castro *et al*., (2005) a dívida ativa é classificada em dívida ativa tributaria e não tributaria:

"Dívida Ativa Tributária, que é o crédito da Fazenda Pública dessa natureza, proveniente de obrigação legal relativa a tributos e respectivos adicionais e multas";

"Dívida Ativa não Tributária, que são os demais créditos da Fazenda Pública, tais como os provenientes de empréstimos compulsórios, contribuições estabelecidas em lei, multas de qualquer origem ou natureza, exceto as tributárias, foros, laudêmios, aluguéis ou taxas de ocupação, custas processuais, preços de serviços prestados por estabelecimentos públicos, indenizações, reposições, restituições, alcance dos responsáveis definitivamente julgados, bem assim os créditos decorrentes de obrigações em moeda estrangeira, de sub-rogação de hipoteca, fiança, aval, ou de garantias de contratos em geral ou outras obrigações legais".

No entanto quando o contribuinte tenta esconder, propositalmente ou não, o fato que gerou o imposto, não escriturado total ou parcialmente o fato, configura-se um crime. A seção a seguir descreve brevemente a sonegação fiscal.

2.3. Sonegação Fiscal

De acordo com Smanio (2005), sonegação fiscal é a ocultação dolosa, mediante fraude, astúcia ou habilidade, do recolhimento de tributo devido ao Poder Público. Note-se, porém, que a lei não conceituou o que seja sonegação fiscal, adotando outro critério de identificação, no artigo 1º da Lei n. 8.137/90, qual seja, considerando delitos contra a ordem tributária a supressão ou redução de tributo ou contribuição social ou acessório, e depois enumerando,

taxativamente, quais as modalidades de conduta que podem levar a tal supressão ou redução, constituindo genericamente o que seja sonegação fiscal.

Segundo Ferreira (2002) a sonegação "funda-se na ocultação voluntária da ocorrência do fato gerador, com o fim de não pagar o tributo devido".

A fraude serve para caracterizar o engano malicioso ou a ação astuciosa, promovida de má-fé, para ocultação da verdade ou fuga ao cumprimento do dever. Nestas condições, a fraude traz consigo o sentido do engano, não como se evidencia no dolo, em que se mostra a manobra fraudulenta para induzir outrem à prática de ato, de que lhe possa advir prejuízo, mas o engano oculto para furtar-se o fraudulento ao cumprimento do que é de sua obrigação ou para logro de terceiros. É a intenção de causar prejuízo a terceiros (Souza, 2003).

2.3.1. Sonegação do ISSQN

Segundo Amaral *et al.* (2009) o imposto sobre serviços (ISS) está entre os tributos mais sonegados pelas empresas brasileiras. A Tabela 2 apresenta uma classificação de tributos sonegados com seus respectivos índices.

Tabela 2. Classificação e índices de sonegação das empresas brasileiras

Colocação	Impostos	Índice de Sonegação (dados 2006, 2007 e 2008)
1°	O INSS	27,75 %
2°	ICMS	27,14 %
3°	Imposto de Renda	26,64 %
4°	ISS	25,02 %
5°	Contribuição Social Sobre Lucro	24,89 %
6°	Imposto de Importação	24,83 %
7°	PIS/ COFINS	22,13 %
8°	IPI	19,08 %
9°	IOF	16,55 %
10°	CPMF	4,03 %

Para a determinação deste índice foi considerada a relação faturamento e recolhimento de tributos dos diversos setores da economia com dados de 2006, 2007 e 2008. Os tributos sonegados pelas empresas somam R$ 200 bilhões por ano. Em valores, a sonegação é maior no setor industrial, seguido das empresas do comércio e das prestadoras de serviços.

2.3.2.
Tipos de fraude do imposto

Em Doehlen (1997), apresentam-se tipos mais comuns de fraude tributária:

O principal tipo de fraude é a não emissão de documento fiscal. Nesse caso, o contribuinte utiliza para acobertar a prestação de serviços documentos não autorizados: recibos, contratos, orçamentos, qualquer documento de controle interno, variando de acordo com sua atividade econômica.

Em segundo lugar, deve-se mencionar o "calçamento" de notas fiscais. Nesse procedimento, o contribuinte lança na via fixa do bloco de notas um valor diferente do constante na via do cliente. O cruzamento com as informações provenientes do tomador do serviço é o meio mais importante para constatar essa fraude. No entanto, esse cruzamento entre pessoas físicas e jurídicas nem sempre é possível, por falta de informação por parte das pessoas físicas que tomam os serviços. A redução é geralmente de 90% do efetivo valor da nota, pela facilidade de cortar um zero e mudar a vírgula na quantia lançada. Mas, em alguns casos, a fraude é superior a estes 90%. A sonegação do imposto pode, assim, corresponder a muitas vezes o valor pago.

No caso de notas frias, paralelas, o contribuinte utiliza outros elementos que simulam o documento verdadeiro. Ele engana o fisco e os clientes que acreditam estarem recebendo um documento válido. Muitas vezes adulteram algum detalhe da nota fiscal ou, no caso da nota fria, utilizam autorização inexistente para a emissão do documento. No caso da nota paralela, ele tem autorização para emitir o documento, mas confecciona, com base nele, mais alguns, conforme sua necessidade. Para os serviços de valor pequeno, ele emite o documento verdadeiro. Para os de valor maior, utiliza documentos falsos.

Em Queiroz (1997) há um caso de sonegação conhecida como "caixa 2" de uma empresa do ramo de alimentação que foi denunciada por uma ex-funcionária de alto nível, que apresentou toda a documentação com cópias de documentos referentes ao caixa 2 da empresa. Esta era mantida em conta corrente em outro banco e chegava a corresponder a 2/3 de todo o movimento financeiro da empresa, incluindo folha de salários, movimentação de mercadorias e todo o movimento financeiro e econômico. Tudo que seria de interesse a qualquer fiscalização estava distribuído em duas contas bancárias e em duas contabilidades, como se fossem duas empresas distintas. Alguns dos melhores e mais preparados fiscais, com bons conhecimentos de contabilidade e de informática, estão debruçados sobre o caso, mas desde já se sabe que a

maior parte da movimentação da empresa ficará de fora do levantamento, se não for “quebrado o sigilo bancário” da segunda conta corrente, referente ao caixa 2. Este é um caso típico em que a flexibilização do sigilo bancário seria altamente proveitosa, evitando a impunidade fiscal.

Outro tipo de fraude é o contribuinte oculto ou “laranjas”. Não é o cadastrado, mas um empresário falsificado, oculto. Ele realiza fraudes de difícil apuração e verificação. Favatto (1998) descreve que essas empresas são constituídas em nome de pessoas estranhas, as “laranjas”. O verdadeiro empresário esconde-se atrás de nomes falsos ou utilizados sem consentimento. Há um procedimento que evitaria esse tipo de fraude: cada empresário, ao constituir uma empresa, requereria seu cadastramento na Secretaria de Fazenda ou na Receita Federal.

Há empresas constituídas em nome de pessoas que sequer têm um teto para morar e, frequentemente, nem sabem que seu nome está sendo utilizado para esse fim. Há empresas constituídas em nome de pessoas inexistentes ou que já morreram. Quando os débitos devidos são constatados, procura-se o proprietário e se deparam com pessoas que não possuem qualquer relação com a firma. Consideram-se os empresários que utilizam esses subterfúgios como verdadeiros fraudadores, que atrapalham a atividade dos contribuintes sérios. Essas manobras são realizadas com o apoio de mentores, como contadores. São profissionais que não passam de veiculadores de ideias, mas é o empresário que aceita essas sugestões ilícitas, sendo o responsável principal.

No que diz respeito ao não cumprimento das normas tributárias, é importante distinguir duas práticas que frequentemente aparecem nas verificações fiscais. Uma dessas práticas é a “fraude”, que pode ser definida, em termos tributários, como toda ação ou omissão, praticada com ardil, astúcia, malícia ou má-fé, que impede ou modifica a ocorrência de fato gerador, visando reduzir ou não pagar o imposto devido. A segunda prática é o “conluio” que pode ser entendido como um ajuste entre duas ou mais pessoas visando à sonegação.

Seja qual for a forma de sonegação, voluntária ou involuntária, quando ela é comprovada tem como consequência a aplicação de punições pelas administrações tributárias. As penalidades vão desde multas pecuniárias, passando por restrições até a limitação de alguns direitos e vantagens (Filho, 2005).

Segundo Andrade (2009), diminuir ao máximo a sonegação fiscal é uma busca constante das administrações fazendárias, pois esta redução traz diversos benefícios para a sociedade, como por exemplo;

- Mais recursos para os investidores do Estado em educação, saúde, transporte, entre outros;
- O pagamento dos impostos devidos por todos possibilita o fim da concorrência desleal. Não é justo que as empresas se beneficiem por não pagar seus impostos em detrimento daquelas que os pagam em dia, pois isso causa um desequilíbrio de mercado.

2.4. Trabalhos relacionados

Na literatura existem vários trabalhos publicados sobre previsão de arrecadação de receitas, utilizando métodos convencionais:

Em Pereira (2007), foi realizado a previsão mensal a curto prazo do ISSQN do município de Teresina, foram manipulados vários modelos, onde o modelo SARIMA demonstrou ser mais robusto em relação ao modelo VAR, com resultados do Erro Quadrático Médio (EQM) de 4,5% e 7,0% respectivamente.

Outro trabalho sobre previsão mensal a curto prazo do ISSQN do Município de Caxias do Sul é apresentado em Caldart (2006), onde foram realizados as previsões para quatro períodos adiante, para os meses de maio a agosto do 2005, as quais foram comparadas com os valores efetivamente arrecadados. Os resultados foram satisfatórios com erro de -0,657% e conclui-se que a aplicação de técnicas econométricas para fazer previsões do ISSQN, e em especial a utilização de modelos de séries temporais do tipo auto-regressivo integrado (ARI) torna-se uma importante alternativa em substituição às técnicas tradicionais.

Em Liebel (2004), foram utilizados modelos para a previsão do Imposto sobre operações relativas à circulação de mercadorias e prestação de serviços (ICMS) do estado do Paraná. No estudo, segundo o autor, não identificou um padrão sazonal relevante nas séries estudadas, porém os resultados apresentados indicam que o modelo de suavização exponencial de Winters aditivo é o mais indicado para realização das previsões de ICMS quando utilizado as últimas 72 observações. Quando se utiliza as últimas 36 observações, o modelo escolhido é o de suavização exponencial de Holt. No

geral, os resultados apontam uma pequena margem de erro, ou seja, os modelos tornam-se um importante subsídio para tomada de decisões por parte dos gestores públicos.

Outro trabalho sobre ICMS do Rio Grande do Sul, com erros inferiores a 2% é apresentado em Guaragna e Mello (2002), onde foram utilizados variáveis endógenas e fatores exógenos (Índice Geral de Preços - Disponibilidade Interna).

Campos (2009), em trabalho de abrangência regional, aplicaram metodologias de modelos dinâmicos univariados e multivariados para a análise de três séries mensais da arrecadação, relativas ao Imposto de Importação (II), Imposto Sobre a Renda das Pessoas Jurídicas (IRPJ) e Contribuição para o Financiamento da Seguridade Social (COFINS), tributos de competência federal. Os resultados foram comparados entre si, por meio da raiz quadrada do erro médio quadrático de previsão (RMSE) e comparados com a modelagem ARIMA e com o método dos indicadores, utilizado pela Secretaria da Receita Federal do Brasil (RFB). Considerados os melhores modelos de cada série, foi alcançada a redução média do RMSE de 42% em relação ao erro cometido pelo método dos indicadores e de 35% em relação à modelagem ARIMA, além da drástica redução do erro anual de previsão.

Poucos trabalhos têm sido publicados na arrecadação de impostos utilizando redes neurais, em (Contreras, 2005) é apresentado um trabalho sobre previsão de arrecadação do ICMS baseado em redes neurais, onde mostra que os resultados foram mais precisos do que aquelas fornecidas por metodologias de previsão tradicionais como o alisamento exponencial e o método Box e Jenkins.

3
Técnicas inteligentes

3.1.
Redes Neurais Artificiais

As Redes Neurais Artificiais (RNA), usualmente denominadas Redes Neurais, são modelos computacionais não lineares, inspirados na estrutura e operação do cérebro humano, que procuram produzir características humanas, tais como: aprendizado, associação, generalização e abstração. As Redes Neurais são compostas por diversos elementos processadores (neurônio artificial), altamente interconectados, que efetuam operações simples, transmitindo seus resultados aos processadores vizinhos. A habilidade das Redes Neurais em realizar mapeamentos não-lineares entre suas entradas e saídas as tem tornado prósperas no reconhecimento de padrões e na modelagem de sistemas complexos. Devido à sua estrutura, as Redes Neurais são bastante eficazes no aprendizado de padrões a partir de dados não-lineares, incompletos com ruído ou compostos por exemplos contraditórios. Foi demonstrado que o algoritmo *Back Propagation* é um aproximador universal (Haykin, 1999), sendo capaz de aprender qualquer mapeamento de entrada-saída.

3.1.1.
Neurônio Artificial

O neurônio artificial é uma estrutura lógico-matemática que procura simular a forma, o comportamento e as funções de um neurônio biológico. Assim sendo, os dendritos foram substituídos por *entradas*, cujas ligações com o corpo celular artificial são realizadas através de elementos chamados de *peso* (simulando as sinapses). Os estímulos captados pelas entradas são processados pela *função de soma*, e o limiar de disparo do neurônio biológico foi substituído pela *função de transferência*. O neurônio biológico, ilustrado de maneira simplificada na Figura 1, serve como inspiração para o modelo do neurônio artificial, cujo esquema é apresentado na Figura 2.

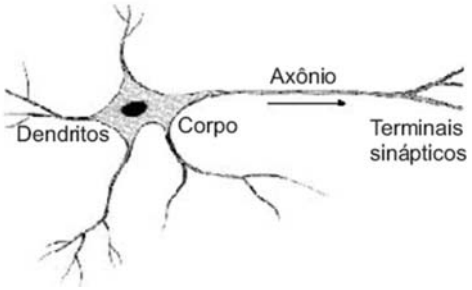


Figura 1. Neurônio Biológico

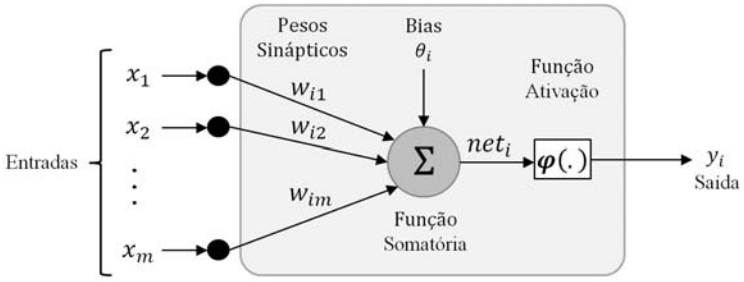


Figura 2. Modelo de Neurônio Artificial (Mcculloch e Pitts, 1943)

Observa-se que o neurônio artificial apresenta um conjunto de entradas, representadas por $x_1, x_2, \ldots, x_m$, simulando os dendritos, e uma saída, representada por y_1, simulando o axônio. As entradas do neurônio são ponderadas por pesos sinápticos, representados por $w_{i,1}, w_{i,2}, \ldots, w_{i,m}$, e somadas, fornecendo o potencial interno do processador, representado por net_i. Especificamente, o sinal de entrada x_j da sinapse j conectado ao neurônio i é multiplicado pelo peso sináptico $w_{i,j}$. O bias, representado por θ_i, é um termo de polarização do neurônio artificial, que pode ser tratado como um peso sináptico cuja entrada é sempre |1|, e seu objetivo é aumentar (ou diminuir) a influência do valor da combinação linear das entradas, deslocando o hiperplano da origem (Haykin, 1999).

A saída do neurônio é obtida pela aplicação de uma *função de ativação*, representada por $\varphi(.)$, ao net_i, como pode ser visto na seguinte equação 3.1:

$$y_i = \varphi(net_i) = \varphi\left(\sum_{j=1}^{m} x_j\, w_{i,j} + \theta_i\right) \tag{3.1}$$

A *função de ativação* é utilizada para restringir a amplitude da saída de um neurônio. A função de ativação é também referida como *função restritiva* já que restringe (limita) o intervalo permissível de amplitude do sinal de saída a um valor finito. Normalmente, o intervalo normalizado da amplitude da saída de um neurônio é escrito como o intervalo unitário fechado $[0,1]$ ou $[-1,1]$.

3.1.2.
Arquitetura de Redes Neurais

Uma rede neural artificial (RNA) ou rede neural é um sistema composto por vários neurônios que estão ligados por conexões, os quais podem formar umas variedades de arquiteturas de redes. Em geral, podem-se identificar três classes de arquiteturas para as redes neurais:

Rede Feed Forward com Camada Única

É a rede mais simples, esta formada por uma camada de entrada e uma camada de saída de neurônios (nós computacionais). Cada neurônio apresenta uma resposta, que constitui uma das saídas da rede. A Figura 3 apresenta a arquitetura de uma rede *Feed Forward* de camada única. O primeiro modelo de rede neural, denominado *Perceptron*, foi proposto por Rosenblatt (1962), e é capaz de aprender somente problemas linearmente separáveis.

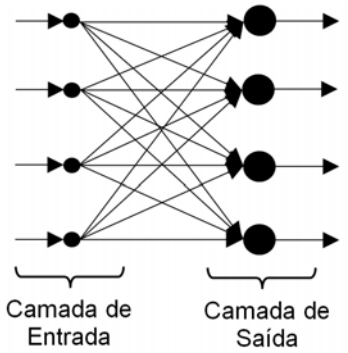


Figura 3. Rede Feedforward com uma única camada de neurônios

Rede Multilayer Perceptron (MLP)

As redes do tipo MLP, também denominado rede *Feed Forward*, são redes multicamadas, formadas por uma camada de entrada, uma ou mais camadas ocultas (escondidas), e uma camada de saída, como pode ser visto na Figura 4. Cada neurônio da camada oculta recebe os sinais de todos os nós da camada de entrada, através de suas conexões. O nó da camada de saída processa a informação transmitida pela camada escondida e gera a saída da rede. As redes MLP, com uma ou duas camadas ocultas e um número suficiente de neurônios, são consideradas aproximadores universais, ou seja, podem representar uma grande gama de funções (Barron, 1993). Segundo Cybenko (1989), apesar de que as Redes MLP possam ter mais de uma camada escondida, uma simples

camada escondida é suficiente para que essas redes aproximem quaisquer funções não-lineares.

Não existe um método que determine o número ideal de camadas ocultas e de neurônios nessa camada, porém a escolha desses parâmetros é muito importante e influencia diretamente o desempenho do sistema, pois o tempo computacional para o cálculo da resposta e para o treinamento da rede aumenta significativamente com o aumento das conexões e de neurônios nas camadas ocultas.

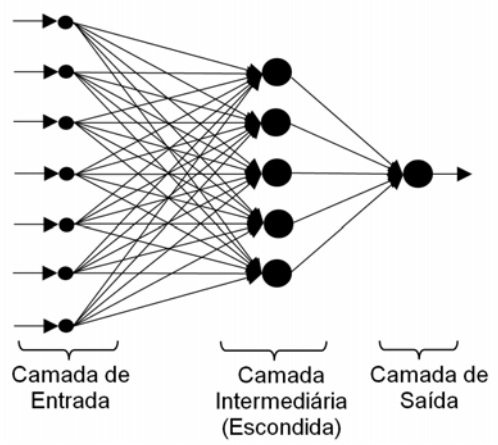


Figura 4. Rede Multilayer Perceptron

As redes MLP são ainda caracterizadas pela elevada possibilidade de aplicações em diversos tipos de problemas relacionados com as mais diferentes áreas do conhecimento, sendo também consideradas uma das arquiteturas mais versáteis quanto à aplicabilidade (Silva *et al*., 2010). Entre essas potenciais áreas, têm-se os seguintes destaques:

- Aproximação universal de funções;
- Reconhecimento de padrões;
- Identificação e controle de processos;
- Previsão de séries temporais;
- Otimização de sistemas.

Redes Recorrentes ou Realimentadas

Uma rede neural recorrente se distingue de uma rede MLP por ter pelo menos um laço de realimentação, onde as saídas dos neurônios são realimentadas como sinais de entrada para neurônios da mesma camada ou de camadas anteriores. A existência de realimentação qualifica tais redes para o processamento dinâmico de informações, isto é, elas podem ser utilizadas em sistemas variantes no tempo. Entre os principais tipos de redes que possuem

realimentação estão as redes de *Hopfield* (Haykin, 1999) e a rede Elman (Silva *et al*., 2010). Uma arquitetura de Rede Recorrente é apresentada na Figura 5.

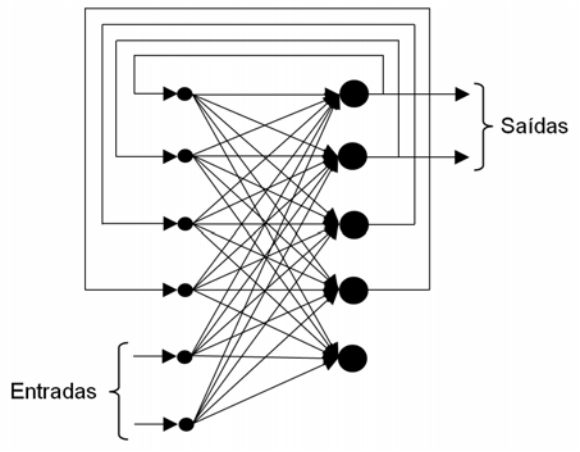


Figura 5. Rede Recorrente ou Realimentada

3.1.3. Aprendizagem e treinamento de redes neurais

As redes neurais são geralmente treinadas de forma iterativa (Silva *et al*., 2010). O processo de treinamento apresentado na Figura 6 pode ser resumido nos seguintes passos (Sánchez, 2011):

1. Inicialização dos *pesos* e *bias* da rede neural, com valores aleatórios pequenos, e do contador de interações em zero.
2. Simulação da rede com os dados de treinamento.
3. Cálculo do erro das estimativas da rede com os seus valores alvo.
4. Comparação do erro com um valor limite especificado pelo usuário.
5. Se a rede não atende aos critérios de desempenho-alvo e se o número máximo de iterações não foi ultrapassado, ir para (a), caso contrário, ir para (b).
    a. Atualização dos *pesos* e *bias*, incrementando do contador e de iterações; ir para a etapa 2.
    b. Parar o treinamento.

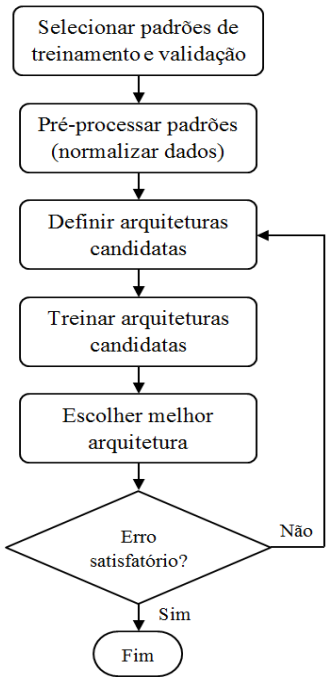


Figura 6. Fase de treinamento (Silva et al., 2010)

O treinamento das redes neurais artificiais pode ser realizado com vários algoritmos além do algoritmo *Backpropagation*. Diversas variações de método *backpropagation* têm sido propostas com o objetivo de tornar o processo de convergência mais eficiente. Entre tais aperfeiçoamentos, tem-se o método de *Levenberg-Marquardt,* um método de otimização publicado primeiramente por *Kenneth Levenberg* (Levenberg, 1944) e aperfeiçoado por *Donald Marquardt* (Marquardt, 1963). Esse algoritmo é considerado o método mais rápido para treinamento de redes *feedforward backpropagation*, que possui uma quantidade moderada de pesos sinápticos (Saini e Soni, 2002).

3.1.4. Previsão multi-step

Em previsão multi-step os valores previstos são acrescentados à base de dados de entrada para prever valores futuros (Chakraborty et al., 1992). Um exemplo de previsão multi-step 12 passos à frente é apresentado na Tabela 3, onde os valores previstos da coluna 14 são acrescentados à janela de entrada a cada passo, para prever valores futuros.

Tabela 3. Previsão multi-step

	Janela de Entrada												
Passos à frente	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Previsto**
1	Jan-09	Feb-09	Mar-09	Apr-09	May-09	Jun-09	Jul-09	Aug-09	Sep-09	Oct-09	Nov-09	Dec-09	**Jan-10**
2	Feb-09	Mar-09	Apr-09	May-09	Jun-09	Jul-09	Aug-09	Sep-09	Oct-09	Nov-09	Dec-09	**Jan-10**	**Feb-10**
3	Mar-09	Apr-09	May-09	Jun-09	Jul-09	Aug-09	Sep-09	Oct-09	Nov-09	Dec-09	**Jan-10**	**Feb-10**	**Mar-10**
4	Apr-09	May-09	Jun-09	Jul-09	Aug-09	Sep-09	Oct-09	Nov-09	Dec-09	**Jan-10**	**Feb-10**	**Mar-10**	**Apr-10**
5	May-09	Jun-09	Jul-09	Aug-09	Sep-09	Oct-09	Nov-09	Dec-09	**Jan-10**	**Feb-10**	**Mar-10**	**Apr-10**	**May-10**
6	Jun-09	Jul-09	Aug-09	Sep-09	Oct-09	Nov-09	Dec-09	**Jan-10**	**Feb-10**	**Mar-10**	**Apr-10**	**May-10**	**Jun-10**
7	Jul-09	Aug-09	Sep-09	Oct-09	Nov-09	Dec-09	**Jan-10**	**Feb-10**	**Mar-10**	**Apr-10**	**May-10**	**Jun-10**	**Jul-10**
8	Aug-09	Sep-09	Oct-09	Nov-09	Dec-09	**Jan-10**	**Feb-10**	**Mar-10**	**Apr-10**	**May-10**	**Jun-10**	**Jul-10**	**Aug-10**
9	Sep-09	Oct-09	Nov-09	Dec-09	**Jan-10**	**Feb-10**	**Mar-10**	**Apr-10**	**May-10**	**Jun-10**	**Jul-10**	**Aug-10**	**Sep-10**
10	Oct-09	Nov-09	Dec-09	**Jan-10**	**Feb-10**	**Mar-10**	**Apr-10**	**May-10**	**Jun-10**	**Jul-10**	**Aug-10**	**Sep-10**	**Oct-10**
11	Nov-09	Dec-09	**Jan-10**	**Feb-10**	**Mar-10**	**Apr-10**	**May-10**	**Jun-10**	**Jul-10**	**Aug-10**	**Sep-10**	**Oct-10**	**Nov-10**
12	Dec-09	**Jan-10**	**Feb-10**	**Mar-10**	**Apr-10**	**May-10**	**Jun-10**	**Jul-10**	**Aug-10**	**Sep-10**	**Oct-10**	**Nov-10**	**Dec-10**

3.1.5. Correlação Cruzada

Em Bourke (1996) a Relação cruzada ou correlação cruzada é um método padrão para estimar o grau no qual duas séries são correlacionadas em função de um atraso aplicado a um deles. Considerando duas séries $x(i)$ e $y(i)$ donde $i = 0,1,2,\dots,N-1$. A correlação cruzada r com atraso d é definida pela equação 3.2:

$$r(d) = \frac{\sum_i [(x(i) - mx) * (y(i-d) - my)]}{\sqrt{\sum_i (x(i) - mx)^2}\ \sqrt{\sum_i (y(i-d) - my)^2}} \tag{3.2}$$

Em que:

r = correlação cruzada

d = atraso entre duas séries

x e y = séries

mx = média da série x

my = média da série y

Os limites da correlação cruzada $r(d)$ estão entre $-1 \leq r(d) \leq 1$:

$r(d) = 0$, indica que não há correlação

$r(d) = 1$, indica correlação elevada

$r(d) = -1$, indica correlação elevada, mas o inverso de uma das séries.

3.1.6.
Método de seleção de variáveis

Na maioria das aplicações reais de previsão, as entradas da rede contêm um grande número de variáveis irrelevantes e ou redundantes, o que aumenta o tempo de processamento computacional. Deste modo, um problema comum nestas aplicações é a seleção das características ou variáveis mais relevantes do ponto de vista do objetivo final, dentre de todas as variáveis de entrada da rede (Dash *et al*., 1997). O modelo de tratamento descrito nesta seção tem essa missão, ou seja, reduzir e otimizar as variáveis de entradas da rede.

Existem inúmeras técnicas para a tarefa de seleção de variáveis. Dentre os algoritmos de seleção de variáveis, que são independentes do modelo (*Model Free*) que possuem capacidade de escolha de variáveis em menor tempo e a um menor custo computacional que os algoritmos dependentes do modelo, podem-se citar: Correlação cruzada, Autocorrelação, Estimador por Mínimos Quadrados (*LSE - Least Squares Estimator*), Método da Efetividade de uma Entrada Singular (*SIE - Single Input Effectiveness*) e (PCAM - *Principal Components Analasys Modified*).

3.1.6.1.
Método do Estimador por Mínimos Quadrados (LSE)

Seja um sistema de n entradas e uma saída. O método LSE calcula a importância da i-ésima variável de entrada x_i estimando o i-ésimo parâmetro b_i da função F que descreve a variação da variável de saída Δy em relação à variação de cada i-ésima variável de entrada Δx_i sobre o conjunto completo de dados. A função F é dada pela equação 3.3:

$$F = \Delta y = b_1[\Delta x_1] + b_2[\Delta x_2] + \cdots + b_n[\Delta x_n] \qquad (3.3)$$

Os componentes do vetor Δy são obtidos subtraindo-se os valores da variável de saída nos padrões da base de dados em combinações de duas a duas, e os componentes do vetor Δx_i são obtidos subtraindo-se os valores da variável de entrada x_i nos padrões da base de dados em combinações de duas a duas.

Da eq. (3.3), pode-se dizer que cada parâmetro b_i representa a importância da i-ésima variável de entrada com respeito à saída no sentido

estatístico. O cálculo dos parâmetros b_i é feito mediante um algoritmo de mínimos quadrados (Chung e Duan, 2000).

3.1.6.2. Método da Efetividade de uma Entrada Singular (SIE)

O objetivo deste método é calcular o ranking das entradas mediante a definição do SIE, que é o grau de efetividade de cada entrada na saída (Cao e Rossiter, 1997; Cao e Biss, 1996).

Considerando o sistema linear $y = Gu$ de m saídas e r entradas, neste método a importância de cada variável de entrada é obtida expressando-se o vetor de entradas u do sistema como a soma de duas projeções ortogonais u_n e u_c, uma pertencente ao espaço nulo da matriz de transferência G do sistema e a outra pertencente ao complemento ortogonal do espaço nulo de G respetivamente. Portanto, a efetividade de cada variável de entrada x_i é calculada como a norma da projeção da entrada no canal i sobre o complemento ortogonal do espaço nulo de G, já que a norma da projeção da entrada no canal i sobre o espaço nulo de G não tem efeito na variável de saída.

Maiores detalhes sobre este método podem ser encontrados em (Cao e Rossiter, 1997; Cao, 1995; Contreras, 2002).

3.1.6.3. ReliefF

O algoritmo ReliefF (Kira e Rendell, 1992) trabalha por meio da amostragem aleatória de exemplos do conjunto de dados e localização do vizinho mais próximo da mesma classe e do vizinho mais próximo da classe oposta. Os valores dos atributos dos vizinhos mais próximos são comparados aos da classe amostrada e utilizados para atualizar os pesos de relevância de cada atributo em relação à classe. Esse processo é repetido um número m de vezes. A ideia do ReliefF é que atributos importantes devem diferenciar exemplos de classes diferentes e possuir valores similares para exemplos da mesma classe. A proposta original do algoritmo ReliefF, a qual permitia trabalhar com duas classes, foi mais tarde estendida no algoritmo ReliefF para lidar com ruído e conjuntos de dados contendo múltiplas classes (Kononenko, 1994). No ReliefF, a influência de ruído nos dados é amenizada por meio da distribuição da contribuição dos k vizinhos mais próximos da mesma classe do exemplo correntemente considerado e de k vizinhos mais próximos de cada uma das

classes diferentes do exemplo amostrado, ao invés de considerar apenas um único vizinho mais próximo.

Maiores detalhes sobre o método encontram-se nas referências (Kononenko, 1994; Santoro, 2005).

3.1.6.4. Método de Análise de Componentes Principais Modificado (PCAM)

O método de análise de componentes principais tradicional é uma transformação de coordenadas, que pode ser usado para redução de dados. O PCAM modificado no algoritmo permite utilizá-lo como método de seleção de variáveis (Hall, 2005).

O PCA original têm as seguintes características:

- Matriz de dados de entrada: $X[n\ x\ k]$
- Descompõe-se como: $X = v_1P_1^T + v_2P_2^T + \cdots + v_kP_k^T$
- Define-se uma quantidade $L \leq k$: $X = v_1P_1^T + v_2P_2^T + \cdots + v_LP_L^T + E$

O PCAM (Modificado): $X = v_1P_1^T + v_2P_2^T + \cdots + v_kP_k^T$, têm as siguintes caracteristicas:

- Vetores de loading: $P_j[k\ x\ 1],\quad j = 1 \ldots L$
- Vetores de score: $v_j[n\ x\ 1],\quad j = 1 \ldots L$
- Matriz de loadings: $P[k\ x\ L]$
- Matriz de scores: $V[m\ x\ L]$

$$v_j = X * p_j \quad \rightarrow \quad V = X * P$$

No PCA original a matriz $X[m\ x\ k]$ é substituída pela matriz $V[m\ x\ L]$. O problema é que se perde o sentido físico com as novas variáveis $V_j[n\ x\ 1], j = 1 \ldots L$, isto por que o PCA tradicional é um método de redução de dimensionalidade, e não de seleção de variáveis.

O PCAM (Modificado) implica em uma seleção sobre as variáveis originais. Onde para o primeiro componente principal: $p1[k\ x\ 1]$, cada elemento de $p1$ indica o peso da variável original xj na combinação linear que define a variável modificada $v1$. Onde maior valor absoluto em $p1$ indica maior importância.

3.2. Fuzzy C-Means (FCM)

Em agrupamentos tradicionais (não nebulosos), o limite de diferentes grupos é *crisp,* ou seja, cada objeto pertence a um e somente um grupo, isto é conhecido como agrupamento *hard.* Assim os grupos nesses tipos de abordagens são disjuntos. Agrupamentos *Fuzzy* (nebulosos) estende essa noção para permitir associar um objeto com todos os grupos usando uma função de pertinência (Zadeh, 1965).

O algoritmo *Fuzzy C-means* (FCM), originalmente introduzido por (Bezdek, 1981), é uma extensão fuzzy do método de clusterização *k-means,* em que cada registro pode pertencer a mais de um agrupamento, de acordo com seu valor de pertinência.

Seja $X = \{x_1, x_2, \dots, x_n\}$ um conjunto de dados, deseja-se particionar os elementos em p conjuntos nebulosos de forma a otimizar a função objetivo da equação 3.4, onde p é o número de *clusters.*

O resultado do agrupamento nebuloso pode ser expresso através da seguinte matriz U:

$$U = \begin{bmatrix} \mu_1(x_1) & \cdots & \mu_j(x_1) \\ \vdots & \ddots & \vdots \\ \mu_1(x_n) & \cdots & \mu_j(x_n) \end{bmatrix}$$

Cada coluna da matriz U está associada a um grupo e cada linha a um registro. Onde $\mu_j(x_i)$ é um valor entre $[0,1]$ que indica o grau de pertinência de cada elemento x_i para um determinado j-ésimo *cluster.*

O algoritmo FCM minimiza a seguinte critério:

$$J\left(\mu_j(x_i), C_k\right) = \sum_{i=1}^{n} \sum_{j=1}^{p} {\mu_{ij}}^m {d_{ji}}^2 \qquad (3.4)$$

Onde $m > 1$ é o coeficiente nebuloso responsável pelo grau de "fuzificação" dos elementos da matriz U e C_k é o centróide do k-ésimo *cluster*.

Quanto maior é o coeficiente nebuloso m, mais nebuloso se torna a matriz U. Quando $m = 1$ a função objetivo $J\left(\mu_j(x_i), C_k\right)$ é reduzida ao caso *crisp* do método de agrupamento k-*means*.

O valor de pertinência μ é calculado utilizando a equação 3.5,

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ji}{}^2}\right)^{\frac{1}{m-1}}}{\left(\frac{1}{d_{k1}{}^2}\right)^{\frac{1}{m-1}}} \tag{3.5}$$

Em que:

$\mu_j(x_i)$: é o valor de pertinência de x_i no j^o cluster

C_j : é o centro do j^o clusters $(j = 1,2, \dots, p)$

d_{ji} : distancia do x_i no cluster C_j

m : é o parâmetro de fuzzificação

P : é o número de clusters especificados

d_{ki} : é a distancia do x_i no cluster c_k

A soma dos valores de pertinência de um registro a todas as classes é igual a um:

$$\sum_{j=1}^{p} \mu_j(x_i) = 1 \tag{3.6}$$

No Quadro 1 são apresentados os passos do algoritmo do Fuzzy c-Means.

Quadro 1. Algoritmo Fuzzy c-Means (FCM)

1. **Inicialização:** definir o número de agrupamentos K; definir o parâmetro de forma $m > 1$ e escolher uma estimativa inicial para os centros de agrupamentos C.
2. Calcular a matriz de partição inicial
3. Enquanto o critério de parada não for alcançado, **repita:**
4. Calcular os novos valores de centros de agrupamento:

$$C_j = \frac{\sum_i \left[\mu_j(x_i)\right]^m x_i}{\sum_i \left[\mu_j(x_i)\right]^m}$$

5. Atualizar a matriz de partição eq. 3.4.
6. **Fim**

3.2.1.
Índice de validação de clusters

Vários índices de validação de clusters têm sido propostos na literatura nos últimos anos (Vendramin *et al*., 2009). Um índice de validação de *clusters* é um valor quantitativo que permite avaliar a análise de agrupamentos quanto:

- Ao número de clusters; e
- À qualidade da partição dos grupos.

A maioria dos índices são baseados na estrutura geométrica da partição, avaliando o quanto clusters diferentes estão separados e a dispersão dos registros em cada clusters.

3.2.1.1.
Método gráfico de avaliação da clusterização - Silhueta

A silhueta é um coeficiente que mede o quão bem alocado cada elemento está ao seu grupo comparado aos outros clusters formados (Rousseeuw, 1987), (Vendramin *et al*., 2009). Esta medida é calculada em termos da média da distância euclidiana entre o elemento i e todos os elementos do grupo de i, comparado com as médias das distâncias entre i e os elementos do grupo vizinho.

Para determinar o valor de silhueta $s(i)$, utilizamos a siguinte equação 3.7:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i),\ b(i)\}} \tag{3.7}$$

A Figura 3, é utilizada para mostrar os elementos da equação, onde:

- $a(i)$ é a média da distância do i-ésimo objeto a todos os objetos do cluster A.
- Considerando um *cluster* C diferente de A, utiliza-se $d(i, C)$.
  $d(i, C)$ é a média da distância do i-ésimo objeto a todos os objetos de C.
- Calcule a distância do i-ésimo objeto a todos os objetos dos outros *clusters* $C \neq A$ e escolha a menor desta, a qual é definida como $b(i)$:

$$b(i) = \min_{C \neq A} d(i, C).$$

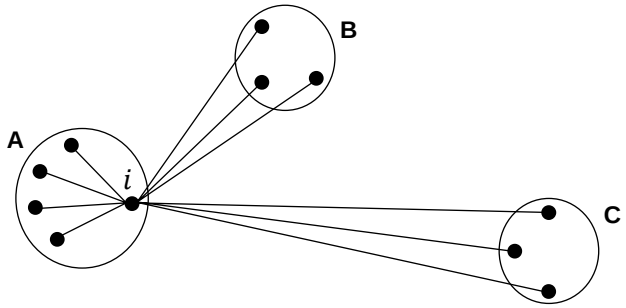


Figura 7. Elementos utilizados para calcular S(i)

Segundo a

Figura 7, o cluster B é mais proximo (em média) do objeto i, deste modo o mínimo é alcançado com $d(i,B) = b(i)$.

O número $s(i)$ é obtido pela combinação de $a(i)$ e $b(i)$, onde:

$$s(i) = \begin{cases} 1 - {a(i)}/{b(i)}, & if\ a(i) < b(i) \\ 0, & if\ a(i) = b(i) \\ {b(i)}/{a(i)} - 1, & if\ a(i) > b(i) \end{cases}$$

Propriedade do número $s(i)$: $-1 \leq s(i) \leq 1$.

Se $s(\mathrm{i})$ é próximo de 1, o objeto i está muito próximo dos objetos do seu grupo em comparação com seu vizinho. De outro lado, quando $s(i)$ é próximo de -1, o valor de $b(i)$ é muito menor do que o valor de $a(i)$. Em outras palavras, o objeto i está muito próximo de seu vizinho do que do grupo ao qual ele foi assinalado, ou seja, i está erroneamente alocado a A.

Para $s(i)$ em torno de zero, $a(i)$ e $b(i)$ são aproximadamente iguais indicando que o mesmo pode ser um objeto intermediário entre A e B.

Assim, para a construção do gráfico de silhueta, os objetos devem ser divididos em grupos de acordo com o resultado do método de agrupamento. Em cada grupo, os elementos são ordenados em ordem decrescente seguindo o valor da silhueta. Cada objeto é representado por uma barra horizontal, cujo comprimento é o valor da silhueta. Desta forma, todos os elementos são expostos em um único diagrama onde a qualidade do agrupamento pode ser analisada.

A silhueta é uma boa ferramenta para a verificação do número de clusters. A média das silhuetas $\bar{s}(k)$ é definida na equação 3.8, e pode ser usada para selecionar o melhor valor do número de grupos (k) pela escolha do valor de quando $\bar{s}(k)$ é máximo. Executa-se o método de agrupamento escolhido para todos os possíveis valores de (k) a saber, $k = \{2,3,\ldots,n-1\}$ onde se tem $max(\bar{s}(k))$ como o número de grupos para amostra. A medida $SM = \bar{s}(k)$ é um coeficiente de qualidade da perda de dimensionalidade da estrutura do agrupamento.

$$\bar{s}(k) = \sum_{i=1}^{n} s(i) \tag{3.8}$$

3.2.1.2. Índice PBM

O índice PBM, proposto por Pakhira *et al.* (2004), cujo nome é acrônimo para as iniciais dos nomes dos autores (Pakhira, Bandyopadhyay e Maulik), é um índice de variação que investiga partições avaliando sua estrutura geométrica e se os agrupamentos gerados são bem definidos e separados. É um índice de maximização, o que significa dizer que quanto maior o índice PBM calculado, melhor é a quantidade da partição gerada. O índice PBM é definido como o produto de três fatores da equação 3.9:

$$PBM(k) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K\right)^2 \tag{3.9}$$

Em que:

K é o número de clusters:

1. $E_1 = \sum_{i=1}^{K} d(\mathbf{x}(t), \mathbf{w}_0)$: dispersão em relação ao centro de todos os dados $\mathbf{w}_0$.
2. $E_K = \sum_{i=1}^{K} \sum_{t=1}^{N} x_i(t) d(x(t), \mathbf{w}_i)$: dispersão de cada cluster ao centro correspondente.
3. $D_K = max_{i,j=1\ldots K} \left(d(\mathbf{w}_i, \mathbf{w}_j)\right)$: maxima da separação entre dois clusters da partição.

4
Modelos para a gestão fazendária

Como já mencionado, um dos objetivos deste trabalho é a criação de modelos para melhorar a gestão fazendária através de previsão de receitas e na identificação de irregularidades do ISSQN, com a possibilidade dos modelos poderem ser adaptados para outros municípios.

A seguir são apresentados os três modelos propostos, que serão aplicados a quatro estudos de caso:

1. Modelo de previsão de receitas utilizando variáveis endógenas
2. Modelo de previsão de receitas utilizando variáveis endógenas e exógenas
3. Modelo de identificação de irregularidades do ISSQN, sem informação de fiscalização anterior.

Definem-se variáveis endógenas como variáveis de entrada criadas a partir da manipulação dos valores originais da série temporal para a qual se deseja realizar a previsão, por exemplo, médias móveis e diferenças. As variáveis exógenas são variáveis distantas da variável que é objeto da previsão, que podem ser agregadas ao modelo visando melhorar os resultados de previsão.

4.1.
Modelo de previsão de receitas utilizando variáveis endógenas

O modelo de previsão utilizando variáveis endógenas desenvolvido permite realizar previsões de até 12 passos à frente. As variáveis endógenas correspondem às variáveis que é objeto da previsão (Receitas Tributárias, Receitas da Dívida Ativa e ISSQN).

Nesta seção, é apresentado o modelo de previsão de receitas tributárias utilizando variáveis endógenas, composto por 3 módulos, conforme mostra a Figura 8. A seguir descrevem-se os módulos do modelo de previsão.

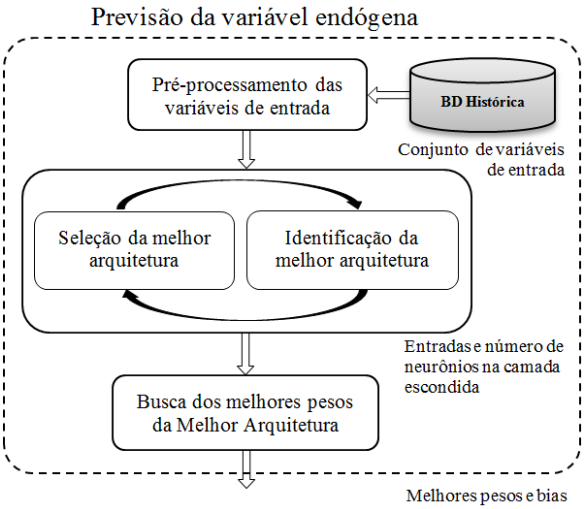


Figura 8. Modelo de previsão utilizando variáveis endógenas

4.1.1. Pré-processamento das variáveis de entrada

O processo inicia-se com o pré-processamento das séries endógenas, a seguir é realizada a construção das variáveis de entrada das redes a partir das séries originais (ver Tabela 4).

Para a construção das variáveis de entrada foram utilizados filtros. Um filtro é uma função matemática que aplicada aos valores da série produz uma nova série com características especificas. Entre os filtros mais comuns têm-se as médias móveis e diferenças (Chatfield, 2003).

Na Tabela 4 são apresentadas as 11 variáveis de entrada, onde as 6 primeiras foram criadas a partir da manipulação dos valores originais da série (como; médias móveis e diferenças). As variáveis de 7 a 11 são os retardos (*Lag's*) dos valores originais da série. O objetivo da criação de tais entradas (criadas a partir de filtros) é reduzir o número de entradas da rede, dada a possibilidade de sazonalidade e tendência crescente da série, facilitando o processamento e o mapeamento entre entradas e saídas realizado pelas redes neurais.

Tabela 4. Variáveis de entrada da rede

Núm.	Variáveis de Entradas	Processamento dos valores da série original (S)
1	MM3	$\frac{S_{t-1}+S_{t-2}+S_{t-3}}{3}$
2	MM2	$\frac{S_{t-1}+S_{t-2}}{2}$
3	MM6	$\frac{S_{t-1}+S_{t-2}+,\ldots,+S_{t-6}}{6}$
4	MM12	$\frac{S_{t-1}+S_{t-2}+,\ldots,+S_{t-12}}{12}$
5	D12	$S_t - S_{t-1}$
6	D13	$S_t - S_{t-2}$
7	$M-1$	S_{t-1}
8	$M-2$	S_{t-2}
9	$M-3$	S_{t-3}
10	$M-6$	S_{t-6}
11	$M-12$	S_{t-12}
12	$TARGET$	$S_{t=13}$

Padrões de variáveis de entrada das redes neurais – O processo de seleção das variáveis que deveriam compor o conjunto de variáveis de entrada do modelo foi avaliado exaustivamente a partir da combinação dessas variáveis. A seguir, descreve-se a forma de como essas variáveis foram combinadas.

Observa-se que o número de possíveis combinações das variáveis de entrada, sem repetição, pode ser definido pela equação 4.1.

$$C_S^n = \binom{n}{s} = \frac{n!}{s! * (n-s)!} \tag{4.1}$$

Em que:

$C_S^n =$ número de padrões de vaiáveis de entrada, dado n e s

$n =$ variáveis de entrada

$s =$ subconjunto de variáveis de entrada

Por exemplo, na Figura 9 pode-se observar um conjunto de cinco variáveis de entrada $\{1, 2, \ldots, 5\}$ (variáveis de entrada $n = 5$) com uma rede de três entradas (subconjunto de variáveis de entrada $s = 3$), uma camada escondida e

um neurônio na saída, que resultaria em 10 conjuntos possíveis de combinações de padrões:

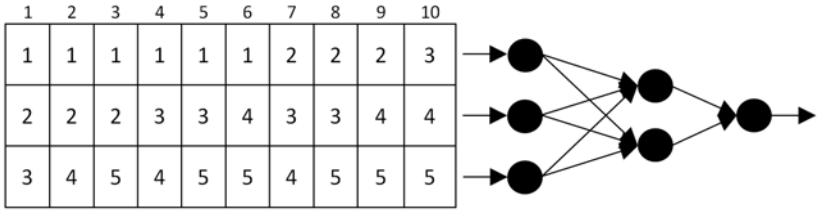


Figura 9. Combinações possíveis com 5 variáveis para uma rede neural com 3 variáveis de entradas

Na Tabela 5 é apresentado o número de padrões de variáveis, formados a partir da combinação de variáveis definidas por subconjuntos variando de 2 a 11 (subconjuntos de no mínimo 2 e no máximo 11 entradas), sem repetição de 11 variáveis.

Tabela 5. Número total de padrões de entrada a partir de combinações com 11 variáveis de entrada

Variáveis	Subconj. de Variáveis de Entrada	Combinação	Número de Padrões
11	2*	C_2^{11}	55
11	3	C_3^{11}	165
11	4	C_4^{11}	330
11	5	C_5^{11}	462
11	6	C_6^{11}	462
11	7	C_7^{11}	330
11	8	C_8^{11}	165
11	9	C_9^{11}	55
11	10	C_{10}^{11}	11
11	11	C_{11}^{11}	1
			2.036

*Combinação variando de no mínimo 2 variáveis até no máximo o total de entradas disponíveis.

Ou seja, o número total padrões (redes) é 2.036, com diferentes conjuntos de entrada a partir das 11 variáveis de entradas endógenas.

Para cada padrão de entrada são avaliadas diversas arquiteturas de redes neurais, ou seja, o número de neurônios na camada escondida varia de 1 a 6 neurônios de forma a se identificar a melhor arquitetura para cada padrão de entrada. Nesse caso, como o processo de ajuste dos pesos das redes neurais é

estocástico (os pesos iniciais são pequenos e definidos aleatoriamente), cada arquitetura das redes neurais (considerando entradas distintas e o número de neurônios na camada escondida) será inicializada 100 vezes. Considerando o número de padrões de entradas (2.036) e o número de neurônios na camada escondida (até 6 neurônios), o número máximo de redes distintas que poderiam ser avaliadas seria 1.221.600 redes neurais (linha 2 da Tabela 6).

A partir dessa metodologia apresentada, a Tabela 6 apresenta os totais de redes a serem avaliadas para diferentes números das variáveis de entradas. Por exemplo, caso existissem apenas 7 variáveis de entradas, seriam avaliados 72.000 modelos de redes neurais (linha 6 da Tabela 6), considerando a combinação das entradas em subconjuntos de 2 a 7 variáveis, número de neurônios na camada escondida variando de 1 a 6 e 100 inicializações de pesos para cada uma dessas configurações.

Tabela 6. Número de modelos com diferentes números de entradas disponíveis

Entradas Disponíveis	Num. Neurônios na Camada Escondida	Inicializações	Total de Combinações	Número de Redes
11	6	100	2.036	1.221.600
10	6	100	1.013	607.800
9	6	100	502	301.200
8	6	100	247	148.200
7	6	100	120	72.000
6	6	100	57	34.200
5	6	100	26	15.600
4	6	100	11	6.600
3	6	100	4	2.400
2	6	100	1	600

Normalização – A normalização das séries é realizada entre $[-1,1]$. O valor -1 corresponde aos valores mínimos e o valor 1 aos valores máximos da série. Para a normalização inicialmente considerou-se a equação 4.2:

$$Y = \frac{(ymax - ymin) * (x - xmin)}{(xmax - xmin)} + ymin \qquad (4.2)$$

Em que:

$xmax =$ máximo valor da série

$xmin =$ mínimo valor da série

$x =$ valor a normalizar

$ymax = 1$

$ymin = -1$

$Y =$ Normalizado

Para evitar que os novos valores previstos ultrapassem os limites da normalização [-1,1], foi predeterminado uma extensão do domínio da normalização, prevenindo o futuro crescimento ou decrescimento dos valores estimados. O novo domínio da série foi calculado com a equação 4.3:

$$Y' = \frac{(ymax - ymin) * (x - xmin * 0.82)}{(xmax * 1.18 - xmin * 0.82)} + ymin \qquad (4.3)$$

Modelagem das redes neurais – A arquitetura e as características de uma rede neural *Multi-Layer Perceptron* (MLP), utilizada neste estudo, compreendem o número de entradas, o número de camadas escondidas e o número de neurônios na camada escondida.

Os estudos teóricos mostram que uma camada escondida é suficiente para que essas redes aproximem quaisquer funções não-lineares (Cybenko, 1989).

Segundo alguns autores como Danh *et. al* (1999), Kisi (2005) e Fayal (2005), há alguns detalhes ainda não resolvidos, em relação a RNA. Não há, ainda, um método sistemático ou teoria para a determinação da arquitetura ótima (número de entradas, número de camadas escondidas, número de neurônios na camada escondida). A geometria varia muito de acordo com o problema em questão, e o método da tentativa e erro tem sido usado para alcançar uma estrutura aceitável.

Nesta dissertação o método para buscar a melhor arquitetura de rede (número de neurônios na camada escondida e peso dos neurônios), dentre os testes realizados, é realizado em dois passos: O primeiro passo é encontrar a melhor arquitetura em média (100 inicializações), onde serão avaliadas as arquiteturas de redes com diferentes padrões de entrada e número de neurônios na camada escondida. O segundo passo é buscar o melhor conjunto de pesos e bias, para lograr o objetivo as melhores arquiteturas em média encontrada no passo anterior, serão inicializadas 1000 vezes.

Treinamento das redes MLP – Os seguintes parâmetros serão utilizados para o treinamento das redes:

- Foi utilizada a função *tansing* (*Hyperbolic Tangent Sigmoid*) como função de ativação da camada intermediaria e na camada de saída foi utilizada a função *purelin (Linear)*.
- Para o treinamento foi utilizado o método *Levenberg-Marquardt.*
- Para determinar a melhor arquitetura, foi utilizado o MAPE *(Mean absolute Percentage Error)* é calculado de acordo com a seguinte fórmula:

$$MAPE = \frac{\sum_{k=1}^{N} \left| \frac{a_k - y_k}{a_k} \right|}{N} \times 100\% \qquad (4.4)$$

  Em que:
  N = Número de padrões da base de dados
  a_k = Saída desejada para o k-ésimo padrão
  y_k = Saída obtida para o k-ésimo padrão

- Número máximo de épocas igual a 1000.
- Número máximo de falhas no conjunto de validação igual a 25.
- Uma camada escondida (oculta) com o número de neurônios variando de 1 a 6.
- Seleção de atributos através da avaliação dos conjuntos de variáveis de entrada, que são formados a partir da combinação das variáveis de entrada desde 2 elementos até no máximo 7 elementos. À medida que o número de entradas aumenta, aumenta o erro de previsão. Por esse motivo o número de entradas foi limitado a 7.
- Tamanho da janela histórica é de 12 meses.

4.1.2. Seleção de variáveis endógenas

Para a seleção de variáveis de entrada serão utilizados os métodos de seleção de variáveis PCAM, LSE, SIE e ReliefF. É importante salientar que os métodos de seleção de variáveis apresentam apenas, uma noção (a eficiência depende da adequabilidade do método aos dados propriamente ditos) de quais variáveis têm mais influencia na previsão (Fayal, 2005).

A seguir se descrevem os procedimentos da seleção de variáveis:

A. Selecionar as variáveis de entrada, utilizando os métodos PCAM, LSE, SIE e ReliefF. Após a seleção avaliar a classificação de cada método e escolher os métodos segundo o seu desempenho.

B. Associar os métodos escolhidos (item A, acima), onde o critério da escolha é a classificação dos métodos, dando prioridade às entradas mais relevantes de cada método, a fim de se obter uma nova configuração de entradas.

4.1.3. Identificação da melhor arquitetura em média

Nesta dissertação para identificar as melhores arquiteturas de rede em média (número de neurônios na camada escondida e peso dos neurônios), será realizada a previsão *multistep* (até 12 passos à frente) das redes selecionadas na seção anterior (4.1.2). Onde cada arquitetura será inicializada com 100 repetições, e avaliadas com diferentes padrões de entradas e número de neurônios na camada escondida.

4.1.4. Busca dos melhores pesos

Nesta seção o objetivo é encontrar o melhor conjunto de pesos e bias para as melhores arquiteturas de rede em média obtida na seção anterior (4.1.3). Para alcançar este objetivo cada rede neural será inicializada 1000 vezes. Após de encontrar o melhore conjunto de pesos e bias será realizado o Teste (Recall), onde serão utilizados os dados da série original do ano de validação e teste respectivamente.

4.2.
Modelo de previsão de receitas utilizando variáveis endógenas e exógenas

O modelo de previsão utilizando variáveis endógenas e exógenas desenvolvido permite realizar previsões de até 12 passos à frente. As variáveis endógenas correspondem às variáveis que são objeto da previsão (por exemplo, o ISSQN). As variáveis exógenas são criadas a partir de dados econômico-financeiros que ajudam a explicar a variável que se deseja prever.

Nesta seção, é apresentado o modelo de previsão de receitas utilizando variáveis endógenas e exógenas, composto por 7 módulos, conforme mostra a Figura 10. A seguir descrevem-se os módulos do modelo de previsão.

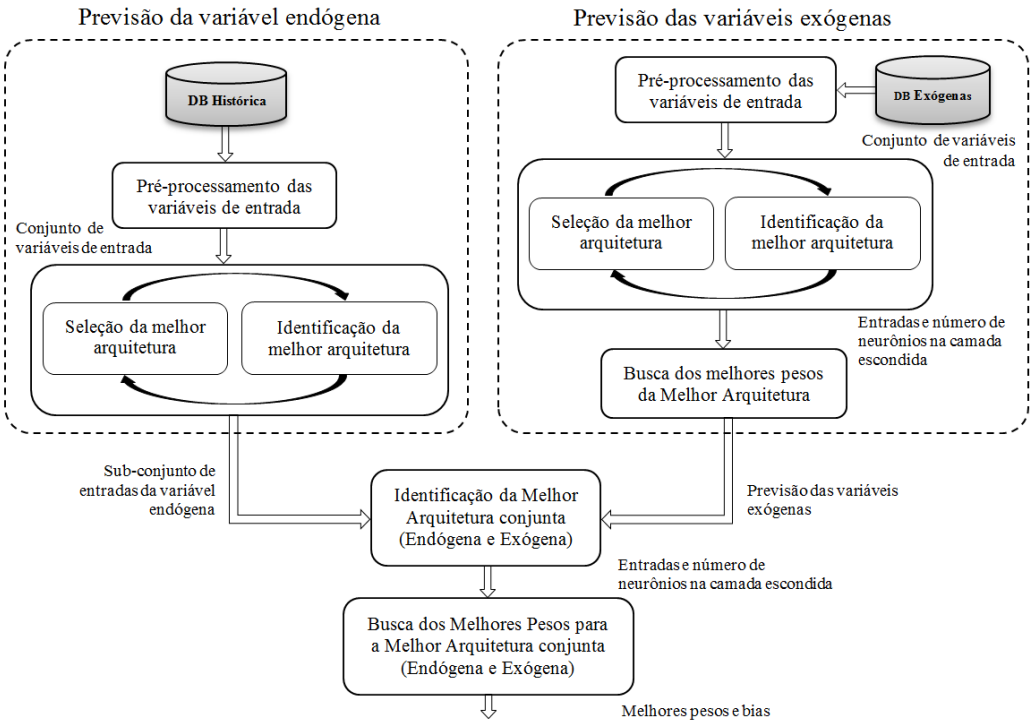


Figura 10. Modelo de previsão utilizando variáveis endógenas e exógenas

4.2.1.
Previsão de variáveis exógenas por redes neurais

Quando o modelo de previsão da variável de interesse usa variáveis exógenas em sincronia com a variável que se deseja prever, é necessário primeiro realizar a previsão para as variáveis exógenas. Por exemplo, na Figura 11, para se prever a variável de saída Z no tempo $t+1$ (futuro), a partir de variáveis exógenas A, B e C sincronizadas à variável Z, é necessário conhecer

os valores dessas variáveis em $t + 1$. A partir desses valores previstos é possível realizar a previsão da variável de interesse.

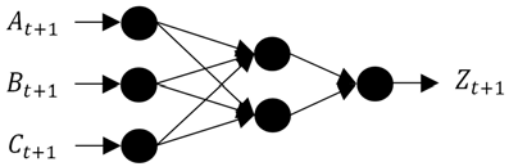


Figura 11. Valores futuros (previstos) das variáveis exógenas

Em métodos de previsão, a inclusão de variáveis exógenas relevantes no modelo pode aumentar a acurácia da previsão. Assim, devem-se investigar as possíveis variáveis exógenas que poderiam contribuir para melhorar a previsão da variável que se deseja prever. No entanto, não se deve simplesmente agregar tais variáveis. Deve-se primeiro avaliar o melhor conjunto de variáveis exógenas que pode agregar valor ao modelo de previsão.

A seguir é realizado o processo de seleção as variáveis exógenas, que podem ser relevantes no modelo.

4.2.1.1. Pré-processamento das variáveis exógenas

Nesta seção são extraídas as séries das variáveis exógenas de bases de dados externas. A seguir as variáveis exógenas são escolhidas através da correlação cruzada da variável que se deseja prever (ISSQN), com cada variável exógena utilizando o software Econométrico EViews (Stock e Watson, 2007). Neste procedimento são selecionadas as variáveis exógenas com correlações significativas.

A seguir são construídas as onze (11) variáveis de entradas (MM3, MM2, MM6, MM12, D12, D13, M-1, M-2, M3, M-6 e M-12) das variáveis exógenas selecionadas, como indicadas na Tabela 4.

4.2.1.2. Seleção das variáveis exógenas

Para a seleção de variáveis de entrada exógenas serão utilizados os métodos de seleção de variáveis PCAM, LSE, SIE e ReliefF. É importante salientar que os métodos de seleção de variáveis apresentam apenas, uma noção (a eficiência depende da adequabilidade do método aos dados

propriamente ditos) de quais variáveis têm mais influencia na previsão (Fayal, 2005).

A seguir se descrevem os procedimentos para a seleção das variáveis exógenas:

A. Selecionar as variáveis de entradas exógenas, utilizando os métodos PCAM, LSE, SIE e ReliefF. Após a seleção avaliar a classificação de cada método de seleção de variáveis e escolher o(s) método(s) segundo o seu desempenho.

B. Associar os métodos escolhidos (item A, acima), onde o critério da escolha é a classificação dos métodos, dando prioridade às entradas mais relevantes de cada método, a fim de se obter uma nova configuração de entradas.

4.2.1.3. Identificação das melhores arquiteturas (em média) das variáveis exógenas

Nesta dissertação para identificar as melhores arquiteturas de rede em média (número de neurônios na camada escondida e peso dos neurônios), será realizada a previsão *multistep* (até 12 passos à frente) das redes selecionadas na seção anterior (4.2.1.2). Onde cada arquitetura será inicializada com 100 repetições, e avaliadas com diferentes padrões de entradas e número de neurônios na camada escondida.

4.2.1.4. Busca dos melhores pesos das variáveis exógenas

Nesta seção o objetivo é encontrar o melhor conjunto de pesos e bias para as melhores arquiteturas de rede em média obtida na seção anterior (4.2.1.3). Para alcançar este objetivo cada rede neural será inicializada 1000 vezes.

Após de realizar as previsões das variáveis exógenas do ano de validação é realizado a previsão conjunta (variáveis endógenas e exógenas).

4.2.2. Identificação da melhor arquitetura conjunta (endógenas e exógenas) em média

Nesta seção, o objetivo é considerar a utilização das variáveis exógenas para melhorar (reduzir o MAPE) a qualidade da previsão. Assim serão

investigados variáveis exógenas econômico-financeiras (relacionadas ao ISSQN) que podem contribuir para aumentar a eficiência do modelo de estimação de valores futuros de arrecadação do ISSQN.

Nesta dissertação para identificar a melhor arquitetura de rede conjunta (endógena e exógena) em média (número de neurônios na camada escondida e peso dos neurônios), será realizada a previsão *multistep* (até 12 passos à frente) das redes selecionadas nas seções anteriores (4.1.3 e 5.3.2.4). Onde cada arquitetura será inicializada com 100 repetições, e avaliadas com diferentes padrões de entradas e número de neurônios na camada escondida.

4.2.2.1. Avaliação do esforço computacional utilizando variáveis conjuntas (endógenas e exógenas)

O teste com variáveis endógenas e exógenas aumenta o número de entradas da rede, o que leva também ao aumento de tempo de processamento computacional. Para se ter uma noção da quantidade de modelos a serem avaliados, a Tabela 7, apresenta os números totais de arquiteturas de redes a serem avaliadas caso fossem, 5 variáveis exógenas (Tabela 7) e 11 entradas endógenas (Tabela 8).

Primeiro é calculado o número de padrão de variáveis exógenas, formados a partir da combinação de variáveis definidas por subconjuntos variando de 0 a 5 (subconjunto com nenhuma exógena e no máximo 5 entradas), sem repetição de 5 variáveis. Para calcular o número total de combinações das vaiáveis endógenas e exógenas, foi utilizado a eq. (4.1).

Tabela 7. Número de combinações possíveis no caso de se considerar 5 variáveis exógenas

Entradas Exógenas	Sub Conj.	Combinação	Número Total de Redes
5	0*	C_0^5	1
5	1	C_1^5	5
5	2	C_2^5	10
5	3	C_3^5	10
5	4	C_4^5	5
5	5	C_5^5	1
			32

* Não será utilizada nenhuma variável exógena (rede sem variáveis exógenas)

Na Tabela 8, é apresentado o número de combinações de arquiteturas de redes com 11 entradas endógenas e subconjuntos variando de 1 a 11.

Tabela 8. Número de combinações com 11 entradas endógenas

Entradas Endógenas	Sub Conj.	Combinação	Número Total de Redes
11	1	C_1^{11}	11
11	2	C_2^{11}	55
11	3	C_3^{11}	165
11	4	C_4^{11}	330
11	5	C_5^{11}	462
11	6	C_6^{11}	462
11	7	C_7^{11}	330
11	8	C_8^{11}	165
11	9	C_9^{11}	55
11	10	C_{10}^{11}	11
11	11	C_{11}^{11}	1

Na Tabela 9, é apresentado o número total de arquiteturas de redes considerando 11 variáveis endógenas (coluna 1) e 5 exógenas (coluna 2) combinadas, número de neurônios na camada escondida (coluna 3), número de inicializações dos neurônios (coluna 4). Os números definidos na coluna 1 indicam as combinações resultantes das variáveis endógenas (combinação de 1 a 11 variáveis). Os números da coluna 2 indicam as combinações resultantes das variáveis exógenas (combinadas de 0 a 5 variáveis). A coluna 3 indica os números de neurônios na camada escondida avaliados no modelo, variando de 1 até 6. E finalmente a quarta coluna apresenta os números de inicializações que cada uma dessas arquiteturas (entradas e número de neurônios na camada escondida) precisa ser avaliada isoladamente. Assim a Tabela 9, mostra o número total de redes com 11 entradas endógenas e 5 variáveis exógenas.

Tabela 9. Número de redes com 11 endógenas e 5 exógenas

Entradas Endógenas	Entradas Exógenas	Neurônios	Inicializações	Número Total de Redes
11	31*	6	100	204.600
55	32	6	100	1.056.000
165	32	6	100	3.168.000
330	32	6	100	6.336.000
462	32	6	100	8.870.400
462	32	6	100	8.870.400
330	32	6	100	6.336.000
165	32	6	100	3.168.000
55	32	6	100	1.056.000
11	32	6	100	211.200
1	32	6	100	19.200
				39.295.800

***Foram contabilizadas 31 variáveis, para o caso de não haver variável exógena.

Na Tabela 10 é apresentado o número total de redes considerando entradas endógenas de 2 a 11(coluna 1) e 5 entradas exógenas (coluna 2). Os números na coluna 1 indicam as entradas endógenas (definido pelos métodos de seleção de variáveis), e os números da coluna 2 indica o número de entradas exógenas (combinadas de 0 a 5 variáveis).

Por exemplo, o número total de redes com 11 entradas endógenas e 5 entradas exógenas é 39.295.800 (Tabela 10), esse valor corresponde ao número total de redes calculado na Tabela 9.

Tabela 10. Número de redes com entradas endógenas variando de 2 a 11 e 5 variáveis exógenas

Entradas Endógenas	Entradas Exógenas	Número Total de Redes
11	5	39.295.800
10	5	19.635.600
9	5	9.805.800
8	5	4.891.200
7	5	2.434.200
6	5	1.206.000
5	5	592.200
4	5	285.600
3	5	132.600
2	5	56.400

4.2.2.2.
Busca do melhor conjunto de pesos da melhor arquitetura conjunta (endógena e exógena)

Nesta seção o objetivo é encontrar o melhor conjunto de pesos e bias para a melhor arquitetura de rede conjunta (endógena e exógena) em média obtida na seção anterior (4.2.2). Para alcançar este objetivo cada rede neural será inicializada 1000 vezes. Após de encontrar o melhor conjunto de pesos e bias será realizado o Teste (Recall), onde serão utilizados os dados da série original do ano de validação e teste respectivamente.

4.3.
Modelo de Identificação de Irregularidades do ISSQN

O modelo proposto permite identificar divergências na classificação de contribuintes do ISSQN em suspeitos e não suspeitos de irregularidades, utilizando técnicas de Mineração de Dados (Data Mining), a partir do valor da emissão do ISSQN dos contribuintes e informações exógenas extraídas de bases de dados do CAGED (Cadastro Geral de Empregados e Desempregados).

O modelo proposto poderia utilizar outras bases de dados, por exemplo, as que incluem informações como: salário dos empregados, consumo de energia elétrica, escrituração fiscal, etc. Essas informações tornariam ainda mais robusta a modelagem. No entanto, devido a entraves que envolvem o sigilo fiscal não foi possível se obter estas outras informações. Além disso, infelizmente não se obteve informações reais para se avaliar essa metodologia, ficando apenas como uma metodologia proposta para futuras análises.

A base do ISSQN é formada pelo identificador do contribuinte (CNPJ) e pelo valor do ISSQN emitidos por mês. A base de dados do CAGED também é formada pelo cadastro do contribuinte (CNPJ) e pelo número de empregados por mês.

O modelo proposto utiliza algoritmo de clusterização Fuzzy C-Means para agrupar contribuintes por ano, a partir de informações do valor do ISSQN emitido e do número de empregados. Após a clusterização do valor emitido e da clusterização segundo o número de empregados, passa-se à fase de análise dos contribuintes nos grupos através de um processo heurístico, que será descrito na seção 4.3.3 em maiores detalhes. O resultado da análise atribui a cada contribuinte avaliado, um grau (0 a 100) de irregularidade. Devido aos problemas

encontrados nas bases de dados utilizados, a granularidade temporal deste modelo é anual, ao invés de ser mensal como poderia ser.

A Figura 12 mostra o diagrama de blocos do modelo, e a seguir descreve-se todo o processo de identificação das irregularidades, a partir das divergências na classificação dos contribuintes nos grupos criados.

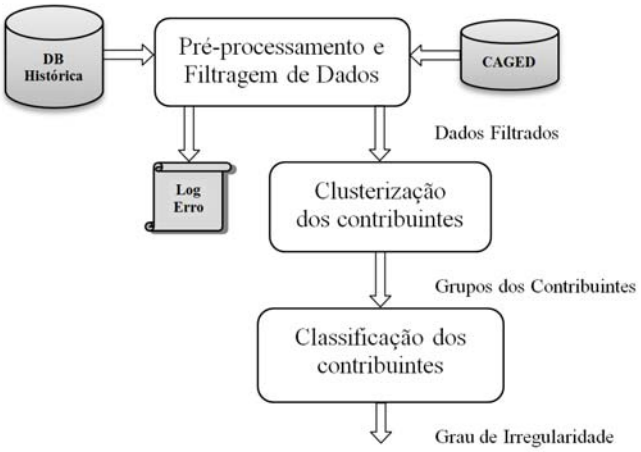


Figura 12. Modelo de Identificação de Irregularidades

4.3.1. Pré-Processamento

O processo inicia-se com a leitura dos registros das bases de dados de contribuintes do ISSQN e do CAGED. O passo seguinte é a realização do pré-processamento dos dados e aplicação de um conjunto de filtros como:

- Cruzamento das bases de dados do Município e do CAGED, para selecionar registros de contribuintes que estejam presentes em ambas as bases de dados;
- Selecionar registros de contribuintes que estejam presentes no ano de referência T e $T+1$ (ano de comparação, ou seja, ano que se deseja verificar a existência de irregularidade);
- Para cada contribuinte somar todos os valores do atributo de Emissão do ISSQN do ano T, repetir a operação para o ano $T+1$;
- Para cada contribuinte calcular a média do atributo Número de Empregados do ano T, repetir a operação para o ano $T+1$;
- Utilizar a heurística da Tabela 11, para filtrar os registros dos anos T e $T+1$.

Tabela 11. Heurísticas para filtrar dados

	Emissão do ISSQN		**Núm. de Empregados**		
	$\boldsymbol{T}$	$\boldsymbol{T+1}$	$\boldsymbol{T}$	$\boldsymbol{T+1}$	**Registro**
1	ISS = 0	ISS = 0	Emp. = 0	Emp. = 0	Ignorar
2	ISS = 0	ISS < 0	Emp. = 0	Emp.= 0	Ignorar
3	ISS = 0	ISS < 0	Emp. = 0	Emp.< 0	Ignorar
4	ISS < 0	ISS = 0	Emp. = 0	Emp.= 0	Ignorar
5	ISS < 0	ISS = 0	Emp. < 0	Emp. = 0	Ignorar

A Tabela 11 apresenta nas colunas 2 e 3 de valores de ISSQN emitidos por ano (T e $T+1$) para cada contribuite. As colunas 4 e 5 mostram, respectivamente, o número de empregados no ano T e $T+1$ para cada contribuinte.

Os contribuintes classificados pela heurística como ignorados, foram removidos da base de dados. O objetivo da heurística é filtrar casos de contribuintes que não se encaixam no padrão de normalidade ou de irregularidade, por exemplo, empresas novas ou que, provavelmente, foram fechadas. Todos os outros casos serão mantidos na base de dados.

Na Tabela 12 apresenta-se um exemplo de registros pré-filtrados e processados do ano $T+1$, onde N representa a quantidade de contribuintes.

A coluna 1 da Tabela 12 apresenta o CNPJ dos contribuintes, a coluna 2 e a coluna 3 apresentam, respectivamente, o somatório do valor da Emissão do ISSQN por ano e o Número médio de Empregados por ano.

Tabela 12. Registros dos contribuintes do ano *T+1*

CNPJ	**Emissão do ISSQN**	**Núm. de Empregados**
1	$EM(T+1)_{11}$	$NE(T+1)_{12}$
2	$\mathrm{EM}(T+1)_{21}$	$NE(T+1)_{22}$
⋮	⋮	⋮
N	$EM(T+1)_{N1}$	$NE(T+1)_{N2}$

Nesse trabalho serão utilizados dois métodos de valiação de clusters. O primeiro foi o método Silhueta (Rousseeuw, 1987), (Vendramin *et a*l 2009). Neste método é gerado um índice que pode ser traduzido como: um coeficiente que mede o quão bem alocado cada elemento está no seu grupo quando

comparado aos outros clusters formados. O segundo método utilizado foi o PBM desenvolvido por Pakhira *et al.,* (2004), o PBM é um índice de validação que investiga partições avaliando sua estrutura geométrica e, se os agrupamentos gerados são bem definidos e separados. Esses métodos visam identificar o número ideal (k) de clusters para os dados avaliados.

A seguir é apresentado o processo de clusterização.

4.3.2. Clusterização dos contribuintes

Esta seção apresenta o processo de clusterização utilizando *Fuzzy C-Means* (FCM), onde os dados obtidos no passo anterior são utilizados para construir grupos *(clusters)* de contribuintes que apresentam comportamentos similares.

Após a realização da clusterização por atributo, têm-se os grupos dos contribuintes para cada atributo (nesse caso emissão do ISSQN e número de Empregados) nos anos T e $T+1$ (ano de avaliação da irregularidade, ou seja, ano que se deseja verificar a existência de irregularidade), e os respectivos centróides. Os centróides dos grupos, para o atributo Emissão do ISSQN do ano $T+1$, podem ser definidos como os apresentados na Tabela 13.

Tabela 13. Centróides do atributo Emissão do ISSQN do ano *T+1*

Grupo	**Centróides de Emissão do ISSQN**
1	$CEISSQN(T+1)_1$
2	$CEISSQN(T+1)_2$
⋮	⋮
g	$CEISSQN(T+1)_g$

Da mesma forma, são criados grupos para o atributo Número de Empregados, com seus respectivos centroides. Esse procedimento deve ser repetido para o ano T.

4.3.3.
Classificação dos contribuintes

A seguir é realizado o processo de classificação dos contribuintes do ISSQN utilizando a equação 4.5, segundo um grau de irregularidade, em suspeitos e não suspeitos de irregularidades, utilizando os grupos obtidos no passo anterior para os anos T e $T+1$.

$$Grau\ de\ Irregularidade = \frac{CEISSQN\ A_i(T)}{CEISSQN\ A_j(T+1)} + \frac{CNEmp\ A_u(T+1)}{CNEmp\ A_v(T)} \qquad (4.5)$$

Em que:

- $CEISSQN\ A_i(T)$ é o centróide do contribuinte A classificado no grupo i quanto a emissão de ISSQN no ano T;
- $CEISSQN\ A_j(T+1)$ é o centróide do contribuinte A classificado no grupo j quanto a emissão de ISSQN no ano $T+1$;
- $CNEmp\ A_u(T+1)$ é o centróide do contribuinte A classificado no grupo u quanto ao número de empregados no ano $T+1$;
- $CNEmp\ A_v(T)$ é o centróide do contribuinte A classificado no grupo v quanto ao número de empregados no ano T.

O objetivo dessa função heurística é avaliar o grau de modificação dos padrões, seja quanto à emissão do ISSQN e/ou quanto ao número de empregados do contribuinte em dois anos consecutivos. O padrão de irregularidade clássico relaciona a redução do ISSQN com o aumento do número de empregados, ou seja, sob o ponto de vista da irregularidade essas duas variáveis são inversamente proporcionais.

A explicação para usar o valor do centróide, ao invés do valor da emissão do contribuinte, por exemplo, se dá pelo fato do centróide representar o valor do grupo, a despeito de pequenas variações das emissões de ISSQN ou número de empregados dos contribuintes que pertençam ao grupo. Essa prática proporciona maior robustez ou considera as incertezas relacionadas às diversas possibilidades que cercam as oportunidades para os negócios dos prestadores de serviços.

Inicialmente considerou-se uma função na qual, para cada contribuinte se observava o centróide do grupo ao qual ele pertence para os atributos emissão de ISSQN e número de empregados em T e $T+1$. Dessa forma percebeu-se, que muitas vezes, de um ano para o outro se verificou uma redução ou crescimento financeiro do setor ao qual o contribuinte está associado, sem com

isso significar, prática de irregularidade. Para eliminar essa componente, ao invés de se utilizar o valor do centroide do grupo ao qual pertencia o contribuinte do ano anterior (T), passou-se a observar o número do grupo ao qual pertencia o contribuinte em T, mas utilizar o valor do centróide do grupo correspondente em $T+1$.

Assim, para efeitos de contabilidade o Grau de Irregularidade é calculado utilizando a eq. (4.6), onde a função passa a efetivamente considerar os seguintes valores:

$$Grau\ de\ Irregularidade = \frac{CEISSQN\ A_{i(T)}(T+1)}{CEISSQN\ A_j(T+1)} + \frac{CNEmp\ A_u(T+1)}{CNEmp\ A_{v(T)}(T+1)} \qquad (4.6)$$

Em que:

- $CEISSQN\ A_{i(T)}(T+1)$ é o centróide do contribuinte A classificado no grupo i em $T+1$ quanto a emissão de ISSQN, dado que em T o contribuinte estava classificado no grupo i;
- $CEISSQN\ A_j(T+1)$ é o centróide do contribuinte A classificado no grupo j quanto a emissão de ISSQN no ano T+1;
- $CNEmp\ A_u(T+1)$ é o centróide do contribuinte A classificado no grupo u quanto ao número de empregados no ano T+1.
- $CNEmp\ A_{v(T)}(T+1)$ é o centróide do contribuinte A classificado no grupo v quanto ao número de empregados no ano T+1, dado que em T o contribuinte estava classificado no grupo v.

Com o objetivo de conhecer todas as relações entre os centróides de dois grupos quaisquer sejam para o valor emitido de ISSQN, ou para o número de empregados, calcula-se a matriz de relações R_1 (Emissão do ISSQN) e R_2 (Número de Empregados), para o ano $T+1$.

Para realizar o cálculo da matriz R_1 (atributo Emissão do ISSQN), serão utilizados os centróides do ano T+1. Esta matriz R_1, apresentada a seguir, tem dimensão $g \times g$:

$$R_1 = \begin{bmatrix} \frac{CEISSQN_1(T+1)}{CEISSQN_1(T+1)} & \frac{CEISSQN_1(T+1)}{CEISSQN_2(T+1)} & \cdots & \frac{CEISSQN_1(T+1)}{CEISSQN_g(T+1)} \\ \frac{CEISSQN_2(T+1)}{CEISSQN_1(T+1)} & \frac{CEISSQN_2(T+1)}{CEISSQN_2(T+1)} & \cdots & \frac{CEISSQN_2(T+1)}{CEISSQN_g(T+1)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{CEISSQN_g(T+1)}{CEISSQN_1(T+1)} & \frac{CEISSQN_g(T+1)}{CEISSQN_2(T+1)} & \cdots & \frac{CEISSQN_g(T+1)}{CEISSQN_g(T+1)} \end{bmatrix}_{g \times g}$$

Para realizar o cálculo da matriz R_2 do atributo Número de Empregados, foram utilizados os centróides do ano $T+1$, esta matriz também tem dimensão $g \times g$. O cálculo da matriz R_2 é apresentado a seguir:

$$R_2 = \begin{bmatrix} \frac{CNEmp_1(T+1)}{CNEmp_1(T+1)} & \frac{CNEmp_1(T+1)}{CNEmp_2(T+1)} & \cdots & \frac{CNEmp_1(T+1)}{CNEmp_g(T+1)} \\ \frac{CNEmp_2(T+1)}{CNEmp_1(T+1)} & \frac{CNEmp_2(T+1)}{CNEmp_2(T+1)} & \cdots & \frac{CNEmp_2(T+1)}{CNEmp_g(T+1)} \\ \vdots & \vdots & & \ddots \ \vdots \\ \frac{CNEmp_g(T+1)}{CNEmp_1(T+1)} & \frac{CNEmp_g(T+1)}{CNEmp_2(T+1)} & \cdots & \frac{CNEmp_g(T+1)}{CNEmp_g(T+1)} \end{bmatrix}_{g \times g}$$

Em seguida é apresentada a matriz de variação μ, onde cada elemento da matriz R_1 é combinado a cada elemento da matriz R_2, através de uma soma desses elementos, gerando a matriz μ, com todas as possíveis combinações entre as relações dos centróides dos grupos formados pelas emissões de ISSQN e pelo número de empregados, de dimensão $g^2 \times g^2$:

$$\mu = \begin{bmatrix} \frac{CEISSQN_1(T+1)}{CEISSQN_1(T+1)} + R_2 & \frac{CEISSQN_1(T+1)}{CEISSQN_2(T+1)} + R_2 & \cdots & \frac{CEISSQN_1(T+1)}{CEISSQN_g(T+1)} + R_2 \\ \frac{CEISSQN_2(T+1)}{CEISSQN_1(T+1)} + R_2 & \frac{CNEmp_2(T+1)}{CNEmp_2(T+1)} + R_2 & \cdots & \frac{CEISSQN_2(T+1)}{CEISSQN_g(T+1)} + R_2 \\ \vdots & \vdots & & \ddots \ \vdots \\ \frac{CEISSQN_g(T+1)}{CEISSQN_1(T+1)} + R_2 & \frac{CEISSQN_g(T+1)}{CEISSQN_2(T+1)} + R_2 & \cdots & \frac{CEISSQN_g(T+1)}{CEISSQN_g(T+1)} + R_2 \end{bmatrix}_{g^2 \times g^2}$$

O objetivo em construir antecipadamente a matriz μ é tornar a avaliação uniforme, a partir da normalização dos seus elementos utilizando eq. (4.2), onde os valores $ymax$ e $ymin$ foram modificados para 100 e 0, respectivamente. Dessa forma o intervalo dos valores da variação da matriz μ está entre 0 e 100.

Para obter o Grau de Irregularidade de cada contribuinte, este será determinado segundo as suas emissões de ISSQN e o número de empregados em T e $T+1$. Por exemplo, para um valor de ISSQN emitido pelo contribuinte A em T, é identificado o centróide mais próximo a esse valor, dentre os centróides dos grupos formados pela clusterização dos valores de ISSQN em T, considerando todos os contribuintes em T.

A seguir na Tabela 14 é apresentados um exemplo de agrupamento de valor do ISSQN emitidos e número de empregados de 12 contribuintes, para os anos T e $T+1$. Na coluna 1 apresenta-se o CNPJ dos contribuintes.

No ano T, a coluna 2 são os valores ISSQN emitidos por ano e na coluna 3 o grupo ao qual pertence o valor ISSQN emitido. Na coluna 4 apresenta a média

do número de empregados por ano e na coluna 5 o grupo ao qual pertence o número de empregados.

No ano $T+1$, a coluna 6 são os valores ISSQN emitidos por ano e na coluna 7 o grupo ao qual pertence o valor ISSQN emitido. A coluna 8 apresenta a média do número de empregados por ano e na coluna 9 o grupo ao qual pertence o número de empregados.

Tabela 14. Contribuintes agrupados nos anos *T* e *T+1*

	T				$T+1$			
CNPJ	**ISSQN**	**Grupo**	**Num. Emp.**	**Grupo**	**ISSQN**	**Grupo**	**Num. Emp.**	**Grupo**
1	48.777,76	6	36	6	0,00	1	46	6
2	48.156,59	6	0	1	984,72	1	0	1
3	21.706,66	5	8	3	0,00	1	8	3
4	6.937,85	3	12	3	0,00	1	20	4
5	8.779,86	3	4	2	224,94	1	4	2
6	6.827,35	3	2	1	0,00	1	2	1
7	4.999,57	3	13	4	0,00	1	10	3
8	4.984,69	3	7	3	0,00	1	2	1
9	3.341,73	2	1	1	807,47	1	3	2
10	1.515,50	2	1	1	1.004,92	1	3	2
11	2.008,58	2	0	1	761,00	1	2	2
12	586,22	1	0	1	2.331,85	2	10	3

Da Tabela 14, pode-se observar o contribuinte com CNPJ 9, no ano T o valor ISSQN emitido foi de 3.341,73 que pertence ao grupo 2, e a média do número de empregados foi de 1 o qual pertence ao grupo 1. O mesmo contribuinte no ano $T+1$, o valor ISSQN emitido foi de 807,47 que pertence ao grupo 1, e a média do número de empregados foi de 3 o qual pertence ao grupo 2.

O grau de irregularidade do contribuinte com CNPJ 9 é significativo, por que no ano T seu valor ISSQN emitido foi de 3.341,73 e o mesmo contribuinte no ano $T+1$ seu valor ISSQN emitido foi de 807,47. Neste caso o valor ISSQN emitido diminuiu em aproximadamente ¼, mas o número de empregados aumento de 1 para 3.

5
Estudos de caso

Neste capítulo serão aplicados os três modelos propostos nesta dissertação, para obter resultados decorrentes dos quatro (4) estudos de casos para o Município de Araruama. A escolha do município foi realizada pela Secretaria de Planejamento e Gestão (SEPLAG) em parceria com a Associação Estadual de Municípios (AMERJ) do projeto de Municípios Eficientes.

Os quatro (4) estudos de caso desenvolovidos foram as seguintes:

1. Previsão da arrecadação mensal de Receitas da Dívida Ativa.
2. Previsão da arrecadação mensal de Receitas Tributárias.
3. Previsão da arrecadação mensal do ISSQN, utilizando variáveis endógenas e exógenas.
4. Identificação de Irregularidades do ISSQN.

5.1.
Previsão da arrecadação mensal de receitas da Dívida Ativa

Além dos Impostos, as Receitas da Dívida Ativa também influencia na composição do orçamento do Município (Tabela 15). Assim para se realizar uma boa gestão, são necessárias ferramentas e funcionalidades, para a previsão de valores futuros da arrecadação de Receitas que estão em Dívida Ativa. Dessa forma neste estudo de caso será realizada a previsão de 12 passos à frente da arrecadação mensal de Receitas da Dívida Ativa do Município de Araruama utilizando variáveis endógenas, ou seja, variáveis criadas a partir da variável que se deseja prever, nesse caso a Dívida Ativa. Para a previsão será aplicado o modelo proposto na Figura 8.

Na Tabela 15 são apresentadas as arrecadações anuais dos anos 2004 a 2010 de Receitas da Dívida Ativa do Município de Araruama.

Tabela 15. Arrecadação anual de Receitas da Dívida Ativa

Anos	Receitas da Dívida Ativa (R$)
2004	2.047.510,61
2005	3.327.961,95
2006	3.693.094,43
2007	4.198.611,43
2008	4.739.891,03
2009	4.832.508,39
2010	5.721.246,32

A seguir descreve-se toda a metodologia usada no processo de previsão da arrecadação de Receita da Dívida Ativa.

5.1.1. Pré-processamento das variáveis de entrada

A série de Receitas da Dívida Ativa utilizada neste estudo de caso foi extraída do Município de Araruama. Essa série corresponde à arrecadação mensal de Receitas das dívidas ativas pagas por mês. As receitas pagas são frações ou o total da dívida de cada contribuinte. Os períodos da série utilizada correspondem de janeiro de 2004 a dezembro de 2010, num total de 84 observações.

Na Figura 13 são apresentados os valores mensais da Receita da Dívida Ativa, observados por ano (janeiro de 2004 a dezembro de 2010).

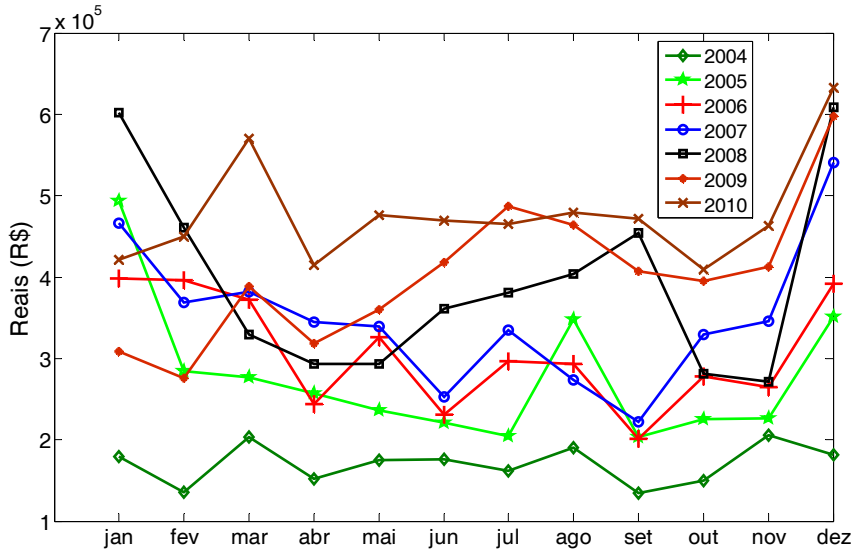


Figura 13. Valores mensais da Receitas da Dívida Ativa por ano

Na Figura 14, pode-se observar que a série de Receitas da Dívida Ativa $(S_1, S_2, \dots, S_{84})$ apresenta uma tendência de crescimento dos valores.

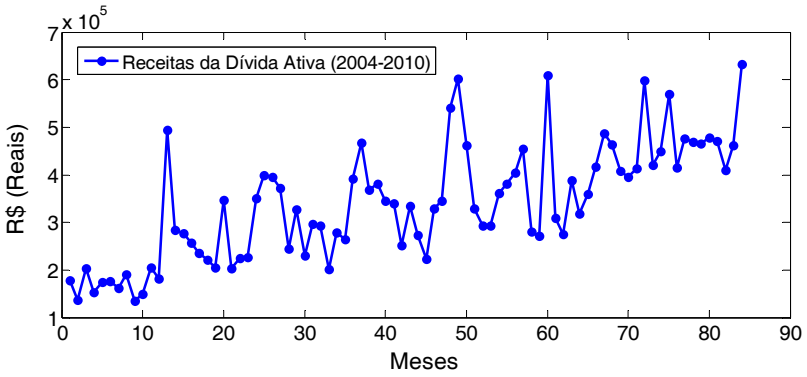


Figura 14. Trajetória da arrecadação mensal de Receitas da Dívida Ativa

Na Figura 15, é apresentada a autocorrelação da primeira diferença dos valores da série original $(D_t = S_{t+1} - S_t)$ da série de Receitas da Dívida Ativa. Pode-se observar que a série apresenta forte evidência de sazonalidade, dado que os retardos (Lag) 12 e 24 das autocorreleções são significativos.

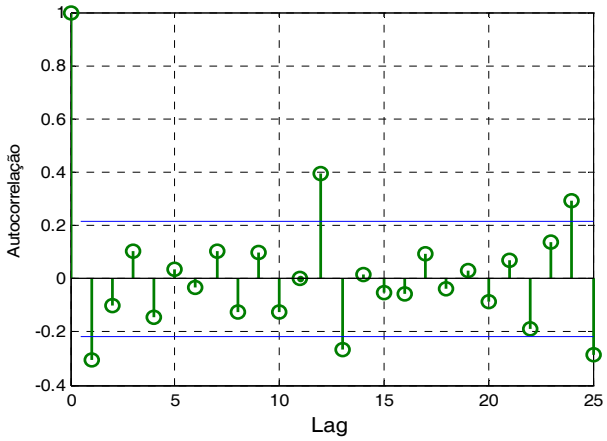


Figura 15. Autocorrelação da série diferenciada de Receitas da Dívida Ativa

A seguir para a variável endógena (Receitas da Dívida Ativa), são construídas as onze (11) variáveis de entradas (MM3, MM2, MM6, MM12, D12, D13, M-1, M-2, M3, M-6 e M-12), como indicadas na Tabela 4.

5.1.2.
Seleção das variáveis endógenas

Devido ao fato do histórico de dados ser curto, faz-se essencial realizar a seleção de variáveis. Essa é uma prática em Mineração de Dados, que ajuda a melhorar os resultados, filtrando as variáveis de entrada que não contribuem para explicar a saída. A seguir são descritos os procedimentos para a seleção de variáveis.

A. Seleção das variáveis de entrada

No capítulo 3 foram apresentados os métodos de seleção de variáveis PCAM, LSE, SIE e ReliefF que serão utilizados, a fim de se obter uma classificação das variáveis de entrada em função das variáveis de saída. A classificação indica as entradas mais relevantes para os diferentes métodos. A ordenação por método teve o objetivo de restringir o número de combinações de entradas (foram utilizadas apenas as sete mais bem avaliadas pelo método dentre as 11 propostas). Os resultados são mostrados na Tabela 16.

A coluna 1 da Tabela 16 indica a variável a ser avaliada, as colunas de 2 a 5 indicam as posições das variáveis para os métodos de seleção.

Tabela 16. Classificação dos métodos

Entradas	LSE	ReliefF	PCAM	SIE
MM3	4°	6°		7°
MM2	6°	10°		6°
MM6		3°	7°	
MM12		2°	6°	
D12	5°	4°	4°	2°
D13	2°	7°	3°	5°
M-1	1°	5°		1°
M-2	7°	8°		3°
M-3	3°	9°	2°	4°
M-6		11°	5°	
M-12		1°	1°	

Por exemplo, o método LSE classificou a variável M-1 (valor da Receita da Dívida Ativa do mês anterior ao mês que se deseja prever) em primeiro lugar, o quer dizer que essa variável é a mais importante, dentre as 11, para esse método. Já para o método ReliefF, a variável que melhor explica a variável de saída, nesse caso a Receita da Dívida Ativa, é a variável M-12, ou seja, o valor

da Receita da Dívida Ativa de 12 meses anterior ao valor da Dívida Ativa referente ao mês M que se deseja prever.

Da Tabela 16 pode-se observar que as classificações dos métodos são diferentes para cada variável de entrada. Nesta dissertação as classificações dos métodos de seleção de variáveis foram avaliadas para determinar qual método é mais eficiente para a variável endógena (Receitas da Dívida Ativa).

Número de redes com 7 variáveis de entrada — Para calcular o número total de combinações com 7 variáveis endógenas e subconjuntos variando de 2 a 7, foi utilizada a eq. (4.1). Na Tabela 17 é apresentado o calculo do número total de redes.

Tabela 17. Número total de combinações das 7 entradas endógenas

Número de Entradas da Rede Neural	Sub Conj. Entradas	Combinação	Conjuntos de Entradas Avaliados
7	2	C_2^7	21
7	3	C_3^7	35
7	4	C_4^7	35
7	5	C_5^7	21
7	6	C_6^7	7
7	7	C_7^7	1
			120

Identificação da melhor arquitetura de rede dos 4 métodos — O objetivo dessa avaliação é identificar os métodos de seleção que estão mais relacionados ao tipo de dado que ele processa, dadas às diferenças encontradas na classificação das variáveis de entrada por método (Tabela 16). Cada rede neural será treinada com dados do período de janeiro de 2004 a dezembro de 2009. O conjunto de dados de janeiro a dezembro de 2010 será utilizado apenas para efeitos de validação.

Foram realizadas muitas avaliações utilizando-se redes MLP (até 12 passos à frente) com entradas geradas a partir das combinações das entradas selecionadas pelos métodos (Tabela 16). As entradas indicadas por cada método foram utilizadas as combinando exaustivamente desde quaisquer duas entradas (das 7 selecionadas) até sete entradas para diferentes números de neurônios na camada escondida (variando de 1 a 6 neurônios).

Na Tabela 18 são apresentados os resultados das previsões das melhores arquiteturas de rede em média (100 inicializações). Pode-se observar que os resultados da variável de Receita da Dívida Ativa foram melhores com os

métodos PCAM e ReliefF. Sendo assim, decidiu-se não utilizar os métodos LSE e SIE, por que seus respectivos resultados não foram considerados satisfatórios, frente aos outros dois.

A Tabela 18 também mostra que os modelos PCAM obteve o melhor resultado para a seleção de entradas MM6, MM12 e M-3. O modelo RiliefF obteve um resultado muito próximo ao obtido pelo PCAM com as variáveis MM6, MM12, D13 e M-1. Essas variáveis foram identificadas após a combinação de variáveis mencionada acima.

Tabela 18. Erro MAPE dos métodos de seleção de variáveis

Método	**Entradas**	**Neurônios**	**MAPE %**
LSE	[MM3, M-1]	3	13,75
PCAM	[MM6, MM12, M-3]	4	9,84
ReliefF	[MM6, MM12, D13, M-1]	3	9,98
SIE	[MM3, M-1]	4	12,83

B. Seleção das variáveis de entrada associando dos métodos PCAM e ReliefF

As classificações das variáveis obtidas a partir dos métodos PCAM e ReliefF foram associadas, dando prioridade as entradas mais relevantes de cada método, a fim de se obter uma nova configuração de entradas. Na Tabela 19 são apresentadas as entradas escolhidas pelas associações dos métodos PCAM-ReliefF (coluna 4).

Tabela 19. Entradas da Rede de Receita da Dívida Ativa

Entradas	**ReliefF**	**PCAM**	**PCAM-ReliefF**
MM3	6°		
MM2	10°		
MM6	3°	7°	X
MM12	2°	6°	X
D12	4°	4°	X
D13	7°	3°	X
M-1	5°		X
M-2	8°		
M-3	9°	2°	X
M-6	11°	5°	
M-12	1°	1°	X

A escolha das variáveis buscou conciliar as 4 melhores variáveis de cada método de forma a totalizar 7 variáveis. Como havia variáveis que estavam entre as 4 melhores de ambos os métodos de seleção (por exemplo, M-12 e D12), optou-se por incluir o M-1, não apenas por ele ser 5° melhor valor para o método ReliefF, mas também por ser o valor imediatamente anterior ao valor que se deseja prever. Essa posição na série temporal costuma ser de grande ajuda no processo de previsão em modelos de previsão de séries temporais.

Identificação da melhor arquitetura de rede associando os métodos PCAM e ReliefF — Foram realizadas muitas avaliações utilizando-se redes MLP com entradas selecionadas pelos métodos PCAM e ReliefF e a associação dos métodos PCAM-ReliefF (Tabela 19). A identificação da arquitetura envolveu as variáveis de entradas que foram combinadas, desde quaisquer duas entradas (das 7 selecionadas) até sete entradas e avaliações com diferentes números de neurônios na camada escondida (variando de 1 a 6 neurônios). Na Tabela 20 são apresentados os resultados das previsões das melhores arquiteturas de rede em média (100 inicializações) para cada arquitetura. Os dois primeiros valores da Tabela 20 vieram da Tabela 18, apenas a terceira linha da Tabela 20 foi acrescentada. Ela mostra uma pequena melhora em média do efeito associativo entre os métodos de seleção de variáveis PCAM e ReliefF.

Tabela 20. Melhores arquiteturas em média de Receitas da Dívida Ativa

Método	Entradas	Neurônios	MAPE %
PCAM	[MM6, MM12, M-3]	4	9,84
ReliefF	[MM6, MM12, D13, M-1]	3	9,98
PCAM-ReliefF	[MM12, M-1]	4	9,69

5.1.3. Busca dos melhores pesos

O passo seguinte, segundo a metodologia adotada, é encontrar o melhor conjunto de pesos e bias para as melhores arquiteturas de rede em média obtida na seção anterior (Tabela 20). Para alcançar o objetivo, as melhores arquiteturas em média foram inicializadas 1000 vezes. Os resultados são apresentados na Tabela 21.

A melhor arquitetura de rede (entradas MM12 e M-1 e 4 neurônios) foi obtida pela combinação dos métodos PCAM-ReliefF, com MAPE igual a 5,63%.

Tabela 21. Resultados previstos de Receitas da Dívida Ativa

Métodos	Entradas	Neurônios	Mape %
PCAM	[MM6, MM12, M-3]	4	6,02
ReliefF	[MM6, MM12, D13, M-1]	3	5,79
PCAM-ReliefF	[MM12, M-1]	4	5,63

Na Figura 16 são apresentados os valores Previstos e Reais da melhor arquitetura de Receitas da Dívida Ativa do ano 2010 (período correspondente aos dados de validação).

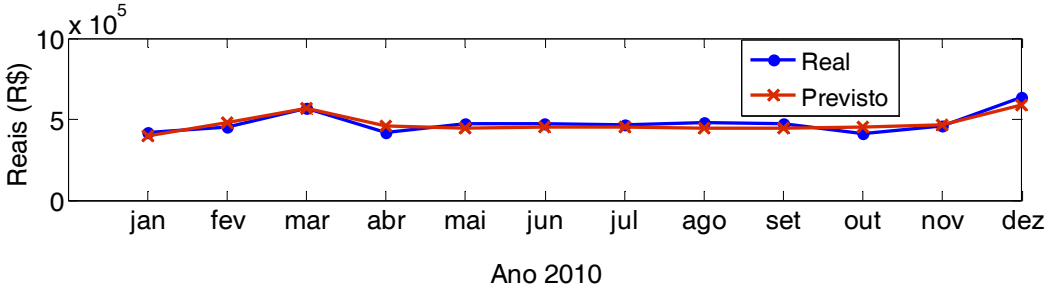


Figura 16. Previsões de Receitas da Dívida Ativa para o ano 2010

5.2. Previsão da arrecadação mensal de Receitas Tributárias

As Receitas Tributárias são compostas por impostos arrecadados diretamente pelo município de Araruama. O valor arrecadado anual influencia na composição do orçamento do Município. Assim para se realizar uma boa gestão, são necessárias ferramentas e funcionalidades, para a previsão de valores futuros da arrecadação de Receitas Tributárias. Dessa forma neste estudo de caso será realizada a previsão de 12 passos à frente da arrecadação mensal de Receitas Tributárias do Município de Araruama utilizando variáveis endógenas, ou seja, variáveis criadas a partir da variável que se deseja prever, nesse caso as Receitas Tributárias. Para a previsão será aplicado o modelo proposto na Figura 8.

As Receitas Tributárias são formadas pelas soma dos seguintes impostos TCE (2010):

- Imposto s/ a Propriedade Predial/Territorial Urbana (IPTU)
- Imposto s/ Serviços de Qualquer Natureza (ISSQN)
- Imposto s/ Transmissão de Bens Imóveis
- Imposto de Renda Retido na Fonte (IRRF)
- Outras Receitas Tributárias

A Figura 17 apresenta o percentual dos impostos que conformam as Receitas Tributárias do Município de Araruama do ano 2010.

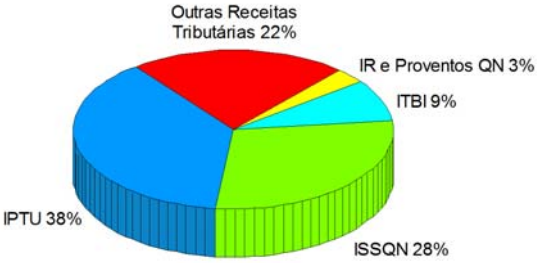


Figura 17. Impostos que conformam as Receitas Tributárias do ano 2010

A seguir descreve-se toda a metodologia usada no processo de previsão de Receitas Tributárias.

5.2.1. Pré-processamento das variáveis de entrada

A série de Receitas Tributárias utilizada neste estudo de caso foi extraída do site do TCE. Esta série corresponde à arrecadação mensal de Receitas Tributárias. Os períodos da série utilizada correspondem de março de 2005 a dezembro de 2010, num total de 70 observações.

Na Figura 18, são apresentados os valores mensais (janeiro a dezembro), dos anos 2005 a 2010, das arrecadações de Receitas Tributárias.

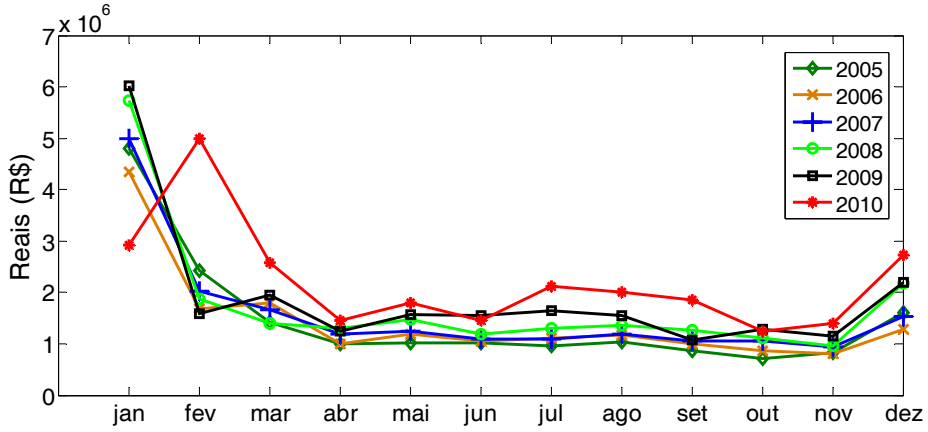


Figura 18. Valores mensais de Receitas Tributárias dos anos 2005 a 2010

Na Figura 19, pode-se observar que a série das Receitas Tributárias $(S_1, S_2, \dots, S_{70})$ apresenta sazonalidade e uma leve tendência de crescimento dos valores.

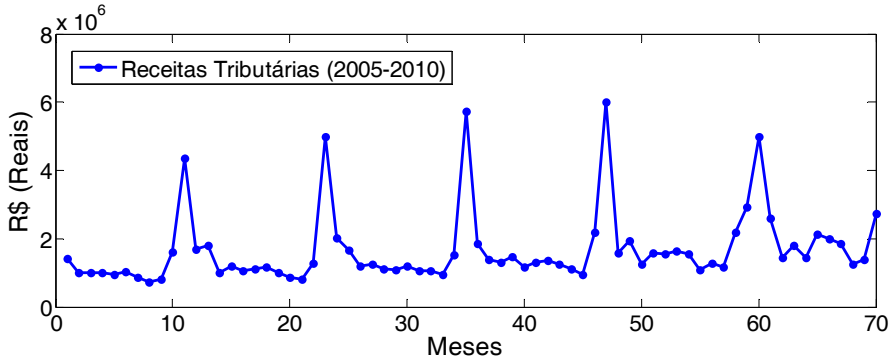


Figura 19. Trajetória da arrecadação mensal de Receitas Tributárias

Na Figura 20 é apresentado a autocorrelação da primeira diferença dos valores da série original $D_t = (S_{t+1} - S_t)$ da série de Receitas Tributárias. Pode-se observar que a série apresenta forte evidência de sazonalidade, dado que os retardos (Lag) 12, 24 e 36 das autocorrelações são significativos.

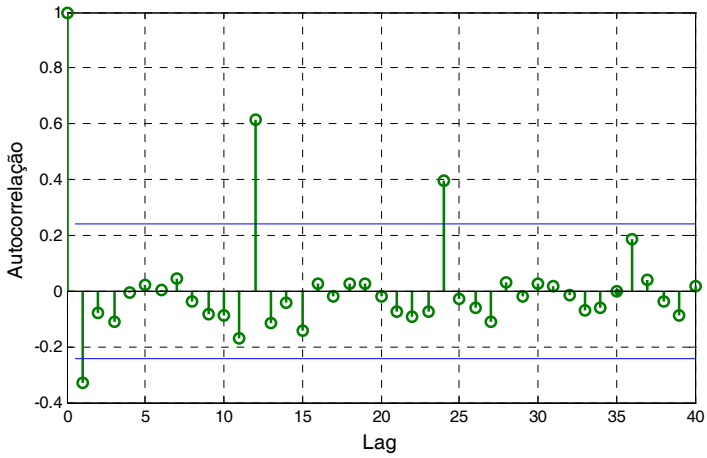


Figura 20. Autocorrelação da série de Receitas Tributárias

A seguir para a variável endógena (Receitas Tributárias), são construídas as onze (11) variáveis de entrada (MM3, MM2, MM6, MM12, D12, D13, M-1, M-2, M3, M-6 e M-12), como indicadas na Tabela 4.

5.2.2.
Seleção das variáveis endógenas

Devido ao fato do histórico de dados ser curto, faz-se essencial realizar a seleção de variáveis. Essa é uma prática em Mineração de Dados, que ajuda a melhorar os resultados, filtrando as variáveis de entrada que não contribuem para explicar a saída. A seguir são descritos os procedimentos para a seleção de variáveis.

A. Seleção das variáveis de entrada

No capítulo 3 foram apresentados os métodos de seleção de variáveis PCAM, LSE, SIE e ReliefF que serão utilizados, a fim de se obter uma classificação das variáveis de entrada em função das variáveis de saída. A classificação indica as entradas mais relevantes para os diferentes métodos. A ordenação por método teve o objetivo de restringir o número de combinações de entradas (foram utilizadas apenas as sete mais bem avaliadas pelo método dentre as 11 propostas). Os resultados são mostrados na Tabela 22.

A coluna 1 da Tabela 22 indica a variável a ser avaliada, as colunas de 2 a 5 indicam as posições das variáveis para os métodos de seleção.

Tabela 22. Classificação dos métodos

Entradas	LSE	ReliefF	PCAM	SIE
MM3	3°	7°		4°
MM2	2°	4°		6°
MM6		10°	7°	
MM12		8°	6°	
D12	4°	3°	4°	5°
D13	6°	5°	1°	2°
M-1	7°	2°		3°
M-2	1°	6°	2°	7°
M-3	5°	9°		1°
M-6		11°	3°	
M-12		1°	5°	

Por exemplo, o método LSE classificou a variável M-2 (valor de Receitas Tributárias de dois meses anteriores ao mês que se deseja prever) em primeiro lugar, o quer dizer que essa variável é a mais importante, dentre as 11, para esse método. Já para o método ReliefF, a variável que melhor explica a saída da variável que se deseja prever, nesse caso a Receita Tributárias, é a variável M-

12, ou seja, o valor da Receita Tributárias de 12 meses anterior ao mês que se deseja prever.

Da Tabela 22 pode-se observar que as classificações dos métodos são diferentes para cada variável de entrada. Nesta dissertação as classificações dos métodos de seleção de variáveis foram avaliadas para determinar qual método é mais eficiente para a variável endógena (Receitas Tributárias).

Identificação da melhor arquitetura de rede dos 4 métodos — O objetivo dessa avaliação é identificar os métodos de seleção que estão mais relacionados ao tipo de dado que ele processa, dada às diferenças encontradas na classificação das variáveis de entrada por método (Tabela 22). Cada rede neural será treinada com dados do período de março de 2005 a dezembro de 2009. O conjunto de dados de janeiro a dezembro de 2010 será utilizado apenas para efeitos de validação.

Foram realizadas muitas avaliações utilizando-se redes MLP (até 12 passos à frente) com entradas geradas a partir das combinações das entradas selecionadas pelos métodos (Tabela 22). As entradas indicadas por cada método foram utilizadas as combinando exaustivamente desde quaisquer duas entradas (das 7 selecionadas) até sete entradas para diferentes números de neurônios na camada escondida (variando de 1 a 6 neurônios).

Na Tabela 23 são apresentados os resultados das previsões das melhores arquiteturas de rede em média (100 inicializações). Pode-se observar que os resultados da variável de Receitas Tributárias foram melhores com os métodos PCAM e ReliefF. Sendo assim, decidiu-se não utilizar os métodos LSE e SIE, por que seus respectivos resultados não foram considerados satisfatórios, frente aos outros dois.

A Tabela 23 também mostra que o modelo PCAM obteve o melhor resultado para a seleção de entradas MM6 e MM12. O modelo RiliefF obteve um resultado maior ao obtido pelo PCAM com as variáveis D12, D13 e M-12. Essas variáveis foram identificadas após a combinação de variáveis mencionada acima.

Tabela 23. Erro MAPE dos métodos de seleção de variáveis

Método	Entradas	Neurônios	MAPE %
LSE	[MM2, M-3]	1	32,90
PCAM	[MM6, MM12]	1	28,55
ReliefF	[D12, D13, M-12]	1	31,12
SIE	[MM3, M-3]	1	33,52

B. Seleção das variáveis de entrada associando os métodos PCAM e ReliefF

As classificações das variáveis obtidas a partir dos métodos PCAM e ReliefF foram associadas, dando prioridade as entradas mais relevantes de cada método, a fim de se obter uma nova configuração de entradas. Na Tabela 24 são apresentadas as entradas escolhidas pelas associações dos métodos PCAM-ReliefF (coluna 4).

Tabela 24. Entradas da Rede de Receitas Tributárias

Entradas	ReliefF	PCAM	PCAM-ReliefF
MM3	7°		
MM2	4°		
MM6	10°	7°	
MM12	8°	6°	X
D12	3°	4°	X
D13	5°	1°	X
M-1	2°		X
M-2	6°	2°	X
M-3	9°		
M-6	11°	3°	X
M-12	1°	5°	X

A escolha das variáveis buscou conciliar os melhores 3 variáveis de cada método de forma a totalizar 7 variáveis. Para completar a variável que falta foi selecionado a variável MM12 já que essa variável foi selecionada por os dois métodos.

Identificação da melhor arquitetura de rede associando os métodos PCAM e ReliefF — Foram realizadas muitas avaliações utilizando-se redes MLP com entradas selecionadas pelos métodos PCAM e ReliefF e a associação dos métodos PCAM-ReliefF (Tabela 24). A identificação da arquitetura envolveu as variáveis de entradas que foram combinadas, desde quaisquer duas entradas (das 7 selecionadas) até sete entradas e avaliações com diferentes números de neurônios na camada escondida (variando de 1 a 6 neurônios). Na Tabela 25 são apresentados os resultados das previsões das melhores arquiteturas de rede em média (100 inicializações) para cada arquitetura. Os dois primeiros valores da Tabela 25 vieram da Tabela 23 e apenas a terceira linha da Tabela 25 foi acrescentada.

O resultado aponta a melhor arquitetura com entradas MM6 e MM12. Nesse caso a junção dos modelos de seleção de variáveis apresentou resultados muito próximos ao resultado obtido a partir do método PCAM.

Tabela 25. Melhores arquiteturas em média de Receitas Tributárias

Método	Entradas	Neurônios	MAPE %
PCAM	[MM6, MM12]	1	28,55
ReliefF	[D12, D13, M-12]	1	31,12
PCAM-ReliefF	[MM12, D13]	1	28,66

5.2.3. Busca dos melhores pesos

O passo seguinte, segundo a metodologia adotada, é encontrar o melhor conjunto de pesos e bias para as melhores arquiteturas de rede em média obtida na seção anterior (Tabela 25). Para lograr o objetivo, as melhores arquiteturas em média foram inicializadas 1000 vezes. Os resultados são apresentados na Tabela 26.

A melhor arquitetura de rede (entradas D12, D13 e M-12 e 1 neurônio) foi obtida com o método ReliefF, com MAPE igual a 10,87%.

Tabela 26. Resultados previstos de previsão de Receitas Tributárias

Métodos	Entradas	Neurônios	Mape %
PCAM	[MM6, MM12]	1	24,46
ReliefF	[D12, D13, M-12]	1	10,87
PCAM-ReliefF	[MM12, D13]	1	16,76

Na Figura 21 são apresentados os valores Previstos e Reais da melhor arquitetura de Receitas de Tributárias do ano 2010 (período correspondente aos dados de validação).

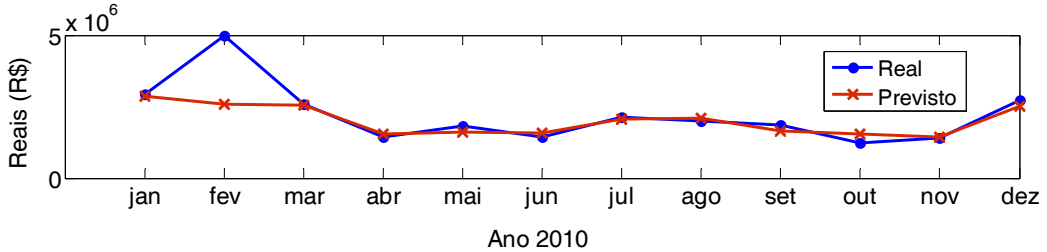


Figura 21. Previsões de Receitas Tributárias para o ano 2010

Na Tabela 27 é apresentada a previsão de Receitas Tributárias realizada pela Secretaria de Fazenda do Município de Araruama (SFMA), que esta disponível no site do TCE, e a previsão do modelo proposto nesta dissertação para o ano 2010 (somatório das previsões mensais), além do valor real, que corresponde à soma da arrecadação mensal de Receitas Tributárias de janeiro a dezembro do ano 2010. O resultado obtido com o modelo proposto foi melhor em comparação ao valor previsto pela SFMA. Infelizmente não se sabe como foi realizada a previsão apresentada pela SFMA. A princípio, o que sabe é que esta previsão é uma estimativa considerando os valores dos anos anteriores e expectativas baseadas em políticas.

Tabela 27. Comparação da previsão de Receitas Tributárias para o ano 2010

Receitas Tributárias (Janeiro a Dezembro)	**Ano 2010**	**Erro Relativo (|Previsto-Real|)/Real**
Valor Real	26.565.500,30	-
Previsão do SFMA	23.649.510,90	10,98%
Previsão do Modelo Proposto	24.081.821,80	9,35%

A seguir foi realizado um teste (Recall) utilizando como entrada da rede, a informação da arrecadação mensal do ano 2010, para prever o ano 2011, com os mesmos pesos e bias da melhor rede obtida (Tabela 26) para essa série.

O resultado do teste para o ano 2011 é apresentado na Tabela 28. Pode-se observar que o resultado com o modelo proposto foi melhor do que a previsão feita pela SFMA.

Tabela 28. Comparação da previsão de Receitas Tributárias para o ano 2011

Receitas Tributárias (Janeiro a Dezembro)	**Total – 2011**	**Erro Relativo (|Previsto-Real|)/Real**
Real	31.995.794,60	-
Previsão do SFMA	25.398.409,30	20,6%
Previsão do Modelo Proposto	27.169.503,90	15,1%

5.3.
Previsão da arrecadação mensal do ISSQN utilizando variáveis endógenas e exógenas

O ISSQN é diretamente arrecadado pelo município de Araruama. O valor arrecadado anual influencia na composição do orçamento do Município. Assim para se realizar uma boa gestão, são necessárias ferramentas e funcionalidades, para a previsão de valores futuros da arrecadação do ISSQN. Dessa forma neste estudo de caso será realizada a previsão de 12 passos à frente da arrecadação mensal do ISSQN do Município de Araruama utilizando variáveis endógenas e exógenas.

O modelo apresentado nessa seção investigou variáveis endógenas (variáveis explicativas construídas a partir da variável que se deseja prever) e exógenas (variáveis explicativas construídas a partir de outras variáveis distintas da variável que se deseja prever) para avaliar os melhores conjuntos de variáveis de entrada para as redes neurais.

Para a previsão da variável endógena (ISSQN) foi aplicado o modelo proposto na Figura 8, e para a previsão utilizando ambas variáveis (endógenas e exógenas) foi aplicado o modelo proposto na Figura 10. A seguir descreve-se toda a metodologia usada no processo de previsão da arrecadação do ISSQN.

5.3.1.
Previsão da arrecadação do ISSQN com variáveis endógenas

Nesta seção é realizado a previsão de 12 passos à frente da arrecadação mensal do ISSQN do Município de Araruama utilizando variáveis endógenas. Para a previsão será aplicado o modelo proposto na Figura 8.

5.3.1.1.
Pré-processamento das variáveis de entrada

As séries do ISSQN utilizada neste estudo de caso foram extraídas do site do TCE, e do Município de Araruama. Esta série corresponde à arrecadação mensal do ISSQN.

Os períodos das séries endógenas e exógenas utilizadas foram de janeiro de 2004 a dezembro de 2010, num total de 84 observações.

Na Figura 22. são apresentados os valores mensais do ISSQN, observados por ano (janeiro de 2004 a dezembro de 2010). Percebe-se que as séries não apresentam um comportamento sazonal, como as outras séries já

apresentadas Da Figura 23, pode-se observar uma tendência de crescimento dos valores $(S_1, S_2, \ldots, S_{84})$, do ISSQN.

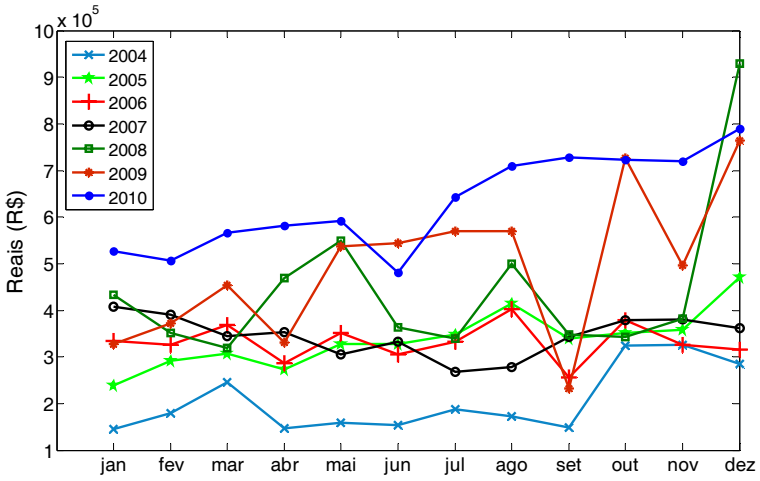


Figura 22. Valores mensais do ISSQN por ano

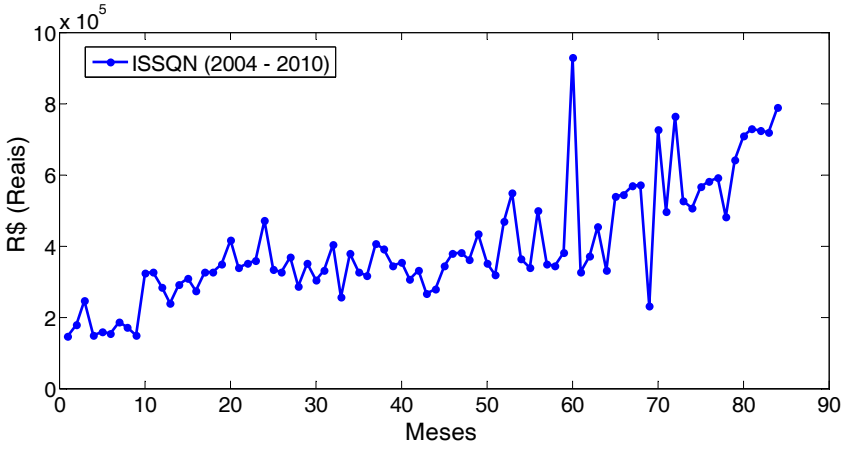


Figura 23. Trajetória da arrecadação da série ISSQN

Na Figura 24, apresenta a autocorrelação da primeira diferença dos valores da série original $(D_t = S_{t+1} - S_t)$ da série ISSQN. Pode-se observar que a série não apresenta evidência de sazonalidade nos retardos (Lag) 12, 24 e 36. A série tem maior correlação com alguns meses dentre os 12 meses anteriores.

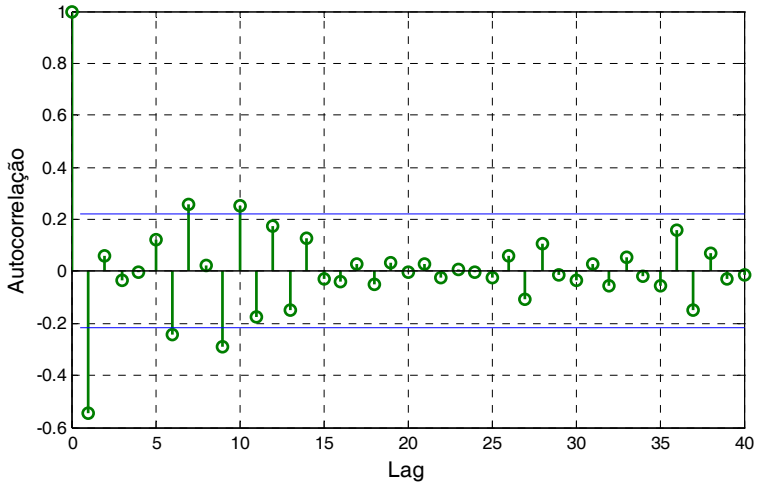


Figura 24. Autocorrelação da série diferenciada do ISSQN

A seguir para a variável endógena (ISSQN), são construídas as onze (11) variáveis de entrada (MM3, MM2, MM6, MM12, D12, D13, M-1, M-2, M3, M-6 e M-12), como indicadas na Tabela 4.

5.3.1.2. Seleção das variáveis endógenas do ISSQN

Devido ao fato do histórico de dados ser curto, faz-se essencial realizar a seleção de variáveis. Essa é uma prática em Mineração de Dados, que ajuda a melhorar os resultados, filtrando as variáveis de entrada que não contribuem para explicar a saída. A seguir são descritos os procedimentos para a seleção de variáveis endógenas.

A. Seleção das variáveis de entrada endógenas (ISSQN)

No capítulo 3 foram apresentados os métodos LSE, PCAM, ReliefF, e SIE, que serão utilizados a fim de se obter uma classificação das variáveis endógenas (ISSQN) segundo esses métodos. A ordenação por método teve o objetivo de restringir o número de combinações de entradas (foram utilizadas apenas as sete mais bem avaliadas pelo método dentre as 11 propostas). Os resultados são mostrados naTabela 29.

A coluna 1 da Tabela 29 indica a variável a ser avaliada, as colunas de 2 a 5 indicam as posições das variáveis para os métodos de seleção.

Tabela 29. Seleção de entradas da variável endógena (ISSQN) pelos 4 métodos de seleção de variáveis

Entradas	LSE	PCAM	ReliefF	SIE
MM3	1°		2°	2°
MM2	3°		6°	3°
MM6		7°	4°	
MM12		6°	1°	
D12			9°	5°
D13	4°	3°	5°	4°
M-1	2°	1°	10°	1°
M-2	5°	2°	7°	7°
M-3	6°		8°	6°
M-6		5°	11°	
M-12		4°	3°	

A partir da Tabela 29 pode-se verificar, por exemplo, que o método PCAM classificou a variável M-1 (valor do ISSQN do mês anterior ao mês que se deseja prever) em primeiro lugar, o quer dizer que essa variável é a mais importante, dentre as 11, para esse método. Já para o método ReliefF (Tabela 29), a variável que melhor explica a variável que se deseja prever, nesse caso o ISSQN, é a variável MM12, ou seja, a média móvel dos 12 últimos valores do ISSQN antes do mês que se deseja prever.

Da Tabela 29 pode-se observar que as classificações dos métodos são diferentes para cada variável de entrada. Nesta dissertação as classificações dos métodos de seleção de variáveis foram avaliadas para determinar qual método é mais eficiente para a variável endógena (ISSQN).

Identificação da melhor arquitetura de rede dos 4 métodos — O objetivo dessa avaliação é identificar os métodos de seleção que estão mais relacionados ao tipo de dado que ele processa, dada às diferenças encontradas na classificação das variáveis de entrada por método (Tabela 29). Cada rede neural será treinada com dados do período de janeiro de 2004 a dezembro de 2009. O conjunto de dados de janeiro a dezembro de 2010 será utilizado apenas para efeitos de validação.

Para a avaliação dos métodos de seleção de variáveis, a partir das variáveis endógenas indicadas por cada método (Tabela 29), foi realizada a previsão do ISSQN através de uma rede *multistep* (até 12 passos à frente). Para essas avaliações considerou-se as combinações das 7 entradas selecionadas e subconjuntos variando de 2 a 7 (Tabela 17). Também se investigou diferentes

neurônios na camada escondida (de 1 a 6 neurônios) para cada uma das combinações de variáveis de entrada, e, além disso, cada arquitetura foi inicializada 50 vezes.

Na Tabela 30 são apresentados os resultados das previsões das melhores arquiteturas de rede em média (50 inicializações). Pode-se observar que os resultados das previsões da variável ISSQN foram bem melhores com os métodos PCAM (13,01%) e ReliefF (13.54%). Sendo assim, decidiu-se não utilizar as classificações dos métodos LSE e SIE para a variável endógena (ISSQN) por que seus respectivos resultados não foram considerados satisfatórios.

Tabela 30. Melhores arquiteturas da variável endógena (ISSQN) em média

Variáveis	LSE	PCAM	ReliefF	SIE
ISSQN	22,84 %	13,01 %	13,54 %	23,34 %

B. Seleção das variáveis de entrada associando os métodos PCAM e ReliefF

Dado que os resultados de previsão (Tabela 30) por redes neurais, utilizando as variáveis endógenas do ISSQN selecionadas pelos métodos PCAM e RefliefF, foram superiores, avaliou-se também outras configurações de variáveis de entrada associando-se as variáveis selecionadas pelos métodos PCAM e RefliefF. Deu-se prioridade às entradas mais relevantes de cada método de seleção e, a partir da associação das variáveis selecionadas, obteve-se uma nova configuração de entradas. Na Tabela 31 são apresentadas as variáveis endógenas escolhidas pelas associações dos métodos PCAM-ReliefF (coluna 4). Por exemplo, a variável de entrada M-1 foi escolhida porque é mais relevante para o método PCAM.

Tabela 31. Entradas das variáveis endógenas (ISSQN)

Entradas	ReliefF	PCAM	Associação ReliefF-PCAM
MM3	2°		X
MM2	6°		
MM6	4°	7°	X
MM12	1°	6°	X
D12	9°		
D13	5°	3°	X
M-1	10°	1°	X
M-2	7°	2°	X
M-3	8°		
M-6	11°	5°	
M-12	3°	4°	X

A arquitetura de rede com as entradas endógenas da Tabela 31 é apresentado na Figura 25. , onde a arquitetura têm 7 entrada (MM3, MM6, MM12, D13, M-1, M-2, M-12), uma camada escondida com seis neurônios e um neurônio na camada de saída.

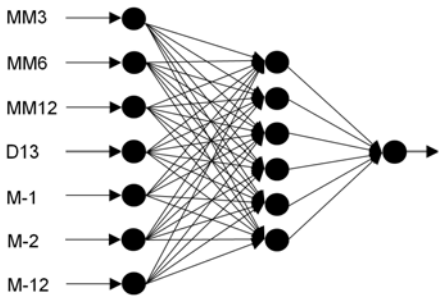


Figura 25. Arquitetura da variável endógena para a previsão do ISSQN

Identificação da melhor arquitetura de rede associando os métodos PCAM e ReliefF — Nesta seção é apresentada a metodologia para identificar a melhor arquitetura endógena em média. Foram avaliadas muitas arquiteturas de redes neurais do tipo MLP: combinações de entradas (desde a combinação de duas em sete variáveis até 7 variáveis de entrada) e diferentes números de neurônios na camada escondida (variando de 1 a 6 neurônios), para cada combinação de variável de entrada. Todas as arquiteturas foram inicializadas 100 vezes.

Para as avaliações das arquiteturas com as variáveis de entradas associadas pelos métodos PCAM e ReliefF foram utilizadas as entradas escolhidas na Tabela 31. O resultado da melhor arquitetura endógena (ISSQN) em média é apresentado na Tabela 32.

Tabela 32. Melhor arquitetura da variável endógena (ISSQN) em média

Variável	Entradas	Neurônios	Mape %
ISSQN	[MM12, D13]	1	12,74

5.3.2. Previsão das variáveis exógenas

Para se avaliar a previsão do ISSQN utilizando variáveis exógenas, também se faz necessária a criação de modelos para prever variáveis exógenas. Dessa forma nesta seção é realizado a previsão de 12 passos à frente das variáveis exógenas. Para a previsão será aplicado o modelo proposto na Figura 10.

5.3.2.1. Pré-processamento das variáveis de entrada exógenas

Neste estudo de caso foram avaliadas onze (11) variáveis exógenas, as quais estão apresentadas no Apêndice A. Na

Tabela 33, são apresentadas as descrições das variáveis exógenas.

Tabela 33. Descrição das variáveis exógenas

Variáveis Exógenas	Descrição
TIOCSFNRJT	Taxa de inadimplência das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Total (%).
TIOCSFNRJPF	Taxa de inadimplência das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Pessoas físicas (%)
TIOCSFNRJPJ	Taxa de inadimplência das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Pessoas jurídicas %.
SOCSFNRJPJ	Saldo das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Pessoas Jurídicas - R$ (milhões).
SOCSFNRJTot	Saldo das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Total - R$ (milhões).
SOCSFNRJPF	Saldo das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Pessoas Físicas - R$ (milhões)
SELIC	Sistema Especial de Liquidação e de Custódia.
IPCAE	Índice Nacional de Preços ao Consumidor Amplo Especial – IPCA-E.
INPC	Índice Nacional de Preços ao Consumidor
SBCRJ	Saldo da Balança Comercial - Rio de Janeiro - US$ (mil)
OCSFPrASPEM	Operações de crédito do sistema financeiro privado (Risco normal) - Ao setor público estadual e municipal

A seguir, é apresentada a primeira etapa de seleção de variáveis exógenas. Essa seleção foi realizada através da correlação cruzada da variável

endógena (ISSQN) com cada variável exógena, utilizando o software econométrico Eviews (Stock e Watson, 2007).

Na Tabela 34 as variáveis exógenas com correlações significativas são: TIOCSFNRJT, TIOCSFNRJPJ, SOCSFNRJPJ, SOCSFNRJTot, SOCSFNRJPF e SELIC. As variáveis exógenas com correlações baixas (TIOCSFNRJPF, IPCAE, INPC, SBCRJ e OCSFPrASPEM) não serão utilizadas no estudo.

As correlações cruzadas com retardos no tempo de 0 a 12 (Lag_0 até Lag_{12}) encontram-se nas tabelas (Apêndice B).

Tabela 34. Correlação cruzada da variável ISSQN com as variáveis exógenas

Exógenas	**Lag_0**	**Lag_1**	**Lag_2**	**Lag_3**	**Lag_4**	**Lag_5**	**Lag_6**
TIOCSFNRJT	-0,575	-0,550	-0,529	-0,519	-0,489	-0,476	-0,447
TIOCSFNRJPF	-0,210	-0,166	-0,143	-0,113	-0,074	-0,039	-0,017
TIOCSFNRJPJ	-0,549	-0,528	-0,510	-0,503	-0,479	-0,469	-0,443
SOCSFNRJPJ	0,785	0,753	0,719	0,697	0,656	0,626	0,590
SOCSFNRJTot	0,791	0,756	0,722	0,698	0,658	0,626	0,590
SOCSFNRJPF	0,798	0,755	0,722	0,693	0,656	0,620	0,585
SELIC	-0,528	-0,532	-0,515	-0,513	-0,491	-0,519	-0,490
IPCAE	-0,107	-0,097	-0,214	-0,204	-0,201	-0,101	0,018
INPC	0,013	0,015	-0,052	-0,161	-0,156	-0,051	0,045
SBCRJ	0,172	0,005	-0,025	0,016	0,086	0,075	0,130
OCSFPrASPEM	0,240	0,234	0,223	0,265	0,248	0,225	0,206

A seguir para as variáveis exógenas (TIOCSFNRJT, TIOCSFNRJPJ, SOCSFNRJPJ, SOCSFNRJTot, SOCSFNRJPF), foram construídas as onze (11) variáveis de entradas (MM3, MM2, MM6, MM12, D12, D13, M-1, M-2, M3, M-6 e M-12) como indicadas na Tabela 4.

5.3.2.2. Seleção das variáveis exógenas

A seleção de variáveis exógenas foi realizada utilizando os métodos LSE, PCAM, ReliefF, e SIE, nas variáveis exógenas pré-selecionadas da seção anterior (item 5.3.2.1).

Seleção das variáveis de entrada exógenas — Para a avaliação das variáveis de entrada das séries exógenas selecionadas (TIOCSFNRJT, TIOCSFNRJPJ, SOCSFNRJPJ, SOCSFNRJTot, SOCSFNRJPF e SELIC), foram utilizadas os

métodos de seleção de variáveis LSE, PCAM, ReliefF e SIE. As seleções das variáveis mais relevantes para cada método são apresentadas nas Tabelas 35, 36, 37 e 38.

A coluna 1 da Tabela 35 indica a variável de entrada a ser avaliada, as colunas de 2 a 7 indicam a ordenação, em ordem de importância, de cada variável para o método LSE. As outras Tabelas 36, 37 e 38, seguem a mesma estrutura.

- **Seleção de variáveis pelo método LSE**

Tabela 35. Classificação das variáveis de entrada exógenas pelo método LSE

Entradas	TIOCSFNRJT	TIOCSFNRJPJ	SOCSFNRJPJ	SOCSFNRJTot	SOCSFNRJPF	SELIC
MM3	2°	4°	1°	5°	2°	1°
MM2	5°	1°	4°	1°	3°	5°
D12	7°	3°	7°	6°	4°	4°
D13	3°	6°	6°	4°		
M-1	4°	7°	3°	7°	5°	6°
M-2	6°	2°	2°	3°	1°	2°
M-3	1°	5°	5°	2°	6°	3°

- **Seleção de variáveis pelo método PCAM**

Tabela 36. Classificação das variáveis de entrada exógenas pelo método PCAM

Entradas	TIOCSFNRJT	TIOCSFNRJPJ	SOCSFNRJPJ	SOCSFNRJTot	SOCSFNRJPF	SELIC
MM6	7°	7°	7°	7°	7°	7°
MM12	6°	6°	6°	6°	6°	6°
D12	5°			5°	5°	
D13	4°	4°		3°	4°	
M-1	2°		1°	1°	1°	1°
M-2		2°	5°			3°
M-3		5°	3°			5°
M-6	3°	3°	4°	4°	3°	4°
M-12	1°	1°	2°	2°	2°	2°

- **Seleção de variáveis pelo método RelifF**

Tabela 37. Classificação das variáveis de entrada exógenas pelo método RelifF

Entradas	TIOCSFNRJT	TIOCSFNRJPJ	SOCSFNRJPJ	SOCSFNRJTot	SOCSFNRJPF	SELIC
MM3	6°	5°	6°	5°	6°	2°
MM2	4°	4°	3°	3°	4°	3°
MM6	8°	8°	8°	8°	9°	9°
MM12	9°	9°	4°	7°	7°	6°
D12	3°	2°	11°	11°	11°	11°
D13	1°	1°	5°	6°	2°	10°
M-1	2°	3°	2°	2°	3°	7°
M-2	5°	6°	7°	4°	5°	1°
M-3	11°	10°	9°	9°	8°	5°
M-6	10°	11°	10°	10°	10°	4°
M-12	7°	7°	1°	1°	1°	8°

- **Seleção de variáveis pelo método SIE**

Tabela 38. Classificação das variáveis de entrada exógenas pelo método SIE

Entradas	TIOCSFNRJT	TIOCSFNRJPJ	SOCSFNRJPJ	SOCSFNRJTot	SOCSFNRJPF
MM3	1°	1°	4°	1°	4°
MM2	3°	3°	5°	3°	5°
D12	7°	4°	2°	4°	2°
D13	6°	6°	6°	6°	7°
M-1	5°	5°	1°	5°	1°
M-2	4°	2°	3°	2°	3°
M-3	2°	7°	7°	7°	6°

Identificação da melhor arquitetura de rede dos 4 métodos (LSE, PCAM, ReleifF e SIE) — O objetivo dessa identificação é selecionar variáveis exógenas com menor MAPE e identificar qual(is) métodos de seleção de variáveis são mais indicados para prever cada variável exógena.

As variáveis selecionadas pelos métodos LSE, PCAM, ReliefF e SIE, apresentadas nas Tabelas 35, 36, 37 e 38, foram avaliadas como entradas para os modelos de cada variável exógena a ser prevista (TIOCSFNRJT, TIOCSFNRJPJ, SOCSFNRJPJ, SOCSFNRJTot, SOCSFNRJPF e SELIC) através de uma rede multistep (até 12 passos à frente). Para esses testes também foram feitas combinações das entradas selecionadas (desde a combinação de duas em sete variáveis até 7 variáveis de entrada). Também foram investigadas diferentes arquiteturas de redes, a partir da variação do número de neurônios na camada escondida (de 1 a 6 neurônios). Além disso, cada arquitetura foi inicializada 50

vezes. O número de inicializações foi menor, em comparação com outros testes apresentados, visando redução do tempo computacional.

Da Tabela 39, pode-se observar que os resultados das variáveis exógenas (TIOCSFNRJT, TIOCSFNRJPJ, SOCSFNRJPJ, SOCSFNRJTot, SOCSFNRJPF) foram semelhantes.

Tabela 39. Melhores arquiteturas das variáveis exógenas em média

Variáveis	LSE	PCAM	ReliefF	SIE
TIOCSFNRJT	2,66 %	2,22 %	2,93 %	2,72 %
TIOCSFNRJPJ	3,85 %	3,77 %	3,79 %	4,08 %
SOCSFNRJPJ	1,67 %	1,26 %	1,28 %	1,74 %
SOCSFNRJTot	1,60 %	0,83 %	0,76 %	1,57 %
SOCSFNRJPF	2,57 %	2,17 %	1,90 %	2,60 %
SELIC *	6,92 %	6,84 %	7,00 %	--

***** Daqui para frente não será utilizada a variável exógena SELIC por que obteve um resultado não satisfatório.

Pode-se observar na Tabela 39 que os resultados para os 4 métodos foram semelhantes, com pequeno destaque para os métodos PCAM e RefliefF. Dessa forma, procedeu-se a duas avaliações:

1. Associando-se as variáveis indicadas pelos métodos PCAM e RefliefF.
2. Realizado uma avaliação sobre as variáveis indicadas pela combinação dos quatro métodos através pesos.

A. Seleção das variáveis de entrada exógenas mediante a associação dos métodos PCAM e ReliefF e os 4 métodos

CASO 1: Associando os métodos PCAM e ReliefF — Para escolher as variáveis entradas associadas dos métodos PCAM e ReliefF para previsão das variáveis exógenas, utilizou-se as entradas mais relevantes de cada método. Na Tabela 40 são apresentadas as variáveis de entrada escolhidas mediante a associação dos métodos PCAM e ReliefF.

Tabela 40. Entradas das variáveis exógenas associando os métodos PCAM-ReliefF

Entradas	TIOCSFNRJT	TIOCSFNRJPJ	SOCSFNRJPJ	SOCSFNRJTot	SOCSFNRJPF
MM2	X		X	X	X
MM12	X	X	X	X	X
D12	X	X		X	X
D13	X	X		X	X
M-1	X		X	X	X
M-2		X	X		
M-3		X	X		
M-6	X	X	X	X	X
M-12	X	X	X	X	X

CASO 2: Escolha de entradas exógenas utilizando os 4 métodos combinados através de pesos — Neste caso para escolher as entradas exógenas, foram utilizadas as ordenações das variáveis dos 4 métodos (LSE, PCAM, ReliefF, e SIE), e para cada método é associado um peso. Na Tabela 41 são apresentados os pesos para os 4 métodos. Esses pesos foram indicados de forma heurística, onde os modelos "com mais características não lineares" tiveram pesos maiores.

Tabela 41. Pesos associados aos métodos de seleção de variáveis

Método	Peso
PCAM	1
LSE	3
SIE	1
ReliefF	2

A seguir são aplicados os pesos às classificações dos 4 métodos utilizando a equação 5.1:

$$Nova\ Classificação\ = Peso * (11 - Classificação\ da\ variável\ pelo\ Método) \quad (5.1)$$

Nas Tabelas 42, 43, 44 e 45, são apresentados os resultados aplicando a eq. (5.1) e os pesos às classificações dos métodos LSE (Tabela 35), PCAM (Tabela 36), ReliefF (Tabela 37), e SIE (Tabela 38).

Tabela 42. Aplicando o peso às classificações do método LSE

Entradas	TIOCSFNRJT	TIOCSFNRJPJ	SOCSFNRJPJ	SOCSFNRJTot	SOCSFNRJPF
MM3	27	21	30	18	27
MM2	18	30	21	30	24
MM6	0	0	0	0	0
MM12	0	0	0	0	0
D12	12	24	12	15	21
D13	24	15	15	21	0
M-1	21	12	24	12	18
M-2	15	27	27	24	30
M-3	30	18	18	27	15
M-6	0	0	0	0	0
M-12	0	0	0	0	0

Tabela 43. Aplicando o peso às classificações do método PCAM

Entradas	TIOCSFNRJT	TIOCSFNRJPJ	SOCSFNRJPJ	SOCSFNRJTot	SOCSFNRJPF
MM3	0	0	0	0	0
MM2	0	0	0	0	0
MM6	4	4	4	4	4
MM12	5	5	5	5	5
D12	6	0	0	6	6
D13	7	7	0	8	7
M-1	9	0	10	10	10
M-2	0	9	6	0	0
M-3	0	6	8	0	0
M-6	8	8	7	7	8
M-12	10	10	9	9	9

Tabela 44. Aplicando o peso às classificações do método ReliefF

Entradas	TIOCSFNRJT	TIOCSFNRJPJ	SOCSFNRJPJ	SOCSFNRJTot	SOCSFNRJPF
MM3	10	12	10	12	10
MM2	14	14	16	16	14
MM6	6	6	6	6	4
MM12	4	4	14	8	8
D12	16	18	0	0	0
D13	20	20	12	10	18
M-1	18	16	18	18	16
M-2	12	10	8	14	12
M-3	0	2	4	4	6
M-6	2	0	2	2	2
M-12	8	8	20	20	20

Tabela 45. Aplicando o peso às classificações do método SIE

Entradas	TIOCSFNRJT	TIOCSFNRJPJ	SOCSFNRJPJ	SOCSFNRJTot	SOCSFNRJPF
MM3	10	10	7	10	7
MM2	8	8	6	8	6
MM6	0	0	0	0	0
MM12	0	0	0	0	0
D12	4	7	9	7	9
D13	5	5	5	5	4
M-1	6	6	10	6	10
M-2	7	9	8	9	8
M-3	9	4	4	4	5
M-6	0	0	0	0	0
M-12	0	0	0	0	0

Na Tabela 46 é apresentado a nova ordenação para as variáveis MM3, MM2, MM6, MM12, D12, D13, M-1, M-2, M-3, M-6 e M-12 considerando os 4 métodos de seleção de variáveis. Os valores da nova ordenação correspondem às somas das Tabelas 42, 43, 44, e 45 para cada variável.

Tabela 46. Soma dos pesos para todas as variáveis

Entradas	TIOCSFNRJT	TIOCSFNRJPJ	SOCSFNRJPJ	SOCSFNRJTot	SOCSFNRJPF
MM3	47	43	47	40	44
MM2	40	52	43	54	44
MM6	10	10	10	10	8
MM12	9	9	19	13	13
D12	38	49	21	28	36
D13	56	47	32	44	29
M-1	54	34	62	46	54
M-2	34	55	49	47	50
M-3	39	30	34	35	26
M-6	10	8	9	9	10
M-12	18	18	29	29	29

São selecionadas, para cada variável exógena, as sete variáveis com maiores valores de pesos. Na Tabela 47, são apresentados as variáveis de entrada exógenas selecionadas.

Tabela 47. Nova ordenação das variáveis de entradas exógenas

Entradas	TIOCSFNRJT	TIOCSFNRJPJ	SOCSFNRJPJ	SOCSFNRJTot	SOCSFNRJPF
MM3	X	X	X	X	X
MM2	X	X	X	X	X
D12	X	X			X
D13	X	X	X	X	X
M-1	X	X	X	X	X
M-2	X	X	X	X	X
M-3	X	X	X	X	
M-12			X	X	X

5.3.2.3. Identificação das melhores arquiteturas (em média) das variáveis exógenas

CASO 1: Associando-se os métodos PCAM e ReliefF — Para os testes foram utilizadas as entradas escolhidas na Tabela 40. Os resultados das arquiteturas com menores erros MAPE em média (100 inicializações) são apresentados na Tabela 48. A busca pela melhor arquitetura, em média, considerou a combinação das variáveis de entrada, variação do número de neurônios e cem (100) inicializações para cada uma das arquiteturas propostas.

Tabela 48. Melhores arquiteturas exógenas em média associando os métodos PCAM e ReliefF

Variáveis	Entradas	Neurônios	Mape %
TIOCSFNRJT	[MM12, M-1]	6	2,19
TIOCSFNRJPJ	[D13, M-2, M-12]	1	4,11
SOCSFNRJPJ	[M-2, M-12]	1	1,39
SOCSFNRJTot	[M-1, M-12]	1	1,33
SOCSFNRJPF	[MM3, M-12]	3	2,52

CASO 2: Utilizando os 4 métodos combinados através de pesos — Para os testes foram utilizadas as entradas escolhidas na Tabela 47. Os resultados da arquitetura com melhor erro MAPE em média (100 inicializações) é apresentado na Tabela 49. A busca pela melhor arquitetura, em média, considerou a combinação das variáveis de entrada, variação do número de neurônios e cem (100) inicializações para cada uma das arquiteturas propostas.

Tabela 49. Melhores arquiteturas exógenas em média para os 4 métodos

Variáveis	Entradas	Neurônios	Mape %
TIOCSFNRJT	[MM3, D12, D13]	6	2,81
TIOCSFNRJPJ	[MM3, MM2, D13, M-2]	1	4,11
SOCSFNRJPJ	[M-1, M-3, M-12]	1	1,29
SOCSFNRJTot	[MM2, M-1, M-2, M-3, M-12]	1	0,75
SOCSFNRJPF	[MM3, M-2, M-12]	3	2,00

5.3.2.4. Busca dos melhores pesos das variáveis exógenas

O passo seguinte, segundo a metodologia adotada, é encontrar o melhor conjunto de pesos e bias para as melhores arquiteturas de redes em média obtidas na seção anterior Tabelas 48 e 49 do ano de validação (2010). Para alcançar o objetivo, as melhores arquiteturas em média foram inicializadas 1000 vezes.

5.3.3. Identificação da melhor arquitetura conjunta (endógena e exógena) em média

Nesta seção são apresentados os resultados da identificação da melhor arquitetura conjunta de variáveis de entrada (endógena e exógena). Para lograr o objetivo foram realizados 2 testes:

1. A primeira avaliação identifica a melhor arquitetura conjunta considerando as variáveis de entrada endógenas do ISSQN da Tabela 31 (métodos PCAM-ReliefF), e para as variáveis exógenas as entradas da Tabela 40.
2. A segunda avaliação identifica a melhor arquitetura conjunta considerando as variáveis de entrada endógena ISSQN da Tabela 31 (métodos PCAM-ReliefF), e as entradas exógenas da Tabela 47 (combinação dos quatro métodos combinados com pesos).

Para dimensionar melhor a extensão dos testes foi avaliado o esforço computacional.

Avaliação do esforço computacional utilizando variáveis conjuntas (endógenas e exógenas) — A Tabela 50 exibe números de combinações de variáveis exógenas do ISSQN considerando que os subconjuntos a serem

avaliados variam de 0 a 5 variáveis de entrada, de um total de cinco (5) variáveis de exógenas.

Tabela 50. Número total de combinações das 5 variáveis exógenas

Número de Entradas da Rede Neural	Sub Conj. de Entradas	Combinação	Conjuntos de Entradas Avaliados
5	0	C_0^5	1
5	1	C_1^5	5
5	2	C_2^5	10
5	3	C_3^5	10
5	4	C_4^5	5
5	5	C_5^5	1
			32

Na Tabela 51, são apresentados os cálculos dos números de combinações de variáveis endógenas do ISSQN (MM3, MM6, MM12, D13, M-1, M-2 e M-12), considerando que os subconjuntos a serem avaliados variam de 1 a 7 variáveis de entrada, de um total de 7 variáveis endógenas.

Tabela 51. Número total de combinações das 7 entradas endógenas

Número de Entradas da Rede	Sub Conj. Entradas	Combinação	Conjuntos de Entradas Avaliados
7	1	C_1^7	7
7	2	C_2^7	21
7	3	C_3^7	35
7	4	C_4^7	35
7	5	C_5^7	21
7	6	C_6^7	7
7	7	C_7^7	1

Na Tabela 52 é apresentado o número total de redes com 7 entradas, subconjuntos variando 1 a 7, considerando 5 variáveis exógenas. Assim, por exemplo, para uma arquitetura de 3 entradas, sendo que o problema pode usar cinco (5) variáveis exógenas e sete (7) variáveis endógenas, tem-se as seguintes configurações de entradas:

- 3 entradas em sete variáveis endógenas $= 35$ combinações de
- 2 entradas em sete endógenas $+ 1$ exógena $= 21$ (combinações de endógenas) $\times 5$ (combinações de exógenas) combinações = 105

- 1 entrada em sete endógena $+ 2$ exógenas $= 7$ (combinações de endógenas) $\times$ 10 (combinações de exógenas) $= 70$.

Ao se somar essas possibilidades tem-se um total de 210 combinações de entradas distintas. Se forem avaliadas redes com números de neurônios na camada escondida variando entre 1 a 6 neurônios, para cada combinação de entrada, e cada uma dessas arquiteturas forem inicializadas cem (100) vezes, será um total de $210 \times 6 \times 100 = 126.000$ redes neurais a serem avaliadas.

Tabela 52. Número de redes com 7 endógenas e 5 exógenas

Número de Combinações de Variáveis Endógenas	Número de Combinações de Variáveis Exógenas	Número de Neurônios	Número de Inicializações	Número de Redes Neurais
7	31*	6	100	130.200
21	32	6	100	403.200
35	32	6	100	672.000
35	32	6	100	672.000
21	32	6	100	403.200
7	32	6	100	134.400
1	32	6	100	19.200
				2.434.200

* Nesse caso como são pelo menos duas entradas não poderia se considerar a possibilidade de não haver pelo menos uma variável de entrada exógena.

Em função da grande quantidade de redes neurais avaliadas. Não se fez maiores investigações combinando um conjunto maior de variáveis.

CASO 1: Associando os métodos PCAM e ReliefF — Antes de realizar a avaliação conjunta para previsão do ISSQN com variáveis endógenas e exógenas, foi feita a previsão dos 12 meses do ano 2010 para as variáveis exógenas, pois no período de validação do modelo, não se consideram conhecidas as variáveis do histórico.

Para as avaliações foram utilizadas as entradas MM3, MM6, MM12, D13, M-1, M-2, M-12, como variáveis endógenas do ISSQN, conforme a Tabela 31, e para a previsão das variáveis exógenas TIOCSFNRJT, TIOCSFNRJPJ, SOCSFNRJTot, SOCSFNRJPJ e SOCSFNRJPF foram utilizadas as entradas exógenas da Tabela 40. Observa-se que a arquitetura apresentada na Figura 26 pode ser desmembrada em redes separadas, visando para reduzir o tempo de computação, a arquitetura foi particionada em 7 partes (entradas endógenas do

ISSQN). O resultado da melhor arquitetura endógena e exógena é apresentado na Tabela 53.

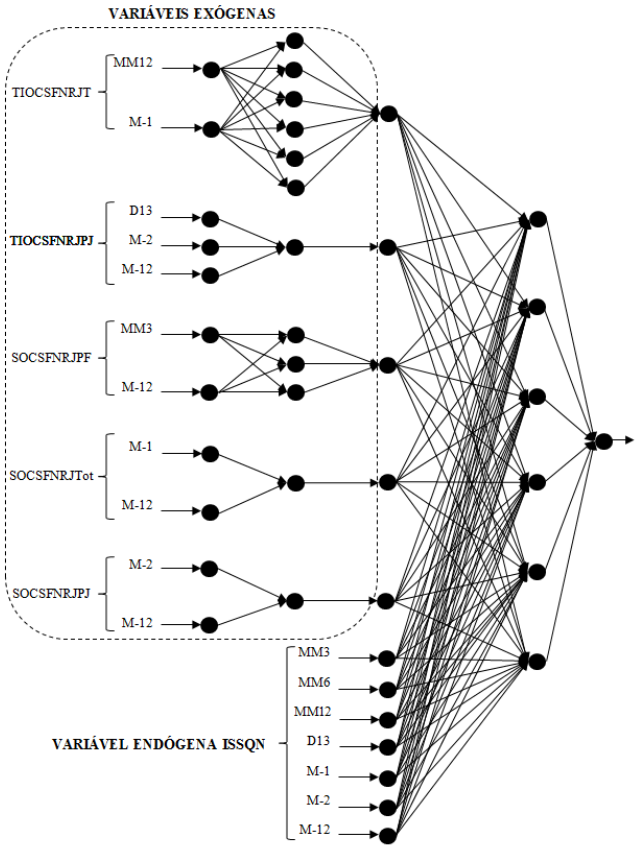


Figura 26. Exemplo de arquitetura de redes neurais considerando todas as variáveis endógenas e exógenas avaliadas

Os resultados (Tabela 53) utilizando variáveis endógenas e exógenas indicou uma pequena melhora dos resultados de previsão (erro MAPE 12,12%), já que utilizando somente variáveis endógenas (Tabela 32) o erro MAPE foi de 12,74%.

Tabela 53. Melhor arquitetura endógena e exógena em média

Variáveis Endógenas	Variáveis Exógenas	Neurônios	MAPE %
[M-12]	TIOCSFNRJPJ SOCSFNRJPJ	4	12,12

CASO 2: Utilizando os 4 métodos combinados através de pesos — Antes de realizar a avaliação conjunta para previsão do ISSQN com variáveis endógenas e exógenas, foi feita a previsão dos 12 meses do ano 2010 para as variáveis

exógenas, pois no período de validação do modelo, não se consideram conhecidas as variáveis do histórico.

Para as avaliações foram utilizadas as entradas MM3, MM6, MM12, D13, M-1, M-2, M-12, como variáveis endógenas do ISSQN, conforme a Tabela 31, e para a previsão das variáveis exógenas TIOCSFNRJT, TIOCSFNRJPJ, SOCSFNRJTot, SOCSFNRJPJ e SOCSFNRJPF foram utilizadas as variáveis de entrada da Tabela 47. Observa-se que a arquitetura apresentada na Figura 27 pode ser desmembrada em redes separadas, visando para reduzir o tempo de computação, a arquitetura foi particionada em 7 partes (entradas endógenas do ISSQN). Os resultados da melhor arquitetura endógena e exógena são apresentados na Tabela 54.

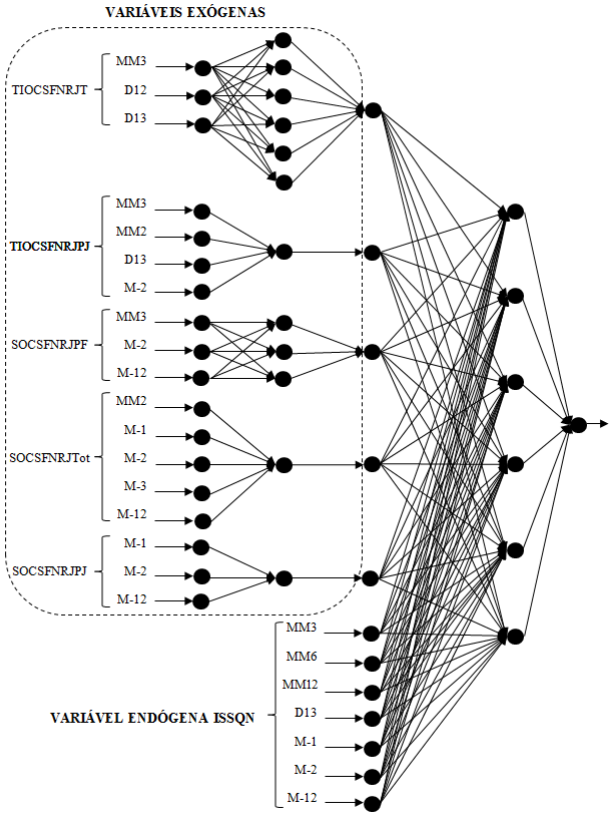


Figura 27. Exemplo de arquitetura de redes neurais considerando todas as variáveis endógenas e exógenas avaliadas

O resultado (Tabela 54) para previsão do ISSQN utilizando variáveis endógenas e exógenas foi melhor em média (11,13%), quando comparado a avaliação considerando os métodos PCAM e ReliefF associando (erro MAPE 12,12%) da Tabela 53.

Tabela 54. Melhor arquitetura endógena e exógena em média

Variáveis Endógenas	Variáveis Exógenas	Neurônios	MAPE %
[M-12]	TIOCSFNRJPJ SOCSFNRJTot SOCSFNRJPJ	6	11,13

Assim, a melhor arquitetura em média possui quatro entradas, uma variável endógena do ISSQN (M-12) e três variáveis exógenas TIOCSFNRJPJ, SOCSFNRJTot e SOCSFNRJPJ, seis neurônios na camada escondida e um neurônio na camada de saída.

5.3.4. Busca dos melhores pesos da melhor arquitetura conjunta

O passo seguinte, segundo a metodologia adotada, é encontrar o melhor conjunto de pesos para melhor arquitetura de rede em média obtida na seção anterior (Tabela 54). Para alcançar o objetivo, a melhor arquitetura em média foi inicializada 1000 vezes.

Antes de testar a melhor arquitetura em média, foram realizadas as previsões do ano 2010 das três variáveis exógenas (TIOCSFNRJPJ, SOCSFNRJTot e SOCSFNRJPJ), cada variável exógena foram testadas com 1000 inicializações e as melhores redes foram escolhidas.

Os resultados da previsão para a melhor arquitetura em média (a que adotou o os 4 métodos combinados com pesos) com os melhores pesos são apresentados na Tabela 55.

Tabela 55. Melhor arquitetura endógena e exógena

Variáveis Endógenas	Variáveis Exógenas	Neurônios	MAPE %
[M-12]	TIOCSFNRJPJ SOCSFNRJTot SOCSFNRJPJ	6	4,86

Na Figura 28 são apresentadas as séries com os valores previstos e reais na melhor arquitetura do ISSQN com os melhores pesos para o ano 2010 (período correspondente aos dados de validação).

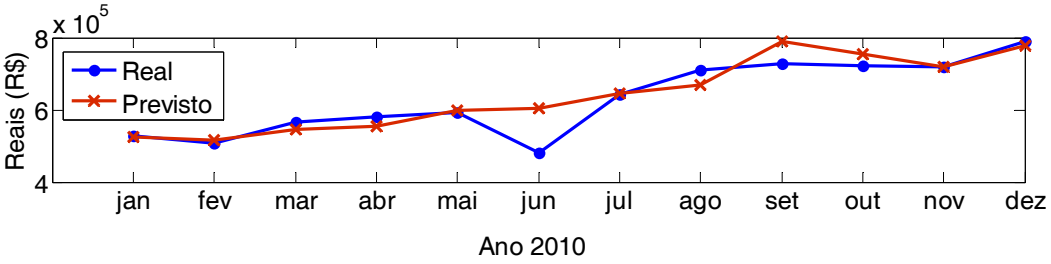


Figura 28. Precisões do ISSQN para o ano 2010

Na Tabela 56, é apresentada a previsão anual do ISSQN realizada pela Secretaria de Fazenda do Município de Araruama (SFMA), que está disponível no site do TCE, e a previsão do modelo proposto nesta dissertação para o ano 2010 (somatório das previsões mensais). Também se incluiu o valor real arrecadado durante o ano de 2010. Os resultados obtidos com o modelo proposto foram melhores em comparação ao valor previsto pela SFMA. Infelizmente não se sabe como foi realizada a previsão apresentada pela SFMA. A princípio, o que sabe é que esta previsão é uma estimativa considerando os valores dos anos anteriores e expectativas baseadas em políticas.

Tabela 56. Comparação da Previsão do ISSQN para o ano 2010

ISSQN (Janeiro a Dezembro)	**Total – 2010**	**Erro Relativo (|Previsto-Real|)/Real**
Real	7.565.208,10	-
Previsão do SFMA	5.967.320,80	21,1%
Previsão do Modelo Proposto	7.701.574,40	1,8%

A seguir foi realizado um Teste (Recall) para prever o ano 2011 com a melhor rede (Tabela 55) para prever o ISSQN. As entradas da rede neural foram construídas com as informações da arrecadação mensal do (ISSQN) ano 2010 (variáveis endógenas) e com os valores previstos das variáveis exógenas para 2011.

Os resultados da previsão do ISSQN para o ano 2011 são apresentados na Tabela 57. Pode-se observar que os resultados com o modelo proposto foi melhor do que a previsão feita pela SFMA.

Tabela 57. Comparação da Previsão do ISSQN para o ano 2011

ISSQN (Janeiro a Dezembro)	**Total – 2011**	**Erro Relativo (|Previsto-Real|)/Real**
Real	8.660.276,10	-
Previsão do SFMA	6.399.849,80	26,1%
Previsão do Modelo Proposto	9.589.742,82	10, 7%

5.4. Identificação de Irregularidades do ISSQN

Neste estudo de caso será realizada a identificação de irregularidades do ISSQN dos contribuintes do Município de Araruama, mediante um grau (0 a 100) que indica a variação do contribuinte com respeito ao grupo, sendo que quanto maior o valor do grau, maior mudança no seu perfil de emissão do ISSQN e Número de empregados. Para a identificação de irregularidades será aplicado o modelo proposto na Figura 12.

A seguir descreve-se toda a metodologia usada no processo de identificação de irregularidades do ISSQN.

5.4.1. Pré-Processamento

Para o estudo caso, foi realizado o cruzamento das bases de dados do Município de Araruama e do CAGED dos anos 2006 e 2007, do cruzamento foram extraídos dois registros de 3.798,00 mensais (janeiro a dezembro) dos contribuintes para os anos 2006 e 2007.

Após o cruzamento dos dados das bases CAGED e ISSQN, agrupamento dos dados por ano e, finalmente, filtragem dos dados segundo a Tabela 11, foram identificados 334 registros dos contribuintes para os anos 2006 e 2007.

5.4.2. Clusterização dos contribuintes

Em problemas de clusterização além da base de dados é necessário conhecer o número de clusters (*k*). Para determinar o número de clusters (*k*) foram utilizados os métodos Silhueta e PBM. Na Tabela 58 são apresentados os índices dos métodos Silhueta e PBM, para número de clusters (*k*) variando de 2 a 10. Pode-se observar que para os dois métodos o maior índice é para *k* igual a 2.

Para poder classificar contribuintes em suspeitos e não suspeitos é necessário contar com maior número de clusters (*k*) e não somente 2. Em problemas de identificação de irregularidades do ISSQN, quando se utiliza maior número de clusters (*k*) se têm maior sensibilidade na classificação dos contribuintes em suspeitos e não suspeitos. Neste estudo o número de clusters (*k*) foi fixado em 10.

Tabela 58. Número de clusters (*K*)

Num. de clusters *(K)*	**Silhueta**	**PBM**
2	0,92	1393,70
3	0,63	1024,51
4	0,58	707,39
5	0,56	644,16
6	0,55	611,24
7	0,54	508,88
8	0,53	435,45
9	0,50	383,68
10	0,49	334,82

A seguir os dois registros (2006 e 2007) foram agrupados utilizando o *Fuzzy C-Means* (FCM) e número de clusters (*k*) igual a 10. Dessa forma cada contribuinte com o atributo valor ISSQN emitido pertence a um grupo nos anos 2006 e 2007, e cada contribuinte com atributo de número de empregados pertence a um grupo nos anos 2006 e 2007.

Na Tabela 59 são apresentados os centroides dos atributos Emissão do ISSQN e Número de Empregados do ano 2007.

Tabela 59. Centroides dos atributos Emissão de ISSQN e Número de empregados do ano 2007

Grupo	Centróides de Emissão do ISSQN	Centróides de Num. de Empregados
1	158,20	1
2	1.884,12	4
3	4.076,59	10
4	7.637,54	18
5	16.012,94	28
6	27.239,25	47
7	43.082,99	67
8	76.879,10	81
9	106.363,07	196
10	212.084,69	979

5.4.3. Classificação dos contribuintes

A seguir são calculadas as matrizes de relações R_1 (atributo Emissão do ISSQN) e R_2 (número de empregados), para o calculo foram utilizados os centroides do ano 2007 da Tabela 59 e as relações R_1 e R_2 da seção 4.3.3.

$$R_1 = \begin{bmatrix} 1{,}00 & 0{,}08 & 0{,}04 & 0{,}02 & 0{,}01 & 0{,}01 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 11{,}91 & 1{,}00 & 0{,}46 & 0{,}25 & 0{,}12 & 0{,}07 & 0{,}04 & 0{,}02 & 0{,}02 & 0{,}01 \\ 25{,}77 & 2{,}16 & 1{,}00 & 0{,}53 & 0{,}25 & 0{,}15 & 0{,}09 & 0{,}05 & 0{,}04 & 0{,}02 \\ 48{,}28 & 4{,}05 & 1{,}87 & 1{,}00 & 0{,}48 & 0{,}28 & 0{,}18 & 0{,}10 & 0{,}07 & 0{,}04 \\ 101{,}22 & 8{,}50 & 3{,}93 & 2{,}10 & 1{,}00 & 0{,}59 & 0{,}37 & 0{,}21 & 0{,}15 & 0{,}08 \\ 172{,}18 & 14{,}46 & 6{,}68 & 3{,}57 & 1{,}70 & 1{,}00 & 0{,}63 & 0{,}35 & 0{,}26 & 0{,}13 \\ 272{,}34 & 22{,}87 & 10{,}57 & 5{,}64 & 2{,}69 & 1{,}58 & 1{,}00 & 0{,}56 & 0{,}41 & 0{,}20 \\ 485{,}97 & 40{,}80 & 18{,}86 & 10{,}07 & 4{,}80 & 2{,}82 & 1{,}78 & 1{,}00 & 0{,}72 & 0{,}36 \\ 672{,}34 & 56{,}45 & 26{,}09 & 13{,}93 & 6{,}64 & 3{,}90 & 2{,}47 & 1{,}38 & 1{,}00 & 0{,}50 \\ 1.340{,}62 & 112{,}56 & 52{,}03 & 27{,}77 & 13{,}24 & 7{,}79 & 4{,}92 & 2{,}76 & 1{,}99 & 1{,}00 \end{bmatrix}$$

$$R_2 = \begin{bmatrix} 1{,}00 & 0{,}16 & 0{,}06 & 0{,}03 & 0{,}02 & 0{,}01 & 0{,}01 & 0{,}01 & 0{,}00 & 0{,}00 \\ 6{,}22 & 1{,}00 & 0{,}37 & 0{,}21 & 0{,}13 & 0{,}08 & 0{,}05 & 0{,}05 & 0{,}02 & 0{,}00 \\ 16{,}96 & 2{,}73 & 1{,}00 & 0{,}56 & 0{,}36 & 0{,}21 & 0{,}15 & 0{,}12 & 0{,}05 & 0{,}01 \\ 30{,}29 & 4{,}87 & 1{,}79 & 1{,}00 & 0{,}64 & 0{,}38 & 0{,}27 & 0{,}22 & 0{,}09 & 0{,}02 \\ 47{,}44 & 7{,}63 & 2{,}80 & 1{,}57 & 1{,}00 & 0{,}60 & 0{,}42 & 0{,}35 & 0{,}14 & 0{,}03 \\ 79{,}22 & 12{,}73 & 4{,}67 & 2{,}62 & 1{,}67 & 1{,}00 & 0{,}70 & 0{,}58 & 0{,}24 & 0{,}05 \\ 113{,}64 & 18{,}26 & 6{,}70 & 3{,}75 & 2{,}40 & 1{,}43 & 1{,}00 & 0{,}83 & 0{,}34 & 0{,}07 \\ 137{,}35 & 22{,}08 & 8{,}10 & 4{,}53 & 2{,}89 & 1{,}73 & 1{,}21 & 1{,}00 & 0{,}41 & 0{,}08 \\ 331{,}70 & 53{,}31 & 19{,}56 & 10{,}95 & 6{,}99 & 4{,}19 & 2{,}92 & 2{,}41 & 1{,}00 & 0{,}20 \\ 1.652{,}60 & 265{,}61 & 97{,}44 & 54{,}55 & 34{,}83 & 20{,}86 & 14{,}54 & 12{,}03 & 4{,}98 & 1{,}00 \end{bmatrix}$$

Em seguida é calculada a matriz de variação μ utilizando a eq. (4.5), onde cada elemento da matriz R_1 é somada com cada elemento da matriz R_2, para formar a matriz μ, a qual contem todas as possíveis combinações entre as

relações dos centroides dos grupos formados pelas emissões de ISSQN e pelo número de empregados, de dimensão 100 x 100.

$$\mu = \begin{bmatrix} 1{,}00 + R_2 & 0{,}08 + R_2 & \cdots & 0{,}00 + R_2 \\ 11{,}91 + R_2 & 1{,}00 + R_2 & \cdots & 0{,}01 + R_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1.340{,}62 + R_2 & 112{,}56 + R_2 & \cdots & 1{,}00 + R_2 \end{bmatrix}_{100 \text{ x } 100}$$

Finalmente, na Tabela 60 são apresentados os Grau de Irregularidades de 17 contribuintes (coluna 6) em ordem de maior a menor, para o cálculo foi utilizada a eq. (4.6). Nas colunas 2 e 3, são apresentados os valores da Emissão do ISSQN dos contribuintes para os anos 2006 (T) e 2007 $(\mathrm{T}+1)$. O Número de Empredados para os anos 2006 e 2007, são apresentados nas colunas 4 e 5.

Tabela 60. Suspeitos de Irregularidades - Graus

	ISSQN		**Num. Empregados**		
Colocação	**2006**	**2007**	**2006**	**2007**	**Grau**
1°	48.777,76	0,00	36	46	**5,79**
2°	48.156,59	984,72	0	0	**5,74**
3°	21.706,66	0,00	8	8	**3,42**
4°	6.675,48	0,00	2	2	**1,07**
5°	6.937,85	0,00	12	20	**0,93**
6°	8.779,86	224,94	4	4	**0,91**
7°	6.827,35	0,00	2	2	**0,91**
8°	4.999,57	0,00	13	10	**0,90**
9°	4.984,69	0,00	7	2	**0,88**
10°	3.341,73	807,47	1	3	**0,61**
11°	1.515,50	1.004,92	1	3	**0,61**
12°	2.008,58	761,00	0	2	**0,61**
13°	586,22	2.331,85	0	10	**0,57**
14°	4.281,99	0,00	5	5	**0,43**
15°	4.028,50	0,00	4	4	**0,43**
16°	3.859,77	0,00	2	2	**0,43**
17°	2.496,51	892,69	3	3	**0,43**

6
Conclusões e recomendações

6.1.
Conclusões

Lembrando os objetivos deste trabalho, exposto no capítulo 1, e os resultados apresentados no capítulo 5, pode-se afirmar que os métodos se seleção de variáveis (PCAM, LSE, SIE e ReliefF) apresentam apenas, uma noção de quais variáveis de entrada das redes têm mais influência na previsão, Nessa dissertação as classificações dos métodos foram avaliadas e selecionados os métodos que agregaram valor ao modelo de previsão, o qual foi bastante benéfico, propiciando melhorias (diminuição do erro MAPE) nas previsões de receitas, conforme nas Tabelas 23, 30 e 39.

Um fato realmente marcante foi a inclusão de variáveis exógenas relevantes no modelo de previsão de ISSQN, Essa inclusão trouxe melhorias consideráveis (diminuição do erro MAPE) na previsão, conforme mostrado na Tabela 54.

Os resultados obtidos com a aplicação dos modelos propostos para a previsão de Receitas Tributárias e ISSQN, foram bem melhores em comparação com os resultados de previsões de receitas disponibilizados no site do TCE (2011), Infelizmente não se tem conhecimento a respeito dos métodos ou modelos utilizados no site do TCE. Informalmente, o que se pode dizer é que as previsões não usam métodos mais elaborados.

Salienta-se que o intervalo de tempo das séries históricas utilizadas para a previsão (2004 a 2010) pode ter prejudicado em muito os resultados. Usar séries históricas maiores para treinamento e validação, normalmente acarretam em melhora dos resultados.

Em relação à identificação de irregularidades no pagamento de ISSQN, foi realizado o cruzamento de bases de dados do Município e do CAGED. O uso de informação sobre o número de empregados foi importante já que ampliou a busca de indícios de irregularidade. Os resultados obtidos (Tabela 60) mediante um grau de 0 a 100 faculta aos fiscalizadores o planejamento das diligências a serem realizadas, priorizando os contribuintes com maiores indícios de

irregularidades (erro ou fraude). E importante destacar que a falta de informações históricas sobre resultados de fiscalizações, ou seja, informações sobre contribuintes fiscalizados e os resultados dessa fiscalização, não permitiu o uso de modelos supervisionados, normalmente com resultados mais robustos comparados aos modelos não supervisados, ajudando inclusive nas escolhas das variáveis exógenas a serem consideradas.

6.2. Trabalhos futuros

Como ideias de possíveis experiências a serem futuramente empreendidas, tendo como subsídio o presente trabalho, pode-se citar:

Espera-se que o modelo proposto para previsão de receitas utilizando variáveis endógenas e exógenas, possa também ser aproveitado para melhorar (diminuição do erro MAPE) a qualidade das previsões de outras receitas como IPTU, Receitas Tributárias, Receitas da Dívida Ativa e, também possa ser aplicado a outros municípios. Mais especificamente também é possível avaliar outras variáveis exógenas e históricas de dados maiores.

Em relação à identificação de irregularidades do ISSQN, espera-se que o cruzamento de bases de dados executada nessa dissertação, possa ser ampliado com a base de dados do tomador de serviços e dados sobre a escrituração fiscal do contribuinte, dessa forma se teria maior informação para a busca de indícios de irregularidades do ISSQN.

Referências bibliográficas

ALBUQUERQUE, B. D. **Receita municipal: experiência e perspectivas, Sonegação, Fraudes e Evasão Fiscal.** ANFIP, Belém-Brasil, VIII: p. 64, 1998.

AMARAL, G. L. D. **Estudo Sobre Sonegação Fiscal das Empresas Brasileiras:** IBPT – Instituto Brasileiro de Planejamento Tributário, 2009.

ANDRADE, H. D. S. **Um processo de mineração de dados aplicados ao combate à sonegação fiscal do ICMS.** Dissertação de Mestrado, Computação Aplicada, Universidade Estadual do Ceará – UECE, 2009.

ARAÚJO, T. S.; OLIVEIRA, P. S. D.; SILVA, E, R, G, D, **Sistemas Inteligentes de Apoio à Tomada de Decisão na Gestão Pública Municipal: Uma Abordagem Conceitual.** IV Conferência Sul-Americana em Ciência e Tecnologia Aplicada ao Governo Eletrônico (IV CONeGOV), Palmas-Brasil, 2007.

BARRON, A. R. **Universal aproximation bound for superpositions of a sigmoidal function.** p. 930 - 945, 1993.

BEZDEK, J. C. **A convergence theorem for the fuzzy c-means clustering algorithms.** IEEE Trans, PAMI: p. 1-8, New York – USA, 1981.

BOURKE, P. **AutoCorrelation - 2D Pattern Identification.** Disponível em: <http://paulbourke.net/miscellaneous/correlate>, Acesso em: 5 nov. 2010.

BREMAEKER, F. E. D. **As Receitas Tributárias Municipais em 2010.** Transparência Municipal, p. 34, Salvador – Brasil, 2011.

CALDART, W. L. **Modelo de previsão de arrecadação do ISSQN para o Município de Caxias do Sul.** Instituto de Pesquisas Econômicas e Sociais - IPES, p. 15, 2006.

CAMPOS, C. V. **Previsão da arrecadação de receitas federais.** Dissertação de Mestrado, Faculdade de Economia, Universidade de São Paulo - USP Ribeirão Preto – Brasil. 2009.

CAO, Y. **Control structure selection for chemical processes using input-output controllability analysis.** (PhD), Exeter, England, 1995.

CAO, Y.; BISS, D. **An extension of singular value analysis for assessing manipulated variable constraints.** Journal of Process Control, p. 37-44, 1996.

CAO, Y.; ROSSITER, D. **An input pre-screening technique for control Structure Selection.** Computers Chem, Elsevier Science Ltd, p. 563-569, Great Britain, 1997.

CARDOSO, R. M. **Imposto sobre serviços: O regime de tributação diferenciado das sociedades de profissionais.** Faculdade de Direito da Pontifícia Universidade Católica do Rio Grande do Sul – PUCRS, p. 27, Rio Grande do Sul – Brasil, 2012.

CARVALHO, P. D. **Curso de direito tributário.** ISBN 9788502087699, São Paulo – Brasil, 2010.

CASTRO, A. L. **Dívida Ativa Tributária: Aspectos Legais e Considerações Sobre a Dívida Ativa Tributária do Estado da Bahia.** Universidade Federal da Bahia – UFBA, p. 11, Bahia – Brazil, 2005.

CHATFIELD, C. **The Analysis of Time Series: An Introduction.** ISBN 978-1584883173, 2003.

CHUNG, F. L.; DUAN, J**. On Multistage Fuzzy Neural Network Modeling.** IEEE Transactions on Fuzzy Systems, The Hong Kong Polytechnic University, p. 125-142, 2000.

CONTRERAS, J. C. S. **Previsão de arrecadação do ICMS através de Redes Neurais no Brasil.** Dissertação de Mestrado, Universidade Federal de Pernambuco - UFPE, Brasil, 2005.

CONTRERAS, R. J. **Técnicas de Seleção de Características aplicadas a Modelos Neuro-Fuzzy Hierárquicos BSP.** Dissertação de Mestrado, Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro - PUC-rio, 2002.

CYBENKO, G. **Approximation by superposition of a sigmoidal function.** Springer-Verlag New York Inc., Mathematics of Control, Signals, and Systems, p. 303-314, 1989.

DANH, N. T.; PHIEN, H. N.; GUPTA, A. D. **Neural network models for river flow forecasting: Asian Institute of Technology,** Thailand, 1999.

DASH, M.; LIU, H.; YAO, J. **Dimensionality Reduction for Unsupervised Data.** In Ninth IEEE International Conference on Tools with AI, ICTAI'97, Newport Beach, CA, USA, 1997.

DOEHLEN, A. M. A**. Arrecadação Tributária de Belo Horizonte, Sonegação, Fraudes e Evasão Fiscal.** ANFIP, p. 76, Belo Horizonte - Brasil, 1997.

FAVATTO, C. S. **As principais formas de fraude, sonegação e evasão, Sonegação, Fraudes e Evasão Fiscal.** ANFIP, p. 64, Belém-Brasil, VIII, 1998.

FAYAL, M. A. **Previsão de Vazão por Redes Neurais Artificiais e Transformadas Wavelet.** Dissertação de Mestrado, Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio, 2005.

FERREIRA, R. D. S**. Crimes contra a ordem tributária.** 2 ed., Malheiros, ISBN 8574203610, São Paulo – Brasil, 2002.

FILHO, A. P.; APPY, B.; LEVY, J. V. **Dívida Ativa: Manual de Procedimento Aplicado à União e aos Estados, Distrito Federal e Municípios.** Secretaria do Tesouro Nacional, Coordenação Geral de Contabilidade, p. 44, Brasília, 2004.

FILHO, E. O. A. **Auditoria de Impostos e Contribuições.** ATLAS, ISBN 9788522446544, 2005.

GUARAGNA, P.; MELLO, M. **Um modelo de previsão da arrecadação do ICMS.** Governo do estado do Rio Grande do Sul, Secretaria da Fazenda, Porto Alegre – Brasil, 2002.

HALL, C. **Mineração de Dados: Laboratório de Inteligência Computacional – ICA,** p.71, 2005.

HAYKIN, S. **Neural Networks: a comprehensive foundation.** United States of America: Tom Robbins, ISBN 0-13-273350-1, 1999.

JÈZE, E. G. **O fato gerador do imposto.** RDA 2-I/50, 1937.

KHAIR, A. A. **Lei de Responsabilidade Fiscal: Guia de Orientação para as Prefeituras.** Ministério do Planejamento Orçamento e Gestão, Brasília – Brasil, 2010.

KIRA, K.; RENDELL, L. A. **A Practical Approach to Feature Selection.** ACM, ML92 Proceeding of the ninth international workshop on Machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1992.

KISI, Ö. **Daily River Flow Forecasting Using Artificial Neural Network and Auto-Regressive Models.** Turkish Journal of Engineering & Environmental Sciences, v. 29, 2005.

KONONENKO, I. **Estimating Attributes: Analysis and Extensions of RelefF.** In: SPRINGER-VERLAG NEW YORK, ECML-94 Proceedings of the European conference on machine learning on Machine Learning, p. 171-182, USA, 1994.

LEVENBERG, K. **A Method for the Solution of Certain Non-Linear Problems in Least Squares.** Quarterly of Applied Mathematics 2: p. 164-168, 1944.

LIEBEL, M. J. **Previsão de receitas tributárias: O caso do ICMS no estado do Paraná.** Dissertação de Mestrado, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre – Brasil, 2004.

MACHADO, H. D. B. **Curso de direito tributário.** Malheiros, ISBN 8574293300, São Paulo-Brasil, 2002.

MARQUARDT, D. W. **An Algorithm for Least-Squares Estimation of Nonlinear Parameters.** Society for Industrial and Applied Mathematics, v. 11(2), p. 11, 1963.

MARTINS, S. P. **Manual do imposto sobre serviços.** ISBN 9788522458899, v. 8, 2010.

MCCULLOCH, W. S.; PITTS, W. **A logical calculus of the ideas immanent in nervous activity.** Bulletin of Mathematical Biophysics, p. 115-133, 1943.

MELO, J. E. **ISS: Aspectos Teóricos e Práticos**, 4ta, Dialética, ISBN 85-7500-128-0, São Paulo – Brasil. 2005.

NOGUEIRA, A. A.; SATO, S. A. **Dívida Ativa na Gestão Pública**. Congresso Brasileiro de Contabilidade, Gramado-RS, 2008.

PAIVA, V. **Tributação, arrecadação e política fiscal** – **Um estudo de caso: O programa de modernização da administração tributária do Município de Vitória.** Dissertação de Mestrado, Ciências contábeis, Fundação Instituto Capixaba de Pesquisas em Contabilidade, Economia e Finanças – FUCAPE, p. 145, Vitória-Brasil, 2005.

PAKHIRA, M. K.; BANDYOPADHYAY, S.; MAULIK, U**. Validity index for crisp and fuzzy clusters.** Pattern Recognition, ISSN 0031-3203. Disponível em: < http://www.sciencedirect.com/science/article/pii/S0031320303002838 >, v. 37, p. 487-501, 2004, Acesso em: 22 dez. 2010.

PEREIRA, J. R. **Previsão de receita do ISSQN de Teresina: Uma abordagem com séries temporais.** Dissertação de Mestrado, Economia, Universidade Federal do Ceará - UFC, Fortaleza-Brasil, 2007.

QUEIROZ, C. R. G. **O combate à sonegação em Minas Gerais, Sonegação, Fraudes e Evasão Fiscal.** ANFIP, p. 76, Minas Gerais - Brasil, 1997.

ROUSSEEUW, P. J. **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.** Journal of Computational and Applied Mathematics, Elsevier Science, p. 53-65, 1987.

SAINI, L. M.; SONI, M. K. **Artificial neural network based peak load forecasting using Levenberg-Marquardt and quasi-Newton.** IEE, Proceding on generation transmision, distribution, p. 578-584, 2002.

SANTORO, D. M. **Sobre o Processo de Seleção de Subconjuntos de Atributos - As Abordagens Filtro e Wrapper.** Dissertação de Mestrado, Ciência da Computação, Universidade Federal São Carlos – UFSCar, 2005.

SILVA, I. N. D.; SPATTI, D. H.; FALUUZINO, R. A. **Redes Neurais Artificiais para engenharia e ciências aplicadas.** ISBN 978-85-88098-53-4, São Paulo, 2010.

SMANIO, G. P. **Sonegação fiscal, Direito Penal.** São Paulo – Brasil, 2005.

SOUZA, E. C. F. **A Fraude à Lei no Direito Tributário Brasileiro.** Fiscosoft, São Paulo – Brazil, 2003.

STOCK, J. H.; WATSON, M. W. **EViews 6: Introduction to Econometrics, Printed in the United States of America.** Quantitative Micro Software, 2007.

SÁNCHEZ, J. E. R. **Sistema de Monitoramento de Múltiplos Sensores por Redes Neurais Auto-Associativas e Lógica Fuzzy.** Dissertação de Mestrado, Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio, Rio de Janeiro – Brasil, 2011.

T.C.E. Tribunal de Contas do Estado de Rio de Janeiro. **Relatório resumido da execução orçamentária do Município de Araruama.** Brasil, 2011.

VENDRAMIN, L.; CAMPELLO, R. J. G.; HRUSCHKA, E. R. **On the Comparison of Relative Clustering Validity Criteria.** In: (SIAM), Proceedings of the Ninth SIAM International Conference on Data Mining, USA, 2009.

ZADEH, L. A. **Fuzzy Sets, Information and control.** Department of Electrical Engineering and Electronics Research Laboratory, p. 16, California - USA, 1965.

APÊNDICE A: Séries Históricas

1. Séries Endógenas

1.1. Imposto Sobre Serviços de Qualquer Natureza – ISSQN, do Município de Araruama.
Fonte: Tribunal de Contas do Estado do Rio de Janeiro
http://www.tce.rj.gov.br/main.asp?View={85778E0F-BCBD-4A61-A1BB-572BB20F8ABB}
No campo Município, coloque Araruama, Selecione o ano da série. E no poder coloque Executivo.

Tabela A1: Série do ISSQN do Município de Araruama - TCE

	2005	2006	2007	2008	2009	2010
Jan.		334.821,8	407.379,0	433.613,6	327.423,3	527.328,0
Fev.		325.495,4	390.225,5	350.758,1	371.951,0	506.762,3
Mar.	307.506,3	368.021,8	344.165,3	318.476,1	454.132,4	565.611,7
Abr.	272.645,9	286.439,7	353.274,5	469.274,9	330.920,0	582.289,4
Mai.	327.360,0	351.482,8	305.111,8	548.322,2	537.756,2	591.498,1
Jun.	327.360,0	304.908,1	331.949,6	363.745,2	543.325,5	480.303,0
Jul.	348.361,8	332.479,6	267.178,7	339.897,7	569.560,2	642.055,0
Ago.	415.223,6	402.876,5	278.286,0	499.974,7	570.446,4	709.687,0
Set.	338.808,6	256.228,8	343.610,2	348.550,2	232.391,2	727.526,1
Out.	350.665,2	379.204,4	379.008,6	343.264,7	726.314,6	722.628,4
Nov.	358.748,1	325.780,0	381.049,4	382.119,5	497.134,7	719.590,6
Dez.	470.067,1	316.009,0	362.354,5	929.968,2	764.431,5	789.928,5

1.2. Imposto Sobre Serviços de Qualquer Natureza – ISSQN, do Município de Araruama.
Fonte: Prefeitura Municipal de Araruama.

Tabela A2: Série do ISSQN do Município de Araruama

	2004	2005	2006	2007	2008	2009
Jan.	145.867,2	239.294,3	334.821,8	407.379,0	433.613,6	296.843,4
Fev.	179.200,4	291.962,1	325.495,4	390.225,5	350.758,1	325.016,6
Mar.	246.119,7	307.506,3	368.021,8	344.165,3	318.476,1	421.868,9
Abr.	147.502,6	272.645,9	286.439,7	353.274,5	469.274,9	298.063,1
Mai.	158.885,3	328.056,6	351.482,8	305.111,8	548.322,2	488.063,0
Jun.	153.807,2	355.540,8	304.908,1	331.949,6	363.745,2	315.564,3
Jul.	187.244,7	348.361,8	332.479,6	267.178,7	339.897,7	506.428,4
Ago.	171.877,7	415.223,6	402.876,5	278.286,0	499.974,7	531.683,4
Set.	148.998,6	338.808,6	256.228,8	343.610,2	348.550,2	366.450,7
Out.	323.723,2	350.665,2	379.204,4	379.008,6	343.264,7	668.899,3
Nov.	326.260,2	358.748,1	325.780,0	381.049,4	386.038,5	484.962,1
Dez.	284.392,2	470.067,1	316.009,0	362.354,5	755.014,9	700.435,0

1.3. Receitas Tributárias do Município de Araruama.
Fonte: Tribunal de Contas do Estado do Rio de Janeiro - TCE.
http://www.tce.rj.gov.br/main.asp?View={85778E0F-BCBD-4A61-A1BB-572BB20F8ABB}
No campo Município, coloque município de Araruama. Selecione o ano da série. E no campo poder, coloque Executivo.

Tabela A3: Série da Receita Tributárias do Município de Araruama

	2005	2006	2007	2008	2009	2010
Jan.		4.345.837,7	4.990.624,9	5.733.237,4	6.016.530,1	2.919.994,8
Fev.		1.684.251,2	2.024.894,9	1.865.116,4	1.586.402,4	4.993.743,0
Mar.	1.410.181,6	1.791.308,3	1.667.330,0	1.394.176,7	1.951.326,6	2.586.732,1
Abr.	1.002.670,6	1.001.794,6	1.195.401,6	1.304.362,0	1.253.891,0	1.454.346,5
Mai.	1.011.631,8	1.191.054,4	1.242.954,1	1.468.328,7	1.573.768,7	1.806.714,7
Jun.	1.011.631,8	1.054.547,6	1.102.956,6	1.180.501,0	1.556.406,4	1.454.724,2
Jul.	953.634,2	1.110.516,5	1.096.585,6	1.303.932,0	1.642.421,9	2.126.392,1
Ago.	1.027.560,7	1.166.598,1	1.185.424,3	1.364.715,0	1.543.120,8	2.003.440,1
Set.	859.245,2	990.480,1	1.062.857,7	1.259.200,1	1.075.017,3	1.847.957,0
Out.	715.095,0	869.171,2	1.060.719,8	1.104.675,0	1.286.106,1	1.245.436,7
Nov.	821.095,8	802.909,1	935.940,5	952.495,4	1.154.733,7	1.397.994,6
Dez.	1.616.091,6	1.276.236,6	1.539.420,0	2.184.895,1	2.196.384,5	2.728.024,5

1.4. Receitas da Dívida Ativa do Município de Araruama.
Fonte: Prefeitura Municipal de Araruama.

Tabela A4: Receitas da Dívida Ativa do Município de Araruama

	2004	2005	2006	2007	2008	2009	2010
Jan.	179.010,4	493.432,0	398.454,7	466.747,7	601.879,6	309.008,6	420.930,3
Fev.	135.937,6	284.142,4	395.842,0	368.298,2	461.171,2	275.376,7	449.518,2
Mar.	203.990,0	277.334,2	372.240,5	381.700,3	328.945,2	388.027,2	569.837,7
Abr.	152.469,1	257.426,6	244.092,9	344.673,3	293.584,3	318.885,7	414.992,3
Mai.	175.239,7	236.293,5	326.376,1	339.185,9	292.921,4	359.750,5	475.986,5
Jun.	176.084,9	220.472,3	230.641,0	252.250,2	360.985,0	417.549,0	469.590,4
Jul.	162.152,7	204.782,3	296.271,6	334.857,3	381.072,4	486.929,5	465.237,6
Ago.	190.273,6	347.828,0	292.828,6	273.661,3	403.985,0	464.020,2	478.900,4
Set.	134.968,2	203.104,2	201.530,6	222.382,5	454.499,0	407.264,4	471.555,3
Out.	149.998,0	225.724,3	278.217,1	328.861,6	280.846,7	394.948,1	409.548,9
Nov.	205.329,2	226.419,9	265.258,7	345.338,7	270.906,1	412.846,4	462.461,8
Dez.	182.057,2	351.002,2	391.340,6	540.654,4	609.095,2	597.902,2	632.687,0

2. Séries Exógenas Socio-econômicas

As séries do 2.1 ao 2.7 foram obtidas do Banco Central do Brasil – BCB.
https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries
No campo por código, insira o código correspondente à série. Será exibida a série. Selecione a série e clique em Consultar série. No campo Período informe a data inicial e final da série. Selecione Visualizar valores.

2.1. Saldo das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Pessoas Físicas - R$ (milhões) – SOCSFNRJPF (14020).

Tabela A5: Série do SOCSFNRJPF

	2004	2005	2006	2007	2008	2009	2010
Jan.	8.383	11.301	15.156	20.027	28.218	36.098	45.051
Fev.	8.514	11.491	15.419	20.402	28.604	36.109	45.489
Mar.	8.718	11.988	15.902	20.777	29.698	36.813	46.656
Abr.	8.885	12.249	16.249	21.715	30.901	37.541	47.558
Mai.	9.090	12.636	16.754	22.393	31.726	38.666	48.543
Jun.	9.104	12.970	17.039	23.029	32.502	39.389	49.037
Jul.	9.301	13.232	17.377	23.767	33.305	40.219	49.648
Ago.	9.597	13.679	17.976	24.592	33.931	41.045	50.833
Set.	9.888	13.834	18.228	25.138	34.693	41.977	51.962
Out.	10.098	14.246	18.813	26.046	35.141	42.926	53.777
Nov.	10.695	14.630	19.151	26.597	35.338	43.603	54.493
Dez.	11.035	14.809	19.658	27.214	35.727	44.383	57.494

2.2. Saldo das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Pessoas Jurídicas - R$ (milhões) – SOCSFNRJPJ (14047).

Tabela A6: Série do SOCSFNRJPJ

	2004	2005	2006	2007	2008	2009	2010
Jan.	30.086	31.856	36.626	51.063	62.883	95.207	128.060
Fev.	29.684	31.564	37.531	51.206	63.653	96.258	130.530
Mar.	29.822	31.834	38.077	51.348	66.531	97.820	127.453
Abr.	29.904	32.455	39.201	54.904	69.456	98.499	126.733
Mai.	30.967	32.093	41.596	52.652	75.073	98.787	126.109
Jun.	31.242	32.111	39.394	49.374	76.926	99.342	132.609
Jul.	31.479	32.886	41.284	50.070	77.723	113.644	133.311
Ago.	32.118	33.029	41.708	53.541	81.029	116.715	135.612
Set.	32.996	32.571	41.947	54.571	83.453	120.240	135.997
Out.	33.403	34.317	42.872	57.374	88.829	121.276	137.715
Nov.	32.268	34.951	45.816	59.679	92.210	123.648	139.969
Dez.	31.345	36.306	48.441	62.325	95.546	125.978	138.467

2.3. Saldo das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Total - R$ (milhões) – SOCSFNRJTot (14074).

Tabela A7: Série do SOCSFNRJTot

	2004	2005	2006	2007	2008	2009	2010
Jan.	38.469	43.157	51.782	71.091	91.101	131.305	173.111
Fev.	38.198	43.056	52.950	71.608	92.257	132.367	176.019
Mar.	38.540	43.822	53.979	72.125	96.230	134.633	174.109
Abr.	38.789	44.703	55.450	76.619	100.357	136.040	174.291
Mai.	40.057	44.730	58.350	75.045	106.799	137.453	174.652
Jun.	40.346	45.081	56.433	72.402	109.428	138.732	181.646
Jul.	40.779	46.118	58.660	73.837	111.028	153.863	182.959
Ago.	41.715	46.707	59.684	78.134	114.959	157.760	186.445
Set.	42.884	46.405	60.176	79.709	118.146	162.217	187.959
Out.	43.502	48.563	61.685	83.420	123.970	164.202	191.492
Nov.	42.963	49.581	64.968	86.276	127.548	167.251	194.462
Dez.	42.380	51.114	68.099	89.540	131.273	170.361	195.960

2.4. Taxa de inadimplência das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Pessoas jurídicas % – TIOCSFNRJPJ (15911).

Tabela A8: Série do TIOCSFNRJPJ

	2004	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Jan.	6,17	4,64	4,17	1,29	0,78	0,97	1,26
Fev.	6,49	4,61	4,09	1,14	0,77	1,10	1,21
Mar.	5,79	4,78	4,39	0,99	0,70	1,33	1,14
Abr.	6,59	4,73	4,00	0,92	0,71	1,44	1,01
Mai.	5,69	4,05	1,65	0,94	0,69	1,60	0,97
Jun.	5,40	4,56	1,62	0,97	0,68	1,56	0,97
Jul.	6,25	4,49	1,01	0,98	0,70	1,51	1,02
Ago.	5,74	4,49	1,07	0,92	0,72	1,52	0,99
Set.	5,26	4,46	1,09	0,84	0,81	1,50	0,98
Out.	5,16	4,30	1,30	0,83	0,80	1,56	1,03
Nov.	5,22	4,16	1,27	0,81	0,80	1,52	0,99
Dez.	4,71	4,19	1,21	0,78	0,85	1,33	0,97

2.5. Taxa de inadimplência das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Total (%) – TIOCSFNRJPJT (15943).

Tabela A9: Série do TIOCSFNRJT

	2004	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Jan.	6,45	4,85	4,68	2,53	2,17	2,44	2,49
Fev.	6,66	4,80	4,68	2,41	2,21	2,60	2,40
Mar.	6,02	4,95	4,98	2,29	2,11	2,79	2,34
Abr.	6,61	4,93	4,75	2,21	2,14	2,90	2,25
Mai.	5,87	4,48	3,12	2,34	2,16	3,08	2,24
Jun.	5,63	4,85	3,02	2,34	2,05	3,03	2,16
Jul.	6,27	4,82	2,60	2,39	2,13	2,87	2,19
Ago.	5,80	4,84	2,66	2,30	2,16	2,86	2,13
Set.	5,45	4,88	2,58	2,18	2,17	2,81	2,09
Out.	5,38	4,80	2,78	2,18	2,16	2,86	2,15
Nov.	5,43	4,73	2,71	2,21	2,20	2,78	2,08
Dez.	4,89	4,65	2,52	2,12	2,25	2,57	2,05

2.6. Taxa de inadimplência das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Pessoas físicas % – TIOCSFNRJPF (15879).

Tabela A10: Série do TIOCSFNRJPF

	2004	2005	2006	2007	2008	2009	2010
Jan.	7,46	5,46	5,96	5,86	5,42	6,53	6,19
Fev.	7,24	5,34	6,14	5,75	5,57	6,82	5,98
Mar.	6,80	5,41	6,43	5,65	5,44	6,88	5,78
Abr.	6,65	5,46	6,63	5,61	5,55	6,92	5,71
Mai.	6,46	5,61	6,88	5,78	5,84	7,07	5,70
Jun.	6,40	5,60	6,36	5,44	5,46	6,94	5,56
Jul.	6,33	5,65	6,54	5,53	5,64	6,93	5,51
Ago.	6,00	5,70	6,53	5,48	5,79	6,86	5,35
Set.	6,08	5,90	6,17	5,24	5,60	6,75	5,17
Out.	6,11	6,05	6,30	5,34	5,79	6,70	5,21
Nov.	6,06	6,13	6,32	5,51	6,06	6,52	5,07
Dez.	5,41	5,81	5,92	5,32	6,21	6,28	4,83

2.7. Operações de crédito do sistema financeiro privado (Risco normal) - Ao setor público estadual e municipal – OCSFPrASPEM (2009).

Tabela A11: Série do OCSFPrASPEM

	2004	2005	2006	2007	2008	2009	2010
Jan.	757	1.663	3.300	3.134	3.488	3.014	2.636
Fev.	862	1.682	3.608	3.134	3.516	2.989	2.592
Mar.	865	1.916	3.632	3.032	3.507	2.963	2.545
Abr.	1.164	2.268	3.609	3.223	3.534	2.929	2.732
Mai.	1.191	2.197	3.649	3.312	3.581	2.902	2.713
Jun.	1.628	2.553	3.542	3.295	3.310	2.827	2.567
Jul.	1.669	2.704	3.678	3.172	3.320	2.786	2.574
Ago.	1.711	2.923	3.227	3.241	3.346	2.766	2.516
Set.	1.712	3.046	3.125	3.298	3.356	2.753	2.481
Out.	1.889	2.924	3.072	3.185	3.106	2.569	2.408
Nov.	1.805	3.090	3.077	3.386	3.116	2.579	2.334
Dez.	1.653	3.235	3.170	3.402	3.135	2.627	2.248

3. Série Exogena SBCRJ (13385)

A série foi obtido do MDIC.
Fonte: Ministério do Desenvolvimento. Indústria e Comércio Exterior (MDIC)
http://www.mdic.gov.br/sitio/interna/interna.php?area=5&menu=1078&refr=1076
Selecione o ano e mês e clique na Unidade da Federação a qual deseja obter as séries correspondentes.
Será gerado um arquivo zipado. Há um arquivo dentro da pasta zipada. RJ-BC-data que contêm as séries de exportação. Importação e saldo da balança comercial.

3.1. Saldo da Balança Comercial - Rio de Janeiro - US$ (mil) – SBCRJ (13385)

Tabela A12: Série SBCRJ

	2004	2005	2006	2007	2008	2009	2010
Jan.	5.853	-65.586	559.222	326.781	74.400	-418.929	167.862
Fev.	32.560	-85.751	43.188	151.901	152.942	85.252	634.894
Mar.	420	67.684	43.776	502.025	44.358	-176.627	270.711
Abr.	-38.887	169.734	404.074	401.679	-380.206	-92.953	291.925
Mai.	-36.864	20.357	118.230	203.357	890.748	95.476	914.792
Jun.	-65.562	-99.061	237.696	169.970	614.584	27.932	146.497
Jul.	-48.054	354.945	506.226	235.299	202.508	725.814	-255.590
Ago.	71.905	75.548	470.810	598.855	832.192	815.960	-344.260
Set.	-136.250	402.395	703.831	773.510	97.393	99.242	-199.719
Out.	421.650	-2.946	258.303	389.160	544.387	492.061	-376.152
Nov.	11.427	362.106	297.188	132.756	965.877	359.480	109.682
Dez.	454.091	301.978	550.602	867.785	252.704	-137.886	1.996.214

4. Séries Exógenas SELIC. IPCAE e INPC

4.1. Sistema Especial de Liquidação e de Custodia – SELIC.
Fonte: Receita Federal.
http://www.receita.fazenda.gov.br/pagamentos/jrselic.htm

Tabela A13: Série SELIC

	2004	2005	2006	2007	2008	2009	2010
Jan.	1,27	1,38	1,43	1,08	0,93	1,05	0,66
Fev.	1,08	1,22	1,15	0,87	0,80	0,86	0,59
Mar.	1,38	1,53	1,42	1,05	0,84	0,97	0,76
Abr.	1,18	1,41	1,08	0,94	0,90	0,84	0,67
Mai.	1,23	1,50	1,28	1,03	0,88	0,77	0,75
Jun.	1,23	1,59	1,18	0,91	0,96	0,76	0,79
Jul.	1,29	1,51	1,17	0,97	1,07	0,79	0,86
Ago.	1,29	1,66	1,26	0,99	1,02	0,69	0,89
Set.	1,25	1,50	1,06	0,80	1,10	0,69	0,85
Out.	1,21	1,41	1,09	0,93	1,18	0,69	0,81
Nov.	1,25	1,38	1,02	0,84	1,02	0,66	0,81
Dez.	1,48	1,47	0,99	0,84	1,12	0,73	0,93

4.2. Índice Nacional de Preços ao Consumidor Amplo Especial – IPCA-E.
Fonte: (Instituto Brasileiro de Geografia e Estatística) – IBGE
http://www.portalbrasil.net/ipca_e.htm

Tabela A14: Série IPCA-E

	2004	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Jan.	0,68	0,68	0,51	0,52	0,70	0,40	0,52
Fev.	0,90	0,74	0,52	0,46	0,64	0,63	0,94
Mar.	0,40	0,35	0,37	0,41	0,23	0,11	0,55
Abr.	0,21	0,74	0,17	0,22	0,59	0,36	0,48
Mai.	0,54	0,83	0,27	0,26	0,56	0,59	0,63
Jun.	0,56	0,12	-0,15	0,29	0,90	0,38	0,19
Jul.	0,93	0,11	-0,02	0,24	0,63	0,22	-0,09
Ago.	0,79	0,28	0,19	0,42	0,35	0,23	-0,05
Set.	0,49	0,16	0,05	0,29	0,26	0,19	0,31
Out.	0,32	0,56	0,29	0,24	0,30	0,18	0,62
Nov.	0,63	0,78	0,37	0,23	0,49	0,44	0,86
Dez.	0,84	0,38	0,35	0,70	0,29	0,38	0,69

4.3. Índice Nacional de Preços ao Consumidor – INPC
Fonte: (Instituto Brasileiro de Geografia e Estatística) – IBGE
http://www.portalbrasil.net/inpc.htm

Tabela A15: Série INPC

	2004	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Jan.	0,83	0,57	0,38	0,49	0,69	0,64	0,88
Fev.	0,39	0,44	0,23	0,42	0,48	0,31	0,70
Mar.	0,57	0,73	0,27	0,44	0,51	0,20	0,71
Abr.	0,41	0,91	0,12	0,26	0,64	0,55	0,73
Mai.	0,40	0,70	0,13	0,26	0,96	0,60	0,43
Jun.	0,50	-0,11	-0,07	0,31	0,91	0,42	-0,11
Jul.	0,73	0,03	0,11	0,32	0,58	0,23	-0,07
Ago.	0,50	0,00	-0,02	0,59	0,21	0,08	-0,07
Set.	0,17	0,15	0,16	0,25	0,15	0,16	0,54
Out.	0,17	0,58	0,43	0,30	0,50	0,24	0,92
Nov.	0,44	0,54	0,42	0,43	0,38	0,37	1,03
Dez.	0,86	0,40	0,62	0,97	0,29	0,24	0,60

APÊNDICE B: Analise da correlação cruzada

1. Correlação Cruzada da Série ISS com a série exógena SOCSFNRJPF.
   Saldo das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Pessoas Físicas - R$ (milhões) – SOCSFNRJPF (14020).

Cross Correlogram of ISS and SOCSFNRJPF

Date: 02/14/12 Time: 16:05
Sample: 2004M01 2010M12
Included observations: 84
Correlations are asymptotically consistent approximations

ISS,SOCSFNRJPF(-i)	ISS,SOCSFNRJPF(+i)	i	lag	lead
		0	0.7983	0.7983
		1	0.7558	0.7483
		2	0.7221	0.7070
		3	0.6928	0.6656
		4	0.6558	0.6204
		5	0.6208	0.5788
		6	0.5846	0.5402
		7	0.5528	0.5299
		8	0.5163	0.5006
		9	0.4790	0.4726
		10	0.4578	0.4471
		11	0.4402	0.4279
		12	0.4159	0.4101

2. Correlação Cruzada da Série ISS com a série exógena SOCSFNRJPJ.
   Saldo das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Pessoas Jurídicas - R$ (milhões) – SOCSFNRJPJ (14047).

Cross Correlogram of ISS and SOCSFNRJPJ

Date: 02/14/12 Time: 16:09
Sample: 2004M01 2010M12
Included observations: 84
Correlations are asymptotically consistent approximations

ISS,SOCSFNRJPJ(-i)	ISS,SOCSFNRJPJ(+i)	i	lag	lead
		0	0.7851	0.7851
		1	0.7534	0.7396
		2	0.7194	0.7043
		3	0.6970	0.6600
		4	0.6563	0.6169
		5	0.6260	0.5711
		6	0.5903	0.5421
		7	0.5620	0.5440
		8	0.5319	0.5202
		9	0.4976	0.4890
		10	0.4747	0.4616
		11	0.4504	0.4485
		12	0.4305	0.4265

3. Correlação Cruzada da Série ISS com a série exógena SOCSFNRJTot
   Saldo das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Total - R$ (milhões) – SOCSFNRJTot (14074).

Cross Correlogram of ISS and SOCSFNRJTOT

Date: 02/14/12 Time: 16:13
Sample: 2004M01 2010M12
Included observations: 84
Correlations are asymptotically consistent approximations

ISS,SOCSFNRJTOT(-i)	ISS,SOCSFNRJTOT(+i)	i	lag	lead
		0	0.7911	0.7911
		1	0.7563	0.7442
		2	0.7223	0.7072
		3	0.6979	0.6635
		4	0.6581	0.6197
		5	0.6264	0.5749
		6	0.5905	0.5432
		7	0.5612	0.5418
		8	0.5292	0.5164
		9	0.4940	0.4860
		10	0.4715	0.4590
		11	0.4489	0.4442
		12	0.4278	0.4233

4. Correlação Cruzada da Série ISS com a série exógena TIOCSFNRJPJ
   Taxa de inadimplência das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Pessoas jurídicas % – TIOCSFNRJPJ (15911).

Cross Correlogram of ISS and TIOCSFNRJPJ

Date: 02/14/12 Time: 16:20
Sample: 2004M01 2010M12
Included observations: 84
Correlations are asymptotically consistent approximations

ISS,TIOCSFNRJPJ(-i)	ISS,TIOCSFNRJPJ(+i)	i	lag	lead
		0	-0.5489	-0.5489
		1	-0.5276	-0.5046
		2	-0.5103	-0.4739
		3	-0.5034	-0.4325
		4	-0.4792	-0.3875
		5	-0.4686	-0.3644
		6	-0.4430	-0.3464
		7	-0.4236	-0.3236
		8	-0.3994	-0.2907
		9	-0.3666	-0.2751
		10	-0.3597	-0.2500
		11	-0.3621	-0.2297
		12	-0.3591	-0.2125

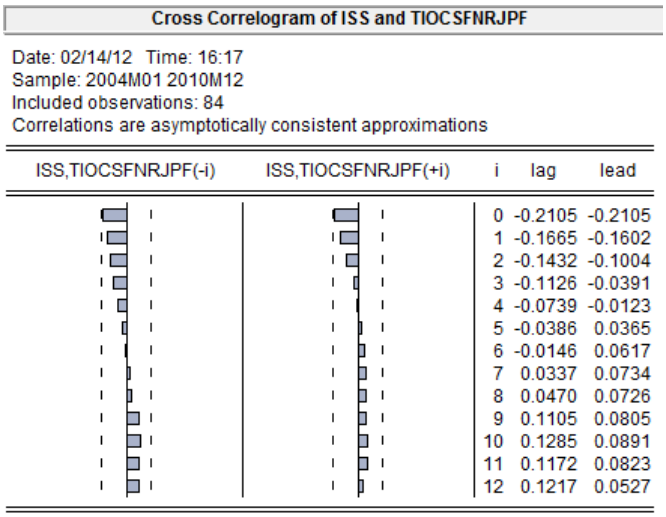
5. Correlação Cruzada da Série ISS com a série exógena TIOCSFNRJPJT
   Taxa de inadimplência das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Total (%) – TIOCSFNRJPJT (15943).

Cross Correlogram of ISS and TIOCSFNRJT

Date: 02/14/12 Time: 16:24
Sample: 2004M01 2010M12
Included observations: 84
Correlations are asymptotically consistent approximations

ISS,TIOCSFNRJT(-i)	ISS,TIOCSFNRJT(+i)	i	lag	lead
		0	-0.5748	-0.5748
		1	-0.5497	-0.5259
		2	-0.5288	-0.4897
		3	-0.5189	-0.4407
		4	-0.4893	-0.3921
		5	-0.4758	-0.3621
		6	-0.4474	-0.3428
		7	-0.4238	-0.3219
		8	-0.3996	-0.2894
		9	-0.3610	-0.2716
		10	-0.3512	-0.2447
		11	-0.3532	-0.2267
		12	-0.3512	-0.2125

6. Correlação Cruzada da Série ISS com a série exógena TIOCSFNRJPF
   Taxa de inadimplência das operações de crédito do Sistema Financeiro Nacional - Rio de Janeiro - Pessoas físicas % – TIOCSFNRJPF (15879).

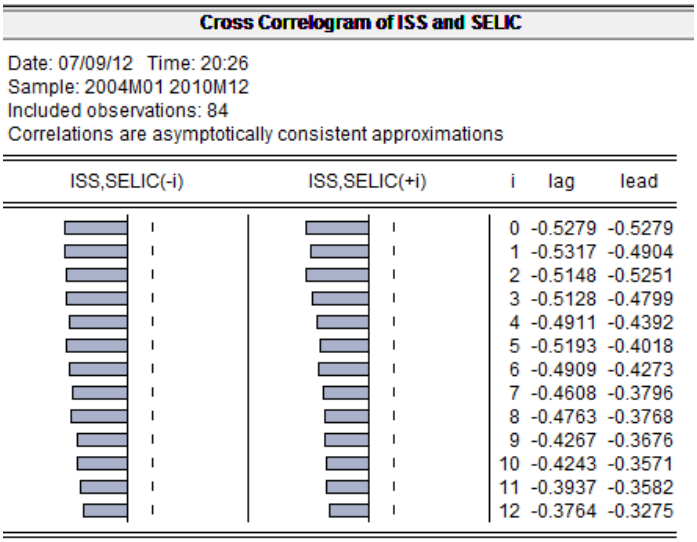


Cross Correlogram of ISS and TIOCSFNRJPF

Date: 02/14/12 Time: 16:17
Sample: 2004M01 2010M12
Included observations: 84
Correlations are asymptotically consistent approximations

ISS,TIOCSFNRJPF(-i)	ISS,TIOCSFNRJPF(+i)	i	lag	lead
		0	-0.2105	-0.2105
		1	-0.1665	-0.1602
		2	-0.1432	-0.1004
		3	-0.1126	-0.0391
		4	-0.0739	-0.0123
		5	-0.0386	0.0365
		6	-0.0146	0.0617
		7	0.0337	0.0734
		8	0.0470	0.0726
		9	0.1105	0.0805
		10	0.1285	0.0891
		11	0.1172	0.0823
		12	0.1217	0.0527

7. Correlação Cruzada da Série ISS com a série exógena SELIC
   Sistema Especial de Custódia – SELIC.



Cross Correlogram of ISS and SELIC

Date: 07/09/12 Time: 20:26
Sample: 2004M01 2010M12
Included observations: 84
Correlations are asymptotically consistent approximations

ISS,SELIC(-i)	ISS,SELIC(+i)	i	lag	lead
		0	-0.5279	-0.5279
		1	-0.5317	-0.4904
		2	-0.5148	-0.5251
		3	-0.5128	-0.4799
		4	-0.4911	-0.4392
		5	-0.5193	-0.4018
		6	-0.4909	-0.4273
		7	-0.4608	-0.3796
		8	-0.4763	-0.3768
		9	-0.4267	-0.3676
		10	-0.4243	-0.3571
		11	-0.3937	-0.3582
		12	-0.3764	-0.3275

8. Correlação Cruzada da Série ISS com a série exógena SBCRJ
   Saldo da Balança Comercial - Rio de Janeiro - US$ (mil) – SBCRJ (13385).

Cross Correlogram of ISS and SBCRJ

Date: 02/14/12 Time: 16:00
Sample: 2004M01 2010M12
Included observations: 84
Correlations are asymptotically consistent approximations

ISS,SBCRJ(-i)	ISS,SBCRJ(+i)	i	lag	lead
		0	0.1720	0.1720
		1	0.0047	0.0813
		2	-0.0249	0.1326
		3	0.0158	0.1000
		4	0.0861	0.1091
		5	0.0752	0.0795
		6	0.1309	0.0124
		7	0.1812	0.1366
		8	0.0368	0.0570
		9	0.0299	-0.0019
		10	0.0044	0.0091
		11	-0.0153	0.0486
		12	0.1857	0.0404

9. Correlação Cruzada da Série ISS com a série exógena OCSFPrASPEM
Operações de crédito do sistema financeiro privado (Risco normal) - Ao setor público estadual e municipal – OCSFPrASPEM (2009).

Cross Correlogram of ISS and OCSFPRASPEM

Date: 02/14/12 Time: 15:39
Sample: 2004M01 2010M12
Included observations: 84
Correlations are asymptotically consistent approximations

ISS,OCSFPRASPEM(-i)	ISS,OCSFPRASPEM(+i)	i	lag	lead
		0	0.2396	0.2396
		1	0.2340	0.2049
		2	0.2229	0.1683
		3	0.2648	0.1411
		4	0.2479	0.1299
		5	0.2247	0.0973
		6	0.2059	0.0825
		7	0.2028	0.0513
		8	0.1884	0.0211
		9	0.1506	-0.0017
		10	0.1684	-0.0379
		11	0.1951	-0.0668
		12	0.1868	-0.1080

10. Correlação Cruzada da Série ISS com a série exógena IPCA-E
Índice Nacional de Preços ao Consumidor Amplo Especial – IPCA-E.

Cross Correlogram of ISS and IPCAE

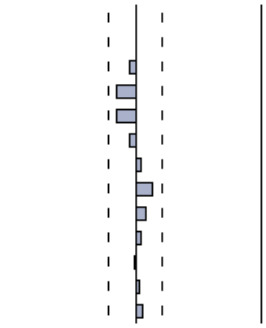
Date: 07/10/12 Time: 15:51
Sample: 2004M01 2010M12
Included observations: 84
Correlations are asymptotically consistent approximations

ISS,IPCAE(-i)	ISS,IPCAE(+i)	i	lag	lead
		0	-0.1071	-0.1071
		1	-0.0971	-0.0400
		2	-0.2141	0.0807
		3	-0.2041	-0.0721
		4	-0.2013	-0.1066
		5	-0.1008	-0.0779
		6	0.0179	-0.0846
		7	-0.0371	-0.1347
		8	0.0287	-0.1461
		9	-0.0683	-0.0844
		10	0.0106	-0.0994
		11	-0.0026	-0.0108
		12	-0.0232	0.0318

11. Correlação Cruzada da Série ISS com a série exógena INPC
    Índice Nacional de Preços ao Consumidor – INPC

Cross Correlogram of ISS and INPC

Date: 07/10/12 Time: 15:54
Sample: 2004M01 2010M12
Included observations: 84
Correlations are asymptotically consistent approximations



ISS,INPC(-i)	ISS,INPC(+i)	i	lag	lead
		0	0.0131	0.0131
		1	0.0156	0.1547
		2	-0.0521	0.0909
		3	-0.1611	0.0285
		4	-0.1562	0.0139
		5	-0.0508	0.0649
		6	0.0452	-0.0107
		7	0.1345	-0.0656
		8	0.0882	-0.0887
		9	0.0547	-0.0196
		10	-0.0043	-0.0202
		11	0.0252	0.0446
		12	0.0647	0.0400