UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

TESE DE DOUTORADO

Classificação de Contribuintes: um modelo em duas fases

Eder Daniel Corvalão

Florianópolis, Setembro de 2009.

Eder Daniel Corvalão

Classificação de Contribuintes: um modelo em duas fases

Tese de Doutorado Apresentada ao
Programa de Pós-Graduação em Engenharia de Produção
da Universidade Federal de Santa Catarina
como requisito parcial para obtenção
do Grau de Doutor em Engenharia de Produção

Orientador: Robert Wayne Samohyl, Ph.D.

Florianópolis, Setembro de 2009.

EDER DANIEL CORVALÃO

Classificação de Contribuintes: um modelo em duas fases

Esta tese foi julgada e aprovada para a obtenção do grau de **Doutor em Engenharia de Produção** no **Programa de Pós-Graduação em Engenharia de Produção** da Universidade Federal de Santa Catarina

Florianópolis, 16 de outubro de 2009.

_____

Prof. Antonio Cezar Bornia
Coordenador do Programa

BANCA EXAMINADORA

Prof. Robert Wayne Samohyl, Ph.D UFSC (Orientador)	Prof. José Leomar Todesco, Dr. UFSC
Prof. Paulo José Ogliari, Dr. UFSC	Prof. Pedro Alberto Barbetta, Dr. UFSC
Prof. Gerson Lachtemacher, PhD. (Examinador externo)	Prof. Gutemberg Hespanha Brasil, Dr. (Examinador externo)

DEDICATÓRIA

Ao meu filho Daniel Henrique e à minha companheira constante Berenice, pelo carinho e compreensão em todos os momentos. A minha mãe, Eudócia pelo apoio incondicional e todo esforço dispensado até este momento. Em especial a lembrança de meu pai, Pedro Antonio seu exemplo de vida e dignidade é a maior herança que eu poderia ter recebido.

AGRADECIMENTOS

A todos os meus colegas, amigos e familiares que, direta ou indiretamente, colaboraram para o desenvolvimento deste trabalho, em especial:

Ao professor Robert Wayne Samohyl, pela orientação constante e confiança depositada nesta pesquisa.

Ao professores membros da banca examinadora desta tese pela contribuição ao aperfeiçoamento deste trabalho e também, ao apoio dos professores do departamento de Engenharia de Produção e do INE da Universidade Federal de Santa Catarina. Ao professor Adriano Mendonça Souza pelas sugestões oportunas.

Aos colegas do Núcleo de Normalização e Qualimetria (NNQ) pelo companheirismos: Andréa Cristina Konrath, Custódia da Cunha Alves, Edson Marcos Leal Soares Ramos, Elisa Henning, Gueibi Peres Souza, Manoel Domingos Filho, Rodrigo Gabriel de Miranda, Rubson Rocha, Silvia dos Santos de Almeida, e Wesley Vieira da Silva

Agradecimento especial ao ex-Gerente de Fiscalização da Secretaria de Estado da Fazenda de Santa Catarina, Sr. Renato Prux e aos auditores fiscais: Valêncio Ferreira da Silva Neto, Olândio Hornburg, Renato Dias Marques de Lacerda e demais integrantes da GEFIS, pelo apoio proporcionado a este trabalho.

Ao Centro de Informática e Automação de Santa Catarina – CIASC, pela oportunidade de desenvolver meus conhecimentos, e aos colegas que colaboraram diretamente para o sucesso da pesquisa: Ademir João da Rosa, André Arantes Capuano Morais, Fabio José do Amaral, Hugo César Hoeschl, José Ernesto Pereira, Luciano Campos da Cunha, Paulo Ricardo Foschiera,

"If we could first know where we are, and whither we are tending, we could then better judge what to do, and how to do it."

Abraham Lincoln (1858)

RESUMO

O termo contribuinte se aplica à pessoa física ou jurídica que a lei obriga ao cumprimento de obrigação tributária. É função da administração tributária acompanhar e fiscalizar a correta execução das obrigações fiscais das empresas contribuintes. Na impossibilidade do acompanhamento de todas as empresas, o processo de seleção de contribuintes a serem auditados torna-se de vital importância.

Com o crescimento do volume de informações apresentadas pelos contribuintes, sistematicamente armazenados em sistemas operacionais; e, com o aparecimento de novas ferramentas de análise de dados aliados à evolução dos recursos computacionais surgem novas alternativas para abordar o problema da seleção de contribuintes. Neste cenário a área de mineração de dados (*data mining*) aparece com diversas aplicações nas mais variadas áreas, entre elas a de detecção de fraude.

Esta tese desenvolve um modelo formal para classificação dos contribuintes a partir dos dados de movimentação mensal que são apresentados ao setor de fiscalização. A proposta busca preservar as características econômicas e regionais de cada empresa, valendo-se da análise de agrupamentos. Na seqüência são construídos modelos probabilísticos que serão usados para relacionar os contribuintes com maiores indícios de irregularidades. Esta relação poderá ser utilizada para direcionar a seleção das empresas a serem auditadas.

Para sua validação, este modelo foi aplicado num estudo de caso junto à Secretaria da Fazenda do Estado de Santa Catarina. A seleção de contribuintes do ICMS (Imposto sobre Circulação de Mercadorias e Serviços) foi o tema analisado utilizando-se dados mensais entre os anos 2005 e 2007.

Palavras-chave: Contribuinte, sonegação, detecção de fraude, *data mining*.

ABSTRACT

Term taxpayer applies to the person or entity that the law requires the fulfillment of tax obligations. It is a task of the state's tax administration to monitor and supervise the proper implementation of the tax obligations of business taxpayers. On the impossibility of monitoring tens of thousands of business enterprises the procedures for selecting taxpayers to be audited is of vital importance.

With the growing volume of information submitted by taxpayers, always stored in operating systems, and with the emergence of new tools of data analysis combined with the development of computational resources arises new alternatives to address the selection of contributors. In this scenario the area of data mining appears as a relevant tool for fraud detection.

This thesis develop a formal model for classification of taxpayers from the monthly data that are presented by the tax payers to the state tax surveillance agency. The first step in model building uses the economic and regional characteristics of each company to define groups through cluster analysis. The next step is to construct probabilistic models that will be used to indicate taxpayers with more likelihood of irregularities. This relationship can be used to guide the selection of firms to be audited.

This model was applied in a case study to the State of Santa Catarina Department of Finance, using monthly data for the period 2005 to 2007.

Keywords: Taxpayer, Tax evasion, Fraud detection, Data Mining.

Lista de Figuras

Lista de Tabelas

Lista de Siglas

AG – Algoritmos Genéticos

CRISP-DM - *CRoss-Industry Standard Process for Data Mining*

CTN – Código Tributário Nacional

ESCELSA - Espírito Santo Centrais Elétricas S. A.

EUA – Estados Unidos da América

GECIN - Gerenciamento de Carteiras de Índices

GEFIS – Gerência de Fiscalização

GES - Grupos Especialistas Setoriais

IA - Inteligência Artificial

IBPT - Instituto Brasileiro de Planejamento Tributário

ICMS - Imposto sobre Circulação de Mercadorias e Serviços

IVA - Imposto sobre valor agregado

KDD – *Knowledge Discovery in Database* (Descoberta de Conhecimento em Base de Dados)

PIS - Programa de Integração Social

RS – Rio Grande do Sul

SANASA - Sociedade de Abastecimento de Água e Saneamento S.A

SC – Santa Catarina

SP – São Paulo

USEFI – Unidade Setorial de Fiscalização

SUMÁRIO

1 - INTRODUÇÃO

Disseminar a educação, garantir saúde, manutenção da ordem e da segurança, e promoção da justiça social são as grandes funções do Estado, considerado como ente político criado pelos indivíduos para atender ao bem comum. Para desempenhar seu papel e atender a estas funções o Estado conta com recursos oriundos da tributação que se apresentam sob a forma de impostos, taxas e contribuições.

Cabe à administração pública a verificação do cumprimento dos dispositivos legais relativos aos tributos e para tal conta com um corpo especializado de funcionários que atuam nas atividades de fiscalização e auditoria. A partir da constatação da desproporção entre a quantidade de auditores e o número de contribuintes[1], todos os esforços do setor de fiscalização devem ser no sentido de procurar potencializar esta atividade. Desta forma, a análise criteriosa das informações apresentadas mensalmente pelos contribuintes é instrumento primordial para a definição das políticas de gestão tributária e a alocação mais efetiva dos recursos disponíveis.

Para o processo de seleção dos contribuintes a serem auditados, diversas experiências foram testadas ao longo dos anos, desde a seleção aleatória das empresas, passando pela divisão em grupos de acordo com o montante arrecadado por cada contribuinte – que numa análise mais simplista, geralmente aponta para os maiores arrecadadores. Chegando, nos últimos anos, ao advento das divisões em grupos setoriais, onde cada grupo de fiscalização fica responsável pelo conhecimento e acompanhamento das empresas de um setor econômico específico.

No processo de acompanhamento das informações financeiras dos contribuintes os órgãos fiscalizadores foram acumulando uma grande quantidade de dados, oriundos dos sistemas operacionais. Em outras épocas estas informações serviram apenas para controle do cumprimento ou não das obrigações fiscais por parte das empresas. Com o avanço e a popularização dos recursos computacionais e a evolução das técnicas de análise de dados novas possibilidades surgem para auxiliar a administração tributária.

[1] No caso do Estado de Santa Catarina, são 443 fiscais para um total de 138.136 empresas (dados da SEF-SC para o ano de 2006).

Neste contexto, vem atraindo a atenção dos envolvidos com a tecnologia da informação e a comunidade acadêmica em geral, as experiências com mineração de dados (*data mining*). Consiste na análise de dados e a descoberta de padrões ou modelos. "*Data mining* é a análise de grandes conjuntos de dados para encontrar relacionamentos não aparentes, para que se tornem úteis e valiosos para os donos da informação" (HAND *et al.*, 1998. Tradução do autor[2]).

As técnicas de *data mining* surgiram da evolução dos sistemas de banco de dados e sua integração com a inteligência artificial (IA) e a estatística tradicional, e são especialmente indicadas para a descoberta de regras ou padrões, prever futuras tendências e comportamentos ou conhecer grupos similares. Atualmente existem diversas referências de utilização de *Data Mining* na detecção de fraudes nos mais variados segmentos econômicos: cartões de crédito, água, telefonia, evasão fiscal, distribuição de energia, entre outros.

1.1 – OBJETIVOS

1.1.1 - Objetivo Geral

Elaborar uma modelo para classificação de contribuintes que auxilie os órgãos de fiscalização no processo de seleção e direcionamento de auditoria tendo como base as probabilidades de indícios de irregularidades contidas nas declarações apresentadas regularmente.

1.1.2 - Objetivos específicos

Para o desenvolvimento do presente trabalho é necessário o estudo dos objetivos específicos citados a seguir:

- Escolher métodos adequados para classificar e agrupar os contribuintes nos diversos setores econômicos e dentro de suas respectivas regiões;

[2] Todas as citações e referências em língua inglesa neste trabalho são traduções do autor.

- Definir um conjunto de equações para calcular indícios de irregularidades nas declarações dos contribuintes; buscando atender o maior número de empresas;
- Disponibilizar algoritmos específicos para auxiliar a aplicação do modelo, especialmente nos casos de dados faltantes, geração de quadros com combinações de variáveis e seleção de equações;
- Aplicar o modelo proposto num contexto real junto ao Estado de Santa Catarina. Neste caso em particular, serão geradas probabilidades e classificadas as empresas contribuintes do ICMS (Imposto sobre Circulação de Mercadorias e Serviços) estadual.

1.2 – JUSTIFICATIVA E IMPORTÂNCIA

Neste trabalho será analisado o processo de extração de conhecimento a partir de grandes bases de dados e o emprego das técnicas de *data mining* para facilitar a atividade de seleção de contribuintes, no sentido de melhor alocar os recursos de auditoria. O foco é a detecção de indícios de irregularidades nas declarações apresentadas.

Sobre o volume do problema causado pela evasão fiscal, Futema (2005) apresenta dados da pesquisa realizada pelo IBPT[3] no ano de 2004 envolvendo 7.437 empresas de todos os setores brasileiros e de portes diferenciados, indicando que 29,45% das empresas pesquisadas apresentam "fortes indícios de sonegação fiscal".

Ao avaliarem a extensão do problema causado pela sonegação, Siqueira e Ramos (2005) apresentam diversos estudos que tentam mensurar os valores sonegados em vários países empregando diferentes abordagens neste processo. Vale destacar as conclusões a seguir:

> A aplicação de diversos métodos de mensuração sugere que nos países industrializados ocidentais a sonegação de impostos atinge entre 5% a 25% da arrecadação tributária potencial, dependendo da técnica adotada e do país, com percentuais mais elevados (até 30% a 40%) para países menos desenvolvidos. (TAMZI e SHOME *apud* SIQUEIRA e RAMOS, *op. cit.*).

[3] IBPT, Instituto Brasileiro de Planejamento Tributário - Organização Não Governamental que analisa e acompanha a arrecadação tributária no País e também divulga periodicamente cálculos da carga global.

Acerca do emprego de técnicas de *data mining* na área fiscal, deve-se destacar o protótipo apresentado pelo Estado de São Paulo analisando informações econômico-fiscais com 70 empresas, onde foram encontradas mais de 3.700 inconsistências, o que representa um valor de R$300.000.000, com ICMS a recuperar calculado em R$15.000.000. Rubin (2006) ainda acrescenta o comentário do gerente do projeto, o fiscal Edson G. de Souza: "[...] Imagine então o que não se consegue detectar com 800 mil empresas, que é a quantidade cadastrada em nossos bancos de dados".

Outro importante trabalho é relatado por Micci-Barreca e Ramachandran (2006) que comparam as formas tradicionais de seleção de contribuintes para auditoria no Estado do Texas (EUA) e o modelo sugerido pelos autores, baseado em técnicas de *data mining*. Neste trabalho é apontada uma melhoria média de aproximadamente 16% com a adoção do novo sistema. No âmbito da arrecadação federal merece destaque o trabalho apresentado por Barreto (2005) que utilizou modelos baseados em *data mining* para previsão do comportamento e classificação dos contribuintes tributários.

Este trabalho, apesar de utilizar técnicas consagradas, visa à construção de um modelo que possibilite o acompanhamento de todas as etapas do processo de análise das informações econômicas e a classificação dos contribuintes a partir dos indícios de irregularidades. Esta estruturação do processo de análise utilizando *data mining* para o problema em questão constitui uma das inovações.

A qualidade dos resultados obtidos está diretamente relacionada com a quantidade de dados disponíveis e as técnicas empregadas na construção dos modelos. Para este trabalho foram disponibilizados dados operacionais de todos os sistemas informatizados de fiscalização e auditoria onde o modelo sugerido foi aplicada.

As questões pertinentes ao volume de dados existente e sua implicação no emprego de técnicas quantitativas por si só justificam a não trivialidade deste trabalho; entretanto, os processos de seleção de equações e classificação das empresas apresentam uma série de desafios nesta pesquisa. As características dos métodos de análise escolhidos apontaram para o desenvolvimento de algoritmos específicos envolvendo grande quantidade de procedimentos estatísticos.

1.3 – ESTRUTURA DO TRABALHO

O trabalho está dividido em cinco capítulos. Os demais capítulos estão assim estruturados:

Capítulo 2 abrange o processo de descoberta do conhecimento, conceitos de *data mining*, ferramentas e as técnicas envolvidas no processo. Finaliza com experiências de *data mining* na detecção de fraude.

Capítulo 3 apresenta o modelo proposto para classificação dos contribuintes a serem auditados, as razões desta proposição e os métodos e técnicas de análise que devem ser utilizados. Também são apresentados os documentos gerados em cada etapa do modelo.

Capítulo 4 ilustra o modelo sugerido com o estudo de caso para classificação de contribuintes do ICMS no Estado de Santa Catarina. Utilizando-se dados dos anos de 2005 a 2007 foi criado um novo agrupamento para as empresas contribuintes e gerado um conjunto de equações para cálculo das probabilidades de ocorrência de irregularidades. O processo é apresentado como um todo e relatam-se os resultados obtidos.

Por fim, o capítulo 5 traz uma síntese dos principais tópicos do trabalho, pressupostos e restrições, bem como sugestões para pesquisas futuras.

2 - EXTRAÇÃO DE CONHECIMENTO (KDD) E *DATA MINING*

2.1 – INTRODUÇÃO

Neste capítulo são apresentados aspectos relacionados com o volume crescente de dados nas organizações e a dificuldade em se obter informações relevantes. Os conceitos de *data mining* são abordados na seção 2.3. Estes conceitos serão fundamentais para a proposta metodológica apresentada. Serão abordadas aplicações práticas de *data mining*, com destaque para a detecção de fraude.

2.2 - PROCESSO DE EXTRAÇÃO DO CONHECIMENTO

A rápida evolução dos recursos computacionais ocorrida nos últimos anos permitiu que simultaneamente fossem gerados grandes volumes de dados. Estima-se que a quantidade de informação no mundo dobra a cada 20 meses e que o tamanho e a quantidade dos bancos de dados crescem com velocidade ainda maior (Dilly, 1999). Segundo Fayyad *et al.* (1996), o processamento de dados é uma ciência que vem progredindo de duas maneiras: em número de objetos de pesquisa ou conteúdos processados e em número de campos ou áreas relacionadas a esses objetos.

Existe uma necessidade econômica e científica de utilizar a tecnologia para facilitar o trabalho humano de coleta e seleção de dados, e mesmo de verificação destes dados, para estabelecer um parâmetro de investigação quando a quantidade de informação a ser digerida é muito grande. Como os computadores possibilitaram um armazenamento de dados superior ao qual o raciocínio humano é capaz de discernir, é natural o uso de técnicas computacionais específicas para ajudar na criação de modelos que estruturem o enorme volume de informação existente.

Surge no início da década de 1990 uma área de estudos específica para Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* – KDD), com metodologia própria para preparação, exploração dos dados e interpretação dos resultados. De acordo com Fayyad *et al.* (*op. cit.*): “KDD é um processo de várias etapas, não trivial,

interativo e iterativo para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados". Outra definição é dada por Berry e Linoff (2004): "KDD é a exploração e análise de dados, por meios automáticos ou semi-automáticos, em grandes quantidades de dados, com o objetivo de descobrir regras ou padrões interessantes".

De acordo com Roger e Geatz, (2003), o processo de KDD pode ser visto como a aplicação do método científico para a descoberta de padrões nos dados. Portanto o KDD deve possuir características que possibilitem a verificação dos resultados e repetição dos experimentos. Nisbet *et al.* (2009) enumeram os passos do método científico e apresentam as semelhanças com o KDD:

1. Definir o problema;
2. reunir as informações existentes sobre o fenômeno;
3. formular uma ou mais hipóteses;
4. coletar novos dados experimentais;
5. analisar a informação no novo conjunto de dados;
6. interpretar os resultados;
7. sintetizar as conclusões, baseadas nos dados antigos, nos novos dados e na intuição;
8. formular novas hipóteses para mais testes; e
9. fazer novamente (iteração).

Os autores apontam que os passos 1 a 5 envolvem dedução, e os passos 6 a 9 envolvem indução. "Apesar do método científico se basear fortemente no raciocínio dedutivo, os produtos finais surgem através do raciocínio indutivo. O *data mining* é muito parecido com isto" (Nisbet *et al., op. cit.*). Como será descrito na seqüência deste capítulo o *data mining* é parte integrante do KDD e em muitas publicações se referem a ambos os termos como sinônimos.

O KDD está focado em todo o processo de descoberta de conhecimento a partir de um sistema de informação, incluindo como essa informação é armazenada e acessada, como os algoritmos podem estabelecer a seleção de dados dentro de um volume massivo de informação de maneira eficiente, como os resultados podem ser interpretados e visualizados e como toda esta interação pode ser modelada e monitorada. A Figura 1 ilustra o processo de extração do conhecimento a partir dos dados disponíveis na organização.

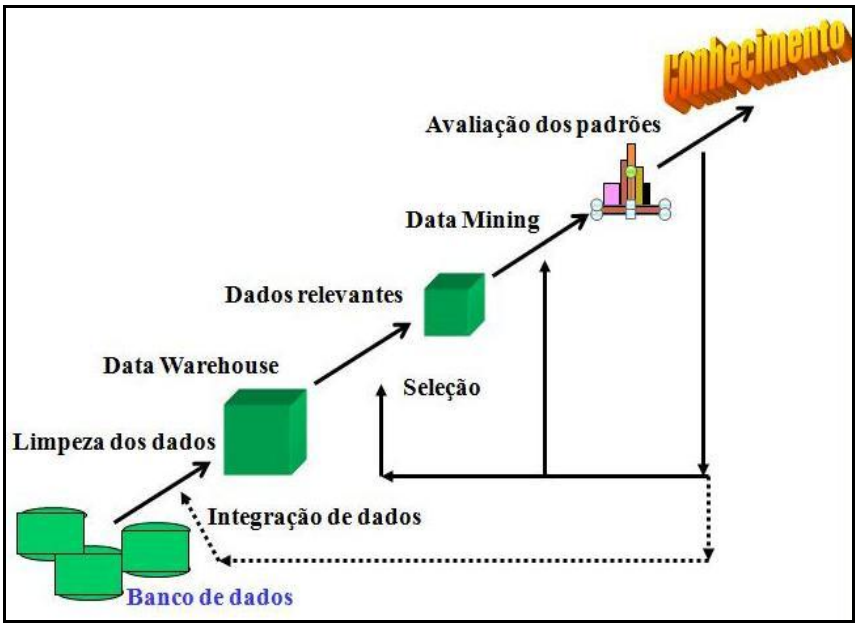

Figura 1 - Processo de extração do conhecimento (KDD). Fonte: Fayyad *et al.* (1996).

No entendimento de Fayyad et al. (*op. cit.*), o KDD se refere a todo o processo de descoberta de conhecimento aplicável via processamento de dados, e envolve também a preparação e seleção dos dados, incorporação de conhecimento prioritário e uma apresentação dos resultados selecionados, que são essenciais para o conhecimento derivado do processamento de dados.

Uma das áreas correlacionadas ao KDD que envolve o processamento de dados é a de *data warehousing*, que se refere à coleta e organização de dados para torná-los disponíveis para análise e decisões de apoio. Segundo Inmon e Hackathorn (1997), *data warehouse* é uma coleção de dados integrados, orientados por assunto, variáveis com o tempo e não voláteis, usados para suporte ao processo gerencial de tomada de decisões.

O processo de *data warehousing* ajuda o processo de KDD de duas maneiras importantes, denominadas limpeza dos dados (*data cleaning*) e integração dos dados (*data access*):

Limpeza dos dados: As organizações são pressionadas a pensar numa visão lógica unificada sobre a grande variedade de dados que elas processam. Desta forma, elas precisam direcionar o mapeamento de dados para uma convenção singular, que represente e organize as informações disponíveis, detectando falhas do sistema.

Integração dos dados: Métodos uniformes e bem definidos devem ser criados para que se consiga ter acesso a informações que, por alguma razão, não estão disponíveis.

Uma vez que as organizações e os indivíduos consigam resolver o problema de como armazenar e acessar os dados, a próxima etapa é definir o que fazer com toda a informação obtida. Ainda segundo Fayyad *et al.* (*op. cit.*), KDD é um processo complicado de identificação de modelos válidos, potencialmente utilitários e atuais em processamento de dados. Os autores salientam que o processo de KDD compreende muitas etapas, que envolvem a preparação dos dados, a procura por modelos, a avaliação e o refinamento do conhecimento, todas repetidas em múltiplas interações.

As principais áreas de aplicação do KDD incluem: marketing, finanças (principalmente investimento), detecção de fraudes, indústria, telecomunicações e sistemas de rede (Internet).

2.3 - *DATA MINING*

Data mining é uma das etapas do processo de KDD e consiste na análise de dados e descoberta de algoritmos que, diante das aceitáveis limitações da eficiência computacional, produz uma particular enumeração de padrões ou modelos sobre os dados. Conforme Giudici (2003): "*Data mining* é o processo de seleção, exploração, e modelagem de grandes quantidades de dados para descobrir padrões ou relações que são, em primeira análise desconhecidos, com o objetivo de obter resultados claros e úteis para o dono do banco de dados".

Também deve ser mencionado o caráter multidisciplinar que originou o *data mining*, tendo incorporado em sua essência técnicas de banco de dados e análise estatística. *"Data mining* é um campo interdisciplinar abrangendo técnicas de aprendizado automático, reconhecimentos de padrões, estatística e visualização para abordar a questão da extração de conhecimento de grandes bases de dados" (SIMOUDIS, E. *in* CABENA *et al.*, 1998). Para estes autores, o *data mining* constituiu uma interseção de diversos campos de pesquisa, a Figura 2 ilustra estas relações.

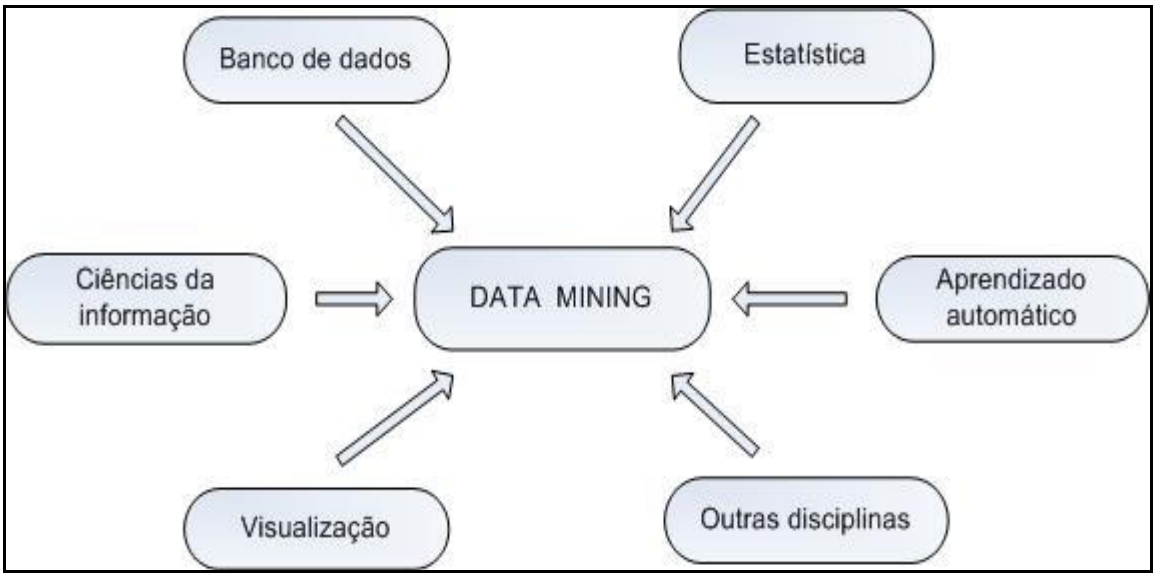

Figura 2 - Processo de extração do conhecimento (KDD). Adaptado de Classificação de Fayyad *et al.* (1996).

Estabelecida a seleção dos dados, o próximo passo é construir algoritmos específicos para a implementação de uma metodologia. Três componentes primários desse processo podem ser identificados: (1) modelo de representação; (2) modelo de avaliação e (3) pesquisa.

O modelo de representação é a linguagem utilizada para descrever os padrões descobertos. Se a representação é muito limitada, o montante de exemplos pode produzir um modelo acurado para os dados. O critério de avaliação consiste em verificar como um padrão em particular (um modelo e seus parâmetros) atinge as metas do processo de extração do conhecimento.

2.3.1 - Principais tarefas

a) Classificação: associa ou classifica um item a uma de várias classes discretas pré-definidas. A classificação busca estabelecer uma nova observação a uma classe já rotulada. De acordo com Sumathi e Sivanandam (2006), objetiva descobrir algum relacionamento entre os atributos de entrada e as classes de saída, de forma que este conhecimento possa ser usado para prever em qual classe um novo objeto não conhecido possa ser classificado.

b) Análise de Agrupamento (*Cluster*): associa um item a uma ou várias classes categóricas (ou *cluster*s), em que as classes são determinadas pelos dados, diversamente da classificação em que as classes são pré-definidas. Pode ser utilizado para avaliar similaridades entre os

dados e analisar correlações entre os atributos. Na Figura 3 pode-se observar um exemplo de agrupamento.

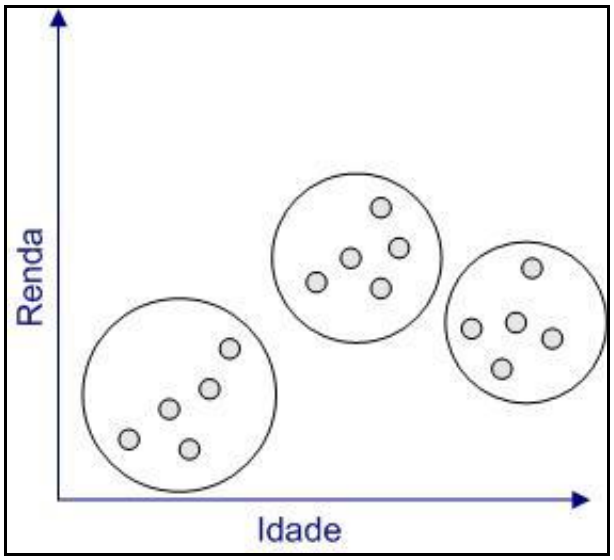

Figura 3 - Exemplo de agrupamento (Cluster)

c) Estimação / Predição: Esta tarefa procura definir os possíveis valores de alguns dados ou a distribuição de valores de certos atributos em um conjunto de objetos. Ela envolve a descoberta de um conjunto de atributos relevantes para o atributo de interesse. Usualmente, análise de regressão, modelo linear generalizado, análise de correlação e árvores de decisão são ferramentas úteis em predição de qualidade. Também são usados algoritmos genéticos e redes neurais com bastante sucesso. (IME, 2005).

d) Análise de ligações ou regras de associação: determinam relações entre campos de um banco de dados. A idéia é a derivação de correlações multivariadas que permitam subsidiar as tomadas de decisões, determinando fatos ou objetos que possam ocorrer juntos numa determinada operação. "Intuitivamente essa tarefa consiste em encontrar conjuntos de itens que ocorram simultaneamente e de modo freqüente em um banco de dados." (VIGLIONI, 2007). Na Figura 4 apresenta-se um exemplo desta tarefa, realizada a partir das observações de vendas num supermercado, o objetivo é investigar mais profundamente os produtos que apresentam similaridades sob o ponto de vista do consumidor. Nota-se na figura uma tendência a comprar vinho o consumidor que comprou queijo, o mesmo entre os clientes que compraram refrigerantes apresentam uma fraca relação com a aquisição de bebidas alcoólicas. Neste exemplo, a tarefa de análise de ligações possibilita ao supermercado dispor os produtos

de forma a facilitar o acesso dos clientes a produtos com este grau de similaridade, o que pode resultar num potencial aumento nas vendas.

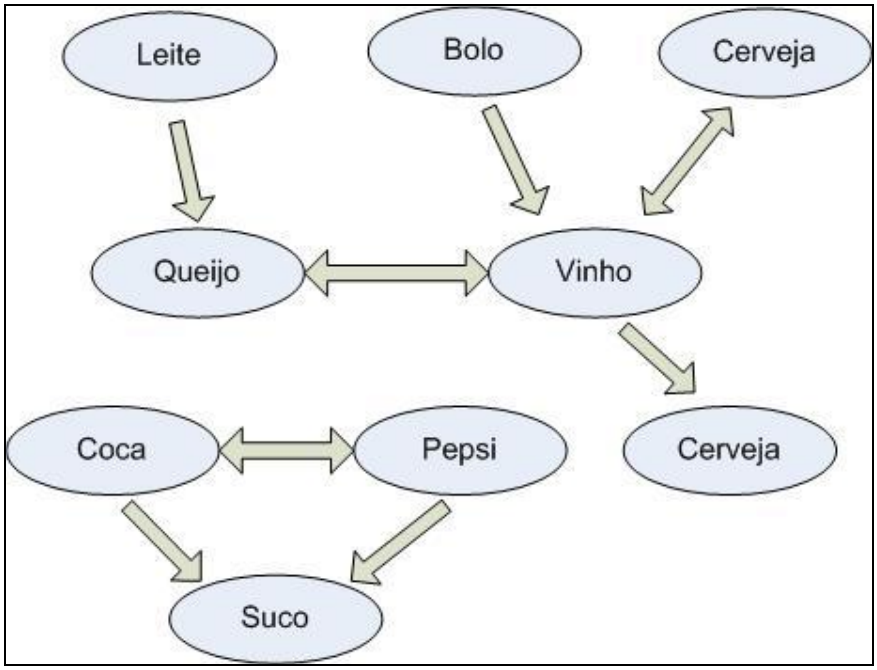

Figura 4 - Exemplo de análise de ligações (Associação de produtos)

e) Sumarização: esse procedimento determina uma descrição resumida de determinado subconjunto, através de técnicas de visualização e a determinação de relações funcionais entre variáveis. As funções de sumarização são responsáveis pela descrição compacta de um conjunto de dados. Esse método é utilizado, principalmente, no pré-processamento dos dados, quando valores inválidos são determinados por meio do cálculo de medidas estatísticas, no caso de variáveis quantitativas e no caso de outras variáveis, por meio da distribuição de freqüência dos valores. Como apontam Sferra e Corrêa (2003), técnicas de sumarização mais sofisticadas são imprescindíveis para se obter um entendimento, muitas vezes intuitivo, do conjunto de dados.

f) Detecção de desvios ou *outliers*: Esta técnica utiliza uma função média, representando o comportamento normal de um sistema, para avaliar possíveis desvios (FAYAAD *et al.*, 1996). Visa descobrir mudanças significativas ou desvios nos dados; conforme Sumathi e Sivanandam (*op. cit.*) incluem a pesquisa por desvios temporais (mudanças importantes nos dados de séries temporais) e desvios em grupos (diferenças não esperadas entre dois subconjuntos de dados).

2.3.2 - Técnicas utilizadas no *Data Mining*

Dependendo da tarefa a ser realizada diversas técnicas e algoritmos podem ser utilizados. Inicialmente desenvolvidos para outras áreas do conhecimento - redes neurais, algoritmos genéticos, modelos estatísticos e probabilísticos entre outros.

a) Análise fatorial – Técnica estatística multivariada utilizada para a redução de variáveis para um conjunto menor de fatores, através de um modelo que procura explicar a correlação entre as variáveis em estudo. A análise fatorial:

> [...] busca, através da avaliação de um conjunto de variáveis, a identificação de dimensões de variabilidade comuns existentes em um conjunto de fenômenos; o intuito é desvendar estruturas existentes, mas que não são observáveis diretamente. Cada uma dessas dimensões de variabilidade comum recebe o nome de FATOR. (BEZERRA, 2007).

No entendimento de Sharrma (1996) os objetivos desta técnica são:

- Identificar o menor conjunto de fatores comuns, que explique as correlações entre as variáveis;
- identificar o conjunto de fatores mais plausível;
- estimar o padrão e estrutura dos coeficientes e variâncias das variáveis;
- oferecer uma interpretação para os fatores calculados; e
- estimar os valores dos fatores.

b) Análise de agrupamentos - De acordo com Sferra e Corrêa (*op. cit.*), os agrupamentos de dados são baseados em medidas de similaridade ou modelos probabilísticos. A análise de agrupamentos é uma técnica que visa detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso de sua existência, determinar quais são eles. Pode-se destacar que o objetivo desta técnica é agrupar os elementos de um grande conjunto de dados em grupos significantes, conforme a sua proximidade ou características comuns, buscando mostrar a homogeneidade dentro do grupo e a heterogeneidade entre os grupos (Johnson e Wichern, 1992). De forma geral a análise de agrupamentos compreende duas etapas:

- A escolha de uma medida de proximidade;
- A escolha de um algoritmo para construção dos grupos.

c) Análise Discriminante - Os objetivos dessa técnica envolvem a descrição gráfica ou algébrica das características diferenciais das observações de várias populações, além da classificação das observações em uma ou mais classes predeterminadas, sendo ideal para identificar quais variáveis melhor separam uma população em grupos distintos. Possibilita a classificação de novos casos, onde a inclusão é feita no grupo onde a nova ocorrência, com base em suas características, apresenta maior probabilidade de pertencer.

d) Regressão Linear – Examina o relacionamento entre uma variável dependente e uma ou mais variáveis independentes. A Figura 5 ilustra o relacionamento entre vendas e propaganda a partir de uma equação de regressão. O objetivo é prever vendas, baseado no total gasto com propaganda, no exemplo extraído de Rud (2001) a medida R^2 indica que 70% da variação apresentada nas vendas pode ser explicada pela inclusão da variável gasto com propaganda no modelo.

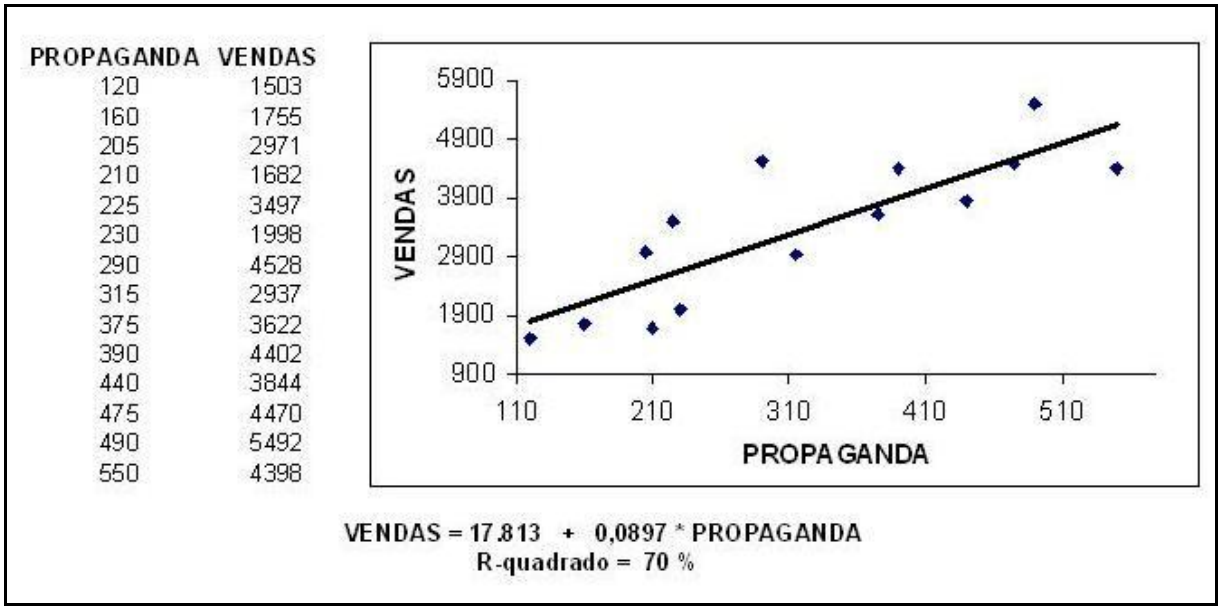

Figura 5 - Regressão linear simples (Extraído de Rud, 2001)

e) Regressão Logística – É uma variação da regressão linear, que permite a predição de um evento, geralmente utiliza variáveis dependentes binárias para se obter a possibilidade probabilística do evento binário ocorrer. As variáveis independentes podem ser discretas ou contínuas. Na Figura 6 é exibido um exemplo de aplicação da técnica de regressão logística para detecção da probabilidade de um indivíduo vir a desenvolver uma determinada doença.

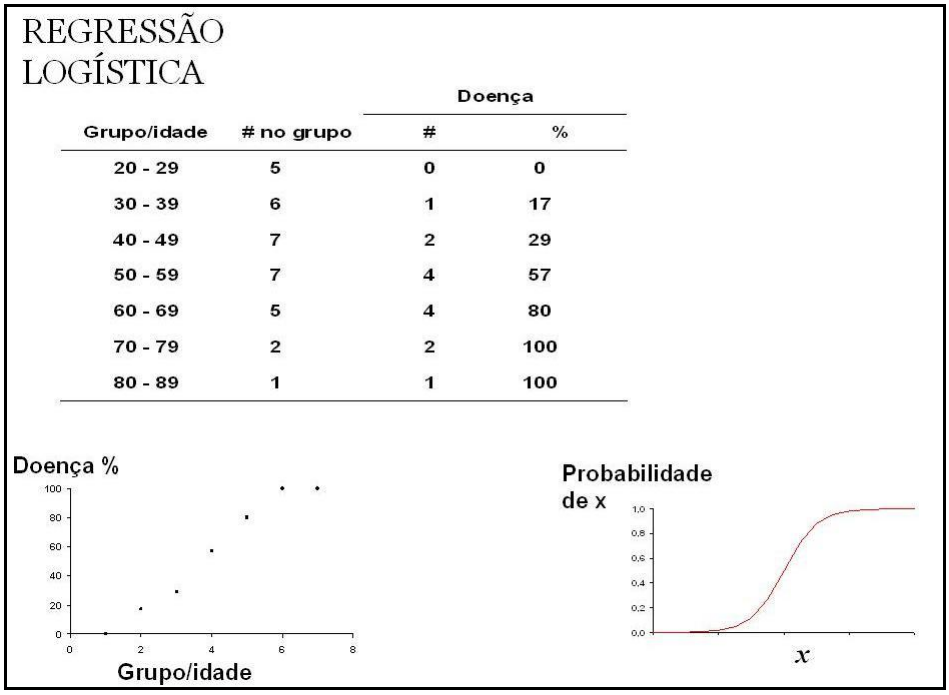

Figura 6 - Exemplo de regressão logística

f) Árvores de decisão - Para Sumathi e Sivanandam (*op. cit.*): "Estruturas em formato de árvore que representam conjuntos de decisões que geram regras para a classificação de um conjunto de dados". O propósito das árvores de decisão é classificar os dados em grupos distintos que formem a mais forte separação nos valores das variáveis dependentes. A Figura 7 apresenta um exemplo de árvore de decisão para classificação de indivíduos em relação à sua altura e sexo.

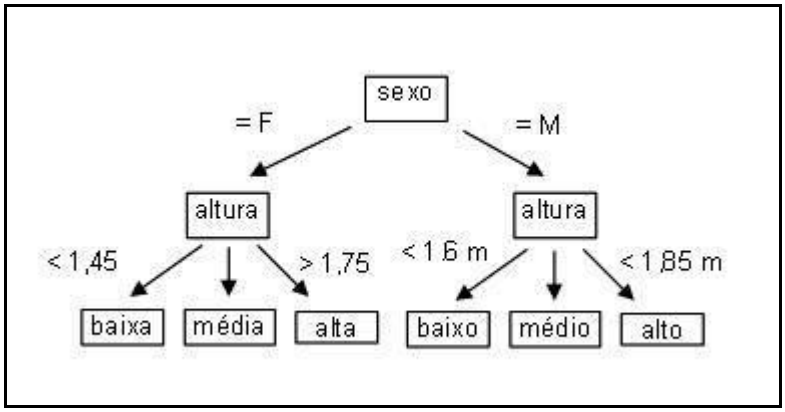

Figura 7 – Árvore de decisão (Extraído de Rud, 2001)

g) Escalonamento Multidimensional – "É uma técnica baseada nas proximidades entre os objetos, temas ou estímulos utilizada para produzir uma representação espacial destes itens" (HARDLE e SIMAR, 2007). Considerada uma técnica de redução de dados, uma vez que procura apresentar as características dos dados em poucas dimensões (geralmente 2 ou 3). Ainda conforme os autores acima, o escalonamento multidimensional geralmente é utilizado para ilustrar como as pessoas percebem e avaliam determinadas informações.

No exemplo apresentado por Herdeiro (2007) foi realizada uma sondagem sobre o posicionamento de algumas marcas nacionais de cerveja. Foram selecionadas 10 marcas e feito o julgamento destas marcas por 20 pessoas apreciadoras de cerveja. Após a análise dos dados e emprego da técnica de escalonamento multidimensional para redução a 2 dimensões os resultados puderam ser exibidos pela Figura 8.

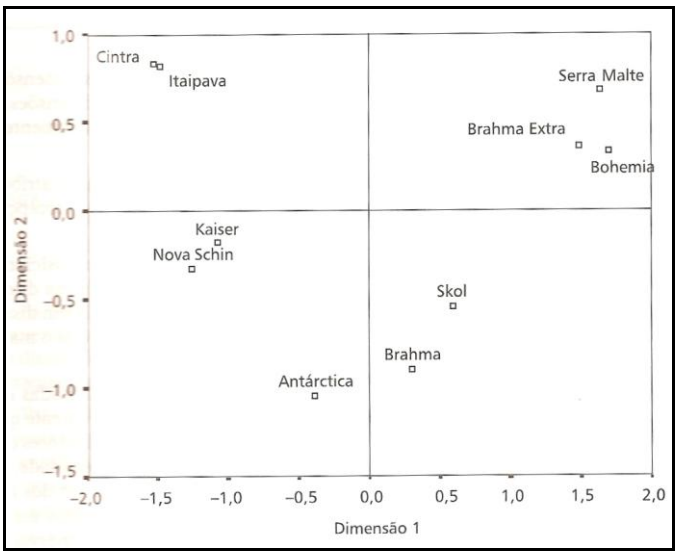

Figura 8 - Exemplo de escalonamento multidimensional (extraído de Herdeiro, 2007)

h) Algoritmos Genéticos – Para Sumathi e Sivanandam (*op. cit.*): "Técnicas de otimização baseadas nos conceitos de combinação genética, mutação, e seleção natural". A partir de uma solução inicial viável os algoritmos genéticos criam iterativamente novas soluções que aos poucos vão melhorando os resultados iniciais. São métodos adaptativos que podem ser usados para resolver problemas de busca e otimização. No desenvolvimento da melhor solução simulam os processos naturais de evolução, utilizando operadores de seleção, cruzamento e mutação para desenvolver sucessivas gerações de soluções.

i) Redes Neurais - Modelos preditivos não lineares que visam aprender com treinamento e em sua estrutura lembram redes neurais biológicas. “Podem ser definidas como sistemas computacionais compostos por inúmeros elementos de processamento, interconectados de acordo com uma topologia especifica (arquitetura) e com capacidade de modificar seus pesos de conexão e parâmetros dos elementos de processamento (aprendizado)” (ZORNETZER *et al.*, 1994).

Aprender, errar e fazer descobertas são os conceitos que norteiam esta técnica. De forma simplificada, uma rede neural artificial possui um sistema de neurônios ou nós e conexões com pesos ponderados. A Figura 9 mostra um exemplo de funcionamento de redes neurais onde os nós de entrada (I) recebem informações das diversas variáveis de entrada; os nós escondidos (H) recebem os valores de entrada *Wn* e as processa segundo uma função de ativação; e, os nós de saída (O) que informam a estimativa calculada pela rede neural.

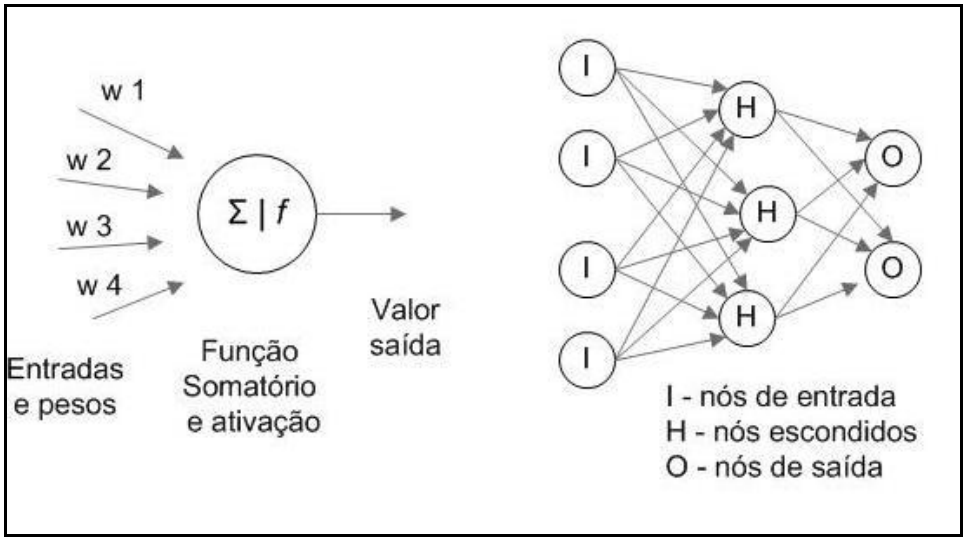

Figura 9 - Exemplo rede neural (Extraído de Rud, 2001)

O processo que busca a melhor definição dos pesos *Wn* é conhecido como processo de aprendizagem ou treinamento da rede. A rede está aprendendo ou treinando para produzir as transformações desejadas, até encontrar um padrão adequado a partir dos inputs apresentados. Cada interação na etapa de treinamento ajusta os pesos associados a cada nó, desta forma a rede neural melhora a “predição” dos dados.

2.4 - METODOLOGIAS DE *DATA MINING* E CRISP-DM

Para se obter melhores resultados no emprego do *data mining*, é necessário adotar regras e padrões formalizados, isto requer que as organizações usem uma abordagem sistemática para conseguir resultados que tenham utilidade.

Em 1996 foi desenvolvida a metodologia CRISP-DM (*CRoss-Industry Standard Process for Data Mining*), por um consórcio de consultores e especialistas em *data mining*, que incluía SPSS, Daimler-Benz (posteriormente DaimlerChrysler) e NCR[4]. Os idealizadores do CRISP-DM utilizaram sua experiência do mundo real para desenvolver um processo em seis fases que incorporasse os objetivos da organização e conhecimento. CRISP-DM é considerado o padrão para a indústria de *data mining*. A Figura 10 ilustra as fases do CRISP-DM.

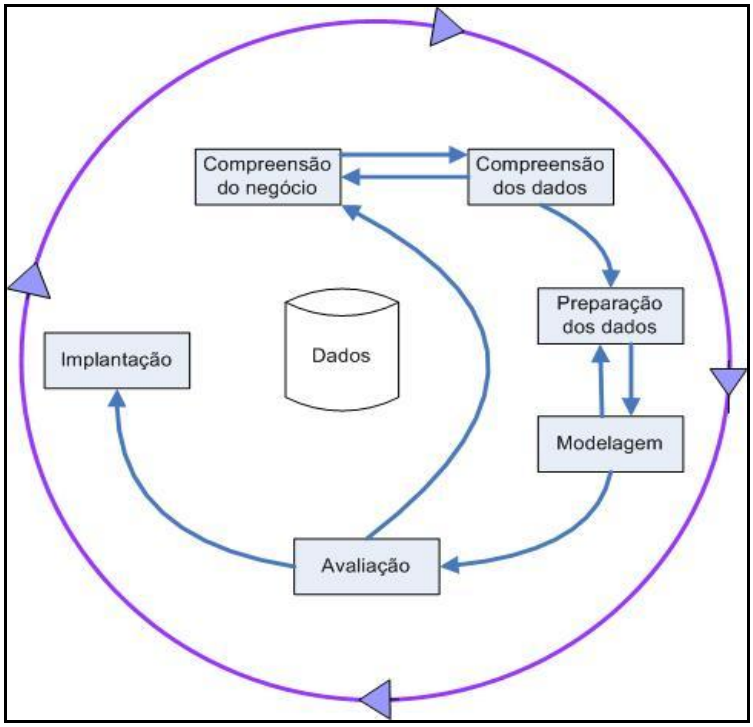

Figura 10 - Fases do CRISP-DM (baseado em the CRISP-DM Consortirum, 2000)

a) Compreensão do negócio - O elemento-chave para um estudo de *data mining* é saber a razão do processo, que inicia com a necessidade de conhecimento novo, onde os objetivos do negócio devem ser considerados. A primeira fase permite que todos os participantes do

[4] CRISP-DM Web site: www.crisp-dm.org.

projeto de *data mining* compreendam os objetivos a partir de uma perspectiva organizacional ou de negócios. Estes objetivos de negócio são então incorporados na definição do problema de *data mining* formando um detalhado plano de projeto.

Para uma agência de tributos, esta fase pode envolver a compreensão do processo de gerenciamento de auditoria, as regras e funções executadas, as informações que são coletadas e gerenciadas, e as mudanças específicas para melhorar a eficiência da auditoria. Esta informação deve ser incorporada na definição do problema de *data mining* e no projeto. (MICCI-BARRECA e RAMACHANDRAN, 2006).

b) Compreensão dos dados – Tem por objetivo a familiarização com o banco de dados pelos participantes do projeto. Envolve a coleta e análise preliminar dos dados.

Existem pelo menos três questões a serem consideradas na seleção dos dados. A primeira é estabelecer uma descrição clara e concisa do problema. A segunda é identificar dados relevantes para o problema descrito. A terceira é assegurar que as variáveis selecionadas serão independentes uma das outras. Uma seleção cuidadosa de variáveis independentes pode facilitar a construção de modelos para a obtenção de conhecimento (OLSON e DELEN, 2008).

A compreensão dos dados é projetada para avaliar as fontes, qualidade e características dos dados. Esta exploração inicial pode também prover introspecções que ajudem a focar no projeto. O resultado é uma compreensão detalhada dos elementos chave que serão usados para construir os modelos. No entendimento de Micci-Barreca e Ramachandran (*op. cit.*): "Esta fase pode consumir muito tempo, mas é crucialmente importante para o projeto".

c) Preparação dos dados - O propósito da preparação dos dados é melhorar a qualidade dos dados selecionados, tornando-os mais claros. Alguns dados selecionados podem ter diferentes formatos por serem escolhidos através de diferentes fontes de dados.

Esta fase envolve a colocação dos dados em um formato apropriado para a construção de modelos, envolve limpeza, transformação, integração e formatação dos dados. Conforme Micci-Barreca e Ramachandran (2006): "O analista usa os objetivos do negócio determinados no passo de conhecimento do negócio para determinar que tipo de dados e algoritmos de *data mining* usar. Esta fase também trata dos problemas com os dados, como dados faltantes".

d) Modelagem - A fase de modelagem envolve a construção de algoritmos de *data mining* que extraem o conhecimento dos dados. Existe uma variedade de técnicas de *data mining*; cada uma é apropriada para descobrir um tipo especifico de conhecimento. De acordo com Larose (2006), cada técnica requer tipos específicos de dados, que podem indicar um retorno à fase de preparação de dados.

A fase de modelagem produz um modelo ou um conjunto de modelos contendo o conhecimento descoberto em um formato apropriado (MICCI-BARRECA e RAMACHANDRAN, *op. cit.*).

e) Avaliação - Esta fase foca na avaliação da qualidade do modelo ou modelos. Os algoritmos de *data mining* podem descobrir um número ilimitado de padrões; muitos deles, entretanto, podem ser sem sentido ou não apresentam interesse para o problema em questão. Esta fase ajuda a determinar quais modelos são úteis para alcançar os objetivos do projeto de negócio, conforme estabelecido na primeira fase. É avaliado, também, se os resultados obtidos com o modelo podem ser efetivamente implantados no ambiente operacional da organização.

f) Implantação - Na fase de implantação, a organização incorpora os resultados do *data mining* ao processo de decisão do dia-a-dia. Dependendo da significância dos resultados, isto pode requerer apenas pequenas modificações, ou ele pode demandar uma maior reengenharia do processo e dos sistemas de suporte à decisão. De acordo com Micci-Barreca e Ramachandran (*op. cit.*), a fase de implantação também envolve a criação de processos repetitivos para o aperfeiçoamento do modelo ou sua re-calibração. A legislação tributária, por exemplo, é comum ser alterada ao longo do tempo. Os analistas precisam de um processo padrão que possibilite atualizações adequadas dos modelos e assim, novas resultados serão obtidos.

Ainda, segundo os autores: “A apresentação apropriada dos resultados garante que os gerentes usem as informações adequadamente. Isto pode ser tão simples quanto a criação de um relatório ou tão complexo quanto a implementação de um processo repetitivo de *data mining* em toda empresa.” (MICCI-BARRECA e RAMACHANDRAN, *op. cit.*). É importante que os responsáveis pelo projeto compreendam desde o início quais ações eles necessitarão tomar com o objetivo de fazer uso dos modelos finais.

As seis fases descritas se aplicam a todo projeto de *data mining*. Embora cada fase seja importante, a seqüência não é rígida e certos projetos podem requerer que se desloque

para frente ou para trás entre as fases. A próxima fase ou próxima tarefa depende do resultado de cada uma das fases anteriores. As setas internas na Figura 10 indicam a mais importante e freqüente dependência entre as fases. O circulo exterior simboliza a natureza cíclica dos projetos de mineração, as lições aprendidas durante um projeto de *data mining* e após a implantação podem trazer novas e mais focadas questões de negócios. Projetos subseqüentes de *data mining*, entretanto, se beneficiam da experiência obtida em projetos passados.

A maior parte do tempo empregado em um projeto de *data mining* típico é gasto em outras fases que não seja a de modelagem, e o sucesso dos modelos depende muito do trabalho executado nestas fases.

2.5 - *DATA MINING* NA DETECÇÃO DE FRAUDE

A detecção de fraude se tornou uma das principais aplicações de *data mining*, com muitas inovações e a inclusão crescente de novas técnicas e metodologias. Bonchi *et al.* (2008) relatam que a principal tarefa, nesse âmbito, é a construção de modelos ou perfis que indiquem um comportamento fraudulento, possibilitando sua averiguação. Os objetivos primários são prevenir fraudes e planejar estratégias de auditoria para a detecção de fraudes existentes. As organizações usam as informações dos bancos de dados e o resultado do *data mining* para detectar fraude existente e não-conformidade ou ainda para prevenir ocorrências futuras.

Phua *et al.* (2005) apresentam uma extensa revisão da aplicação de *data mining* na detecção de fraude a partir de 51 trabalhos publicados sobre o assunto. Os setores com mais referências foram: seguros, cartão de crédito, telecomunicações, crimes financeiros e fiscais, detecção de envio de mensagens eletrônicas indesejadas (*spam*). Os autores apresentam também as principais técnicas utilizadas nesta área, quantidade de variáveis por publicação, percentual de fraude encontrada nos referidos trabalhos. Dentre as técnicas referenciadas obtiveram destaque: árvore de decisão, redes neurais, sistemas especialistas, regras de associação, algoritmos genéticos e regressão logística.

Todas as experiências indicadas na seqüência contemplam casos publicados no Brasil e também os trabalhos não apresentados na pesquisa citada acima.

Kotsiantis *et al.* (2006) relatam um estudo para detecção de fraude financeira na Grécia com 164 declarações (sendo 41 casos confirmados de fraude) foram utilizadas 26

variáveis para elaboração dos modelos de redes neurais, árvores de decisão e regressão logística entre outros. Na avaliação dos autores os modelos utilizando árvore de decisão e redes neurais obtiveram os melhores resultados no quesito acurácia na detecção de casos com e sem fraude.

Virdhagriswaran e Dakin (2006) descrevem um sistema de *data mining* para detectar fraudes "camufladas" como dados normais em grandes volumes de relacionamentos conhecidos. Neste sistema foram criados 2 estágios para se obter o resultado desejado: 1) detecção de exceção, selecionar características e classificar os dados baseados nestas características; e 2) previsão, onde as características detectadas são utilizadas para prever as direções das mudanças nas variáveis selecionadas. Para o primeiro estágio foram empregadas árvores de decisão e regressão logística e no segundo estágio, previsão, utilizou-se regressão linear.

Em sua dissertação de mestrado, Queiroga (2005) propõe a utilização de redes neurais e bayesianas para detecção de fraude em energia elétrica, a experiência foi realizada na ESCELSA, Espírito Santo Centrais Elétricas S. A., em 70 municípios do Estado, utilizando dados dos anos de 2003 e 2004. O autor destaca uma melhoria significativa na identificação de perdas comerciais, saindo de um índice de acerto nas inspeções selecionadas pelo método tradicional de 7% a 25%, para 25% a 45% com a utilização de técnicas de classificação exploradas neste trabalho.

Passini e Toledo (2008) aplicaram técnicas de *data mining* para detecção de fraudes em ligações de água na Sociedade de Abastecimento de Água e Saneamento S.A. (SANASA) em Campinas (SP). Os dados deste estudo foram extraídos dos sistemas operacionais, a partir de 10 arquivos com dados históricos. O objetivo era a redução, inicialmente de 51% para 41%, das visitas infrutíferas para detecção de fraudes. Apesar de não atingirem esta meta, houve um ganho em relação ao conhecimento adquirido da tecnologia e à experiência na utilização de uma ferramenta de mineração. As variáveis utilizadas foram divididas em três categorias: informações cadastrais (ou domiciliares), relacionamento do cliente com a empresa, e as informações sobre o comportamento do cliente ao longo do tempo. As técnicas empregadas no estudo foram: análise fatorial, redes neurais e árvore de decisão.

No caso de fraude tributária, o procedimento de *data mining* visa selecionar as empresas que apresentam um comportamento fora do padrão, que não está seguindo o fluxo mercadológico das outras empresas ou que apresenta informações de movimentação financeira diferentes das apresentadas nos anos anteriores, levantando suspeitas de um comportamento anormal. Gupta e Nagadevara (2007) apresentam uma estratégia para seleção

de auditoria com utilização de *data mining* realizada na Índia visando identificar os contribuintes com maior probabilidade de estarem registrando impostos com valores menores dos que os realmente ocorridos. Foram selecionadas 28 variáveis para a análise e os modelos se concentraram em tarefas de classificação: árvores de decisão, regressão logística e análise discriminante, sendo que esta última técnica resultou em importante melhoria no desempenho do processo de seleção dos contribuintes.

Outro importante trabalho na área de seleção de contribuintes para fins de auditoria fiscais é o de Micci-Barreca e Ramachandran (2006) descrito no capítulo anterior.

2.6 – CONCLUSÕES DO CAPÍTULO

Neste capítulo foram apresentadas características gerais sobre *Data Mining*, as suas principais tarefas e técnicas. Por reunir conhecimento de diversas áreas científicas e possibilitar o processamento de grandes volumes de dados, o *data mining* tem despertado interesse crescente dos pesquisadores. Este interesse se reflete numa vasta gama de aplicações para a matéria, e na seção 2.5 deste capítulo é feita uma revisão das aplicações mais relevantes na área de detecção de fraudes.

Partindo-se do referencial exposto, pode-se acompanhar na seqüência o desenvolvimento do modelo sugerido, objeto desta tese.

3 - MODELO PROPOSTO

3.1 - INTRODUÇÃO

O modelo apresentado baseia-se em duas fases principais: a construção de agrupamentos que permitam uma caracterização homogênea das empresas dentro de cada grupo, e a construção de uma série de equações de regressão logística para se obter as probabiliddes de uma empresa estar incorrendo em irregularidades nas suas declarações de movimentação financeira mensal.

Para uma correta execução das técnicas de *data mining* deve-se trabalhar as fontes de dados disponíveis, a primeira atividade do presente modelo aborda este aspecto. Ao final do capítulo é apresentado um quadro resumo que ilustra todos os passos do modelo.

3.2 – CONSIDERAÇÕES GERAIS

No entendimento de Dempster (1998) o termo modelo representa uma estrutura abstrata, um conjunto definido de ligações entre esta estrutura e o mundo real, veiculado em parte por nomes dados a entidades do modelo juntamente com o material descritivo sobre os metodos experimentais e de amostragem.

Para construção do presente modelo serão utilizados diversos métodos e técnicas disponíveis na literatura. O método é, segundo Lakatos e Marconi (1982, p. 41), "o conjunto das atividades sistemáticas e racionais que, com maior segurança e economia, permite alcançar o objetivo, traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do cientista". Conforme Bomfim (1999):

> A palavra método deriva do vocabulário grego "méthodos" e significa "caminho para alguma coisa", "seguir alguma coisa" ou "andar ao longo de um caminho". Neste sentido, método é a previsão de alguma tarefa que se desenvolve de um modo consciente e objetivo, ou seja, no senso comum um método é o planejamento que antecede uma tarefa.

Para que se possa seguir um método é preciso utilizar técnicas. A técnica se refere ao modo de fazer de forma mais hábil, mais segura e mais perfeita algum tipo de atividade, arte ou ofício. Desta forma método trata sobre “o que fazer”, sobre a orientação geral da atividade e a técnica trata sobre o “como fazer” a atividade.

O desenvolvimento do presente modelo de classificação de contribuintes não tem como objetivo desenvolver ferramentas computacionais ou criar novas técnicas de análise, mas preestabelecer conjuntos ordenados de regras e tarefas a serem seguidas, a fim de realizar processos de KDD e alcançar resultados satisfatórios.

3.3 - DADOS DISPONÍVEIS

As empresas que estão constituídas para desempenharem atividades econômicas devem estar legalmente regularizadas para tal e precisam seguir as regras definidas pelos órgãos que regulamentam estas atividades. Para acompanhar e regular este processo, os órgãos fiscalizadores exigem que diversos documentos e declarações sejam apresentados pelas empresas contribuintes. Dentre estas informações merecem especial atenção para os propósitos deste modelo:

- Dados estruturais - com dados das atividades econômicas, onde constam informações da variação dos estoques, faturamento, receita bruta, entre outras.
- Dados financeiros - com informações que devem refletir os lançamentos financeiros efetuados pelo contribuinte. Esta declaração tem por finalidade demonstrar o imposto calculado em cada período de apuração, bem como apresentar outras informações de interesse econômico-fiscal;
- Registro de auditorias realizadas nas empresas contribuintes com indicativo da ocorrência de sonegação se for o caso.

Cabe ressaltar que em cada entidade fiscalizadora estas declarações se apresentam com características próprias, inclusive com nomenclaturas e quantidades diversas de campos a serem informados. Entretanto, em sua essência todas as declarações contêm dados similares

que possibilitam a classificação e enquadramento das empresas, bem como o acompanhamento da evolução histórica do seu comportamento econômico.

Outras importantes fontes de dados também estão disponíveis e se originam fora da estrutura das entidades fiscalizadoras, tais como: dados de consumo de energia elétrica, informações de outros órgãos públicos, especialmente de entidades trabalhistas que possibilitam acompanhar a utilização, ou alocação, destes recursos.

3.4 - DETALHAMENTO DO MODELO

A aplicação do modelo proposto divide-se em quatro etapas, detalhadas a partir do item 3.4.1. O primeiro nível do modelo é organizado em etapas, que são divididas em atividades. O modelo proposto define um processo completo e sistemático de seleção de contribuintes utilizando técnicas de *data mining*.

Formulários específicos deverão ser elaborados para este modelo de classificação subsidiar a escolha dos procedimentos a serem adotados. Os artefatos gerados acima servirão para a documentação de todo o processo.

3.4.1 - Obtenção dos dados

A primeira etapa do modelo constitui-se na obtenção das informações necessárias para compor o banco de dados que irá subsidiar todas as demais etapas do processo. De acordo com Berry e Lynoff (2004), devem ser identificadas as fontes de dados e efetuado o tratamento necessário para melhorar a assimilação dos dados por modelos quantitativos. Esta etapa é composta de quatro atividades:

a) Acesso aos dados: esta atividade compreende a identificação de quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o processo. Deve-se considerar a regularidade com que os dados estarão disponíveis. Ao final desta atividade deve-se produzir uma relação com todos os dados que serão utilizados, bem como sua fonte e forma de obtenção.

b) Limpeza dos dados: de posse dos dados disponíveis deve-se criar uma base de dados para análise. Nesta atividade serão utilizados procedimentos para assegurar a qualidade dos dados (completude, veracidade e integridade). Informações ausentes, errôneas ou inconsistentes nas bases de dados devem ser corrigidas de forma a não comprometer a qualidade dos modelos a serem desenvolvidos.

c) Análise de relevância: todos os documentos de arrecadação possuem uma quantidade grande de dados quantitativos e nem todos os contribuintes são obrigados a preencher todos os campos, com isto diversas informações deixam de apresentar interesse para as próximas etapas. Outro problema que pode ocorrer é a existência de correlação entre os campos utilizados para a análise. Tal situação pode ocorrer devido às regras de preenchimento das declarações que muitas vezes exigem o registro de valores que são meras operações aritméticas entre outros campos já informados no mesmo documento. Para tanto esta atividade deve selecionar quais variáveis serão utilizadas em cada etapa seguinte da metodologia.

d) Transformação dos dados: para a elaboração dos modelos de *data mining* os dados devem estar dispostos no formato de matriz de dados (em linhas e colunas). Nesta atividade, além da criação da matriz deve-se proceder à transformação, caso necessário, dos dados visando à melhoria da capacidade preditiva dos modelos. Esta transformação pode-se dar por intermédio da padronização dos dados, como no caso de campos com grande disparidade nos seus valores.

Na Tabela 1 é apresentada a relação das atividades e respectivas saídas produzidas nesta primeira etapa.

Tabela 1 - Entradas e saídas das atividades da etapa 3.4.1

Atividades	Saídas
Acesso aos dados	Relação de tabelas e dados disponíveis
Limpeza dos dados	Base de dados para análise (*datamart*)
Análise de relevância	Relação de variáveis selecionadas
Transformação dos dados	Matriz de dados para análise

3.4.2 - Criação de grupos

Essa etapa é responsável pela criação de agrupamentos onde possam ser divididas as empresas contribuintes. Na formação dos grupos deve-se levar em consideração as características regionais de cada empresa, bem como as atividades econômicas de cada contribuinte. De acordo com Sumathi e Sivanandam (2006): "Agrupamento é freqüentemente um dos primeiros passos na análise de *data mining.* Ele identifica grupos de registros relacionados que podem ser usados como ponto de partida para futuros relacionamentos". Esta etapa é composta de duas atividades:

a) Escolha do método de agrupamento: conforme a quantidade e a natureza dos dados disponíveis deve-se escolher qual a melhor forma de agrupamento utilizar. Os métodos para montagem dos grupos, basicamente se dividem em dois algoritmos:

a.1) Algoritmos hierárquicos – "Os procedimentos hierárquicos envolvem a construção de uma hierarquia semelhante a uma árvore. Existem basicamente dois tipos de procedimentos hierárquicos: aglomerativos e divisivos" (POHLMANN, 2007):

- Aglomerativo - os grupos são formados a partir dos ramos até a raiz. Cada ponto forma um grupo; após o cálculo da matriz de distância, os grupos mais próximos são unidos, o procedimento se repete até que todos os pontos estão unidos em um único grupo.
- Divisivo - são formados a partir da raiz até chegar aos ramos. Inicialmente é formado um único grupo com todas as *n* observações em seguida são divididas em dois grupos, a partir de então cada grupo é novamente dividido em dois, até chegarmos a grupos com apenas uma observação.

a.2) Algoritmos não Hierárquicos (de Particionamento) – permitem obter uma partição de *n* observações em *g* grupos ($g < n$), sendo *g* definido *a priori*. Os algoritmos não hierárquicos geralmente são muito mais rápidos que os hierárquicos, sendo mais indicados para grandes volumes de dados. É necessário estabelecer o número desejado de grupos, que deve ser precedido de estudo da natureza do problema e testes de formação de grupos com diversos valores de *g*. De acordo com Rencher (2002) a estratégia de particionamento deve examinar

todos os caminhos possíveis para particionar as *n* observações em *g* grupos e otimizar a criação dos grupos baseados em um dado critério.

O método mais utilizado nos algoritmos não hierárquicos é o de *k-médias* (BOTTOU e BENGIO, 1995), onde inicialmente é escolhida a alocação de algumas observações nos *g* grupos (esta alocação inicial pode ser aleatória, ou das primeiras observações, ou dos pontos mais distantes entre si, entre outros) então é criado um vetor de médias (centróide) para cada grupo. Na seqüência cada elemento dos *g* grupos é realocado para o grupo cujo centróide se encontre à menor distância euclidiana. Após a realocação o vetor de médias para o grupo deve ser recalculado e o processo é reiniciado.

O método de aglomeração denominado de *two-step* permite a utilização de variáveis com características contínuas e categóricas e recebe este nome, pois a aglutinação é realizada em dois estágios. Quando as variáveis são contínuas o método utiliza medidas como média e variância e quando são categóricas utiliza a medida de contagem das variáveis. Este método além de tratar variáveis contínuas e categóricas, é útil para formar grupos quando não se tem idéia de quais variáveis pertencem a cada grupo, pois é possível determinar um número de agrupamentos que se deseja.

De acordo com Zhang *et al.* (1996) realizada esta diferenciação o método realiza uma primeira aglutinação dos dados originais em subgrupos de modo a obter um número de subgrupos possível de realizar a análise. Logo após o método utiliza o agrupamento hierárquico de modo que paulatinamente os grupos formados na primeira etapa são fundidos até formar um grande cluster.

A distância entre dois grupos *j* e *s* é definida como a redução em log-verossimilhança devido à união em dois clusters:

$$D(j, s) = \xi_j + \xi_s - \xi_{<j, s>} \qquad (1)$$

Conforme Bacher et al. (2004), ξv pode ser interpretado como um tipo de dispersão (variância) dentro do cluster *v* sendo *(v = j, s, <j, s>)*, e:

$$\xi_v = N_v \left(\sum_{k=1}^{k^A} \frac{1}{2} \log(\sigma_k^2 + \sigma_{vk}^2) + \sum_{k=1}^{k^B} \widehat{E}_{vk}^2 \right) \qquad \text{e} \qquad \widehat{E}_{vk} = -\sum_{l=1}^{l_k} \frac{N_{vkl}}{N_v} \log \frac{N_{vkl}}{N_v}$$

e onde k^A é o total de variáveis contínuas, k^B é o total de variáveis categóricas, l_k é o número de categorias para a *k -ésima* variável categórica, N_j é o número de observações no cluster *j* , σ_k^2 é a variância da *k -ésima* variável contínua no conjunto de dados original, σ_{jk}^2 é a variância da *k -ésima* variável contínua no cluster *j*, N_{jkl} é o número de observações no cluster *j* cuja *k -ésima* variável categórica possui a *l -ésima* categoria, e < *j*, *s* > representam o cluster formado pela união dos grupos *j* e *s* .

No cálculo da log-verossimilhança, assume-se que as variáveis contínuas possuem distribuição normal, e as variáveis categóricas tenham distribuição multinomial. O primeiro passo, pré-agrupamento, adota a abordagem de cluster conforme desenvolvida por Zhang *et al.* (*op. cit.*). A característica típica CF_j para um cluster C_j é, segundo Chiu *et al.*, (2001):

$$CF_j = \left\{ N_j, s_{Aj}, s_{Aj}^2, N_{Bj} \right\} \qquad (2)$$

Onde S_{Aj} é a soma das variáveis contínuas no cluster C_j , S^2_{Aj} é a soma das variáveis contínuas ao quadrado no cluster C_j , e $N_{Bj} = (N_{Bj1}, N_{Bj2}, ..., N_{Bjk})$ é um vetor dimensional $\sum_{k=1}^{k^B}(L_k - 1)$

cujo *k -ésimo* sub-vector é de tamanho (L_k -1).

Quando dois clusters C_j e C_s são unidos, a característica $CF_{<j\ s>}$, pode ser obtida usando a equação 3 (Chiu *et al.*, *op. cit.*):

$$CF_{<j,s>} = \left\{ N_j + N_s, s_{Aj} + s_{As}, s_{Aj}^2 + s_{As}^2, N_{Bj} + N_{Bs} \right\} \qquad (3)$$

Comparados com as técnicas de cluster K-médias e hierárquico (Chiu *et al.*, *op. cit.*), a estrutura CF ganha uma grande quantidade de tempo para a análise de cluster *TwoStep*. O número ótimo de cluster pode ser determinado usando os critérios Bayesianos ou Informação Akaike (BIC e AIC). Para *J* clusters, estes critérios podem ser calculados, conforme Chiu *et al.*(*op. cit.*):

$$BIC(J) = -2\sum_{j=1}^{j} \xi_j + m_j \log(N) \qquad (4)$$

$$AIC(J) = -2\sum_{j=1}^{j} \xi_j + 2m_j \qquad (5)$$

onde:

$$m_j = J\left\{2K^A + \sum_{k=1}^{K^B}(L_k - 1)\right\} \qquad (6)$$

Sendo que o menor valor (BIC ou AIC) representa o melhor modelo. Em alguns softwares estatísticos também é apresentada a razão das distâncias medidas (*Ratio of Distance Measures*) que é baseado no número corrente de grupos contra a quantidade anterior de grupos, deve-se optar pelos maiores valores para termos o melhor modelo.

Tanto os métodos hierárquicos quanto o *two-step* possuem suas vantagens e desvantagens na aplicação, cabendo ao pesquisador administrar este *tradeoff* entre os métodos. O primeiro é capaz de rapidamente realizar o agrupamento, mas perde na objetividade dos resultados, pois depende da subjetividade do analista para decidir o número de subgrupos a serem formados para uma posterior análise. O método *two-step* possui a vantagem de se pré-determinar o número de cluster a ser utilizado, mas possui a desvantagem de demandar mais tempo para sua execução (Park *et al.*, 2005).

b) Elaboração dos grupos: aplicar o método selecionado e gerar uma relação com os grupos elaborados. Nesta atividade é importante a verificação da disponibilidade de dados, em quantidade suficiente, de auditorias realizadas em empresas pertencentes a cada grupo gerado. Deve-se considerar a restrição anterior no processo de escolha da quantidade de agrupamentos que serão elaborados.

Ao término desta atividade, a matriz de dados deverá ser atualizada com o respectivo grupo onde a empresa contribuinte foi classificada. A Tabela 2 ilustra as atividades e suas saídas produzidas nesta segunda etapa.

Tabela 2 - Entradas e saídas das atividades da etapa 3.4.2

Atividades	Saídas
Escolha do método de agrupamento	Método selecionado
Elaboração dos grupos	Relação dos grupos

3.4.3 - Construção dos modelos probabilísticos

Nesta etapa serão gerados modelos estatísticos para uma empresa ter apresentado declarações com indícios de sonegação. Dentre as técnicas estatísticas multivariadas será utilizada regressão logística, conforme indicações de trabalhos anteriores nesta área de fraude fiscal. Acerca desta escolha, Spathis (2002) apresentou um modelo para detectar declarações financeiras com indícios de sonegação utilizando regressão logística, com resultados satisfatórios. Neste trabalho a precisão do modelo ultrapassou 84%.

Dentre as técnicas de análise disponíveis para detecção de fraude as mais utilizadas são árvores de decisão, redes neurais, regras de associação, regressão logística e algoritmos genéticos. Phua *et al.* (*op. cit.*), relatam uma extensa revisão da aplicação destas técnicas, avaliando que tanto as soluções baseadas em árvore de decisão quanto regressão logística obtiveram resultados satisfatórios. Observa-se também na construção de *Balanced Scorecard* (sistema de gerenciamento baseado em indicadores e estratégias) a utilização da técnica de regressão logística em grande parte das aplicações, Hand (2009) destaca: "Apesar de uma grande variedade de abordagens estatísticas e de aprendizado baseado em automação ter sido investigado, por exemplo, redes neurais [...] entre outros, de longe o tipo mais popular de modelo é a regressão logística".

Para a realização desta etapa são necessários dados de auditorias já realizadas e que tenham campos categóricos informando se a empresa estava fraudando ou não, neste caso utiliza-se a regressão logística binária. Também é possível serem utilizados campos que representem uma escala de resultados referente à auditoria, geralmente codificações relativas a diversos delitos fiscais que podem ser classificados em categorias (por exemplo: sem infração, leve delito, mediano ou grande infração) neste caso deve ser utilizada regressão logística multinomial ordinal. Esta etapa apresenta duas atividades:

a) Identificação de variáveis: inicialmente devem ser identificadas variáveis que estejam associadas com as declarações incorretas. Pode ser utilizada a análise de correlação e também uma análise preliminar com regressão logística passo a passo (*stepwise*) para destacar as variáveis com mais relevância estatística para esta análise em particular.

b) Geração dos modelos estatísticos: a partir das variáveis mais significativas e utilizando também o código categórico do agrupamento de cada empresa, da etapa anterior, deve ser

elaborada uma equação de regressão logística que irá indicar um valor de probabilidade dos dados mensais da empresa estar com indícios de irregularidade.

O modelo de regressão logística é recomendado quando o interesse está na ocorrência ou não de um determinado evento, ou seja, a variável resposta Y é qualitativa dicotômica (assumindo os valores 0 e 1). No caso do processamento de grande volume de dados a Regressão Logística tem a vantagem de não ter pressupostos sobre a distribuição das variáveis preditoras (normalidade, linearidade ou variância igual dentro de cada grupo), sendo esta uma das principais vantagens apontadas por Dias Filho e Corrar (2007) para a utilização do modelo logístico.

Conforme apresentado por Giudici (*op. cit.*), o modelo logístico aplica-se para variáveis dependentes com valores 0 ou 1. O nível 1 usualmente representa a ocorrência de um evento de interesse. O modelo de regressão logística é definido a partir de valores ajustados que podem ser interpretados como probabilidades que um evento possa ocorrer em diferentes sub-populações.

$$\pi_1 = P(Y_i = 1) \quad \text{para i = 1, 2, …, n} \qquad (7)$$

Um modelo de regressão logística que especifica uma função apropriada das probabilidades ajustadas de um evento é uma função linear dos valores observados das variáveis explanatórias disponíveis.

$$Ln\left[\frac{\pi_i}{1-\pi_i}\right] = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} \qquad (8)$$

O lado esquerdo da equação define a função *logit* das probabilidades ajustadas.

$$\text{logit}(\pi_i) = \log\left[\frac{\pi_i}{1-\pi_i}\right] \qquad (9)$$

Invertendo a definição da função *logit*:

$$\pi_i = \frac{e^{(\alpha+\beta 1 Xi1+\beta 2 Xi2+...+\beta k Xik)}}{1+e^{(\alpha+\beta 1 Xi1+\beta 2 Xi2+...+\beta k Xik)}} \qquad (10)$$

Para uma única variável explanatória:

$$\pi_i = \frac{e^{(\alpha+\beta 1 Xi1)}}{1+e^{(\alpha+\beta 1 Xi1)}} \quad (11)$$

De acordo com Neter *et al.*(1996). Uma das propriedades da função logística é que ela pode ser linearizada. Denotando-se E(Y) por π, pois a resposta média é a probabilidade quando a variável resposta é binária. Após a transformação:

$$\pi' = \log_e\left(\frac{\pi}{1-\pi}\right) \quad (12)$$

Obtêm-se:

$$\pi' = \beta_0 + \beta_1 X \quad (13)$$

Este procedimento é chamado transformação *logit* da probabilidade π. A razão π/(1- π) na transformação *logit* é chamada de *Odds* (chance). A função resposta transformada (x) é denominada como função resposta *logit*, e π' de resposta média *logit*. A função logística pode ser empregada para:

DESCRITIVO – descrever a natureza do relacionamento entre a resposta média e uma (ou mais) variáveis regressoras.

PREDITIVO – saber se um evento irá ocorrer, dadas as características preditivas.

Embora a probabilidade de sucesso seja uma função logística, e, portanto não linear nas variáveis explanatórias, o logaritmo da chance é uma função linear:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta X \quad (14)$$

A expressão da razão de chance estabelece que o *logit* aumenta em β unidades para uma unidade de acréscimo em X. Para os modelos de regressão logística a chance de sucesso pode ser expressa:

$$\frac{\pi(x)}{1-\pi(x)} = e^{\alpha+\beta X} \qquad (15)$$

Este relacionamento exponencial oferece uma interpretação útil do parâmetro β: um aumento de uma unidade em X multiplica a chance por um fator e^{β}.

O procedimento usual para estimar os parâmetros é o da Máxima Verossimilhança e o objetivo é encontrar a melhor combinação linear de preditores que maximizem a verossimilhança de obter as freqüências observadas da variável de interesse. Este procedimento é iterativo, e começa com valores arbitrários para os coeficientes e determina a direção e o volume das mudanças nos coeficientes que irão maximizar a função de verossimilhança.

b.1) Teste de Wald - No processo de definição do modelo de regressão deve-se selecionar quais variáveis, dentre as disponíveis, serão utilizadas para se obter o melhor modelo. Para auxiliar a seleção destas variáveis existem procedimentos estatísticos específicos. Nesta pesquisa empregou-se o teste de Wald, conforme descrito em Kleinbaum e Klein (2002).

O teste de Wald é usado para testar a significância estatística de cada coeficiente β em um modelo. Este teste calcula uma estatística Z, que é um teste estatístico aproximadamente normal. O teste de Wald utiliza os erros padrões obtidos na estimação dos parâmetros, sendo calculado pela divisão dos coeficientes estimados pelo seu erro padrão. O quadrado desta estatística Z é aproximadamente uma estatística *qui* quadrado com um grau de liberdade.

c) Obtenção de um Conjunto de Equações: Da mesma forma que ocorre na elaboração dos grupos onde nem todos os dados da declaração anual são preenchidos por todas as empresas, nesta etapa também ocorre o mesmo com as declarações mensais. Como a técnica utilizada nesta metodologia, regressão logística, não processa os casos que contenham dados faltantes optou-se pela elaboração de um conjunto pequeno de equações que envolvam um grande número de empresas a serem processadas.

O procedimento sugerido baseia-se na redução do número de casos da base de dados de auditorias realizadas. Inicialmente devem-se verificar todas as combinações de variáveis

preditoras possíveis e os casos que atendem a cada combinação[5]. Em cada etapa do processo devem ser analisadas as combinações que tenham a mesma quantidade de variáveis e, selecionadas para elaborar as regressões logísticas, as combinações que apresentem uma quantidade de observações razoável para a obtenção de inferências de boa qualidade.

De acordo com Dias Filho e Corrar (op. cit.): "Entretanto, um ponto em relação ao qual a literatura ainda não apresenta consenso é a quantidade de observações necessárias à realização de inferências de boa qualidade." Mais adiante os autores prosseguem com algumas alternativas para facilitar a tarefa: "... uma regra razoável é obter um número de observações equivalente a pelo menos trinta vezes a quantidade de parâmetros que se deseja estimar. Em geral, há um certo consenso no sentido de que o modelo logístico requer amostras mais amplas do que os lineares.". Tanto Hosmer e Lemeshow (1989) quanto Peduzzi *et al.* (1996) recomendam ao menos 10 casos por variável independente.

Após diversos experimentos, para este modelo de classificação adotou-se o fator de 15 casos por variável independente, portanto devem ser elaboradas equações de regressão logística para todas as combinações com fator igual ou superior a 15. Serão utilizados todos os casos que atendam a cada combinação selecionada, sendo que na análise de regressão inicia-se com todos os casos da combinação e deve-se anotar os valores obtidos para o percentual de acertos da equação. Na seqüência procede-se a eliminação da variável que tenha menor contribuição estatística para explicar o comportamento dos dados (estatística Wald).

Após a retirada de uma variável, deve-se trazer da base de dados os casos que são contemplados pela nova formação de variáveis e novamente procede-se a análise de regressão. Neste ponto deve-se comparar o resultado dos percentuais de acertos de cada equação e caso a nova equação esteja com melhor índice de acertos procede-se novamente com a eliminação da variável menos significativa, se agregam novos casos da base de dados e é feita nova análise de regressão. Este processo se encerra quando não houver mais ganho no percentual de acerto, ou então que se alcance determinada quantidade mínima de variáveis explanatórias. Para esta metodologia a quantidade mínima de variáveis em cada equação de regressão será de quatro variáveis explanatórias, esta restrição visa evitar a obtenção de modelos simplistas.

Quando encerrar o processo de redução de variáveis teremos uma equação selecionada com melhor índice de acerto percentual. Caso duas ou mais equações possuam índices de

[5] No decorrer do texto será utilizado o termo COMBINAÇÃO para designar todas as combinações de variáveis preditoras possíveis.

acerto iguais deve-se selecionar o modelo com menor quantidade de variáveis. O próximo passo será retirar os casos já contemplados por esta equação da base de auditorias e o processo deve ser reiniciado com a verificação das combinações de variáveis preditoras.

Este ciclo se repete até que tenham sido processados pelo menos 85% dos casos existentes na base inicial. Quando o percentual de casos ultrapassar este valor o processamento será executado somente mais uma vez e sem a restrição da quantidade mínima de quatro variáveis[6].

Teremos então uma série de equações que deverão contemplar uma quantidade significativa dos casos a serem analisados. A prioridade para escolha das equações será exatamente igual à ordem em que elas forem selecionadas. O algoritmo a seguir apresenta as etapas a serem seguidas para obtenção do conjunto de equações:

---

1 – A partir da base de dados, gerar uma relação com todas as combinações de variáveis independentes e a quantidade de contribuintes que apresentaram valores a todas as variáveis da combinação, calcular também o fator com a quantidade de casos por variável independente.

2 – Dividir a relação do passo 1 em grupos com mesma quantidade de variáveis na combinação.

3 - Selecionar o grupo com maior número de variáveis e que possuam pelo menos uma combinação com fator igual ou superior a 15 casos por variável independente.

4 – No grupo selecionado, para cada combinação com fator igual ou superior a 15, construir uma base de dados com os casos que informaram valores para todas as variáveis da combinação;

5 – Utilizar análise de regressão logística para processar as bases de dados geradas no passo 4. Iniciando com todas as variáveis da combinação e empregar o procedimento stepwise backward para reduzir a quantidade de variáveis até se obter o modelo que melhor ajuste a variável dependente pelo percentual de acerto médio.

5.1 - Em cada iteração após retirar uma variável, agregar novos casos (registro de auditoria) que satisfaçam a condição de ter valor informado em todas as variáveis restantes no modelo;

---

[6] Valores empregados como restrições foram obtidos após inúmeras experiências com diversas quantidades de parâmetros.

5.2 – Encerrar o processo de redução quando o novo modelo possuir um valor percentual de acerto médio menor que o modelo anterior;

5.3 – No processo de redução a quantidade mínima de variáveis será de 4, exceto quando a quantidade de casos processados atingir 85% dos casos na base de dados inicial, nesta situação procede-se a mais uma redução de variáveis, e este último modelo pode ter a quantidade de variáveis inferior a quatro.

5.4 – Selecionar o modelo com melhor ajuste para a combinação. Caso exista mais de um modelo com mesmo valor percentual de acerto médio, selecionar o modelo com menor quantidade de variáveis;

6 – Após processar todas as combinações do grupo selecionado no passo 3, deve-se localizar qual combinação atingiu o maior valor percentual de acerto médio e selecionar a equação de regressão logística respectiva para integrar o conjunto de equações.

7 - Eliminar da base de dados todos os casos que se enquadram na equação selecionada no passo anterior. Com a nova base procede-se novamente a partir do passo 1, até que todos os casos tenham sido processados, ou se alcance os limites do item 5.3.

8 – Para o processamento mensal, a prioridade das equações será a seqüência em que as equações forem selecionadas neste algoritmo.

Algoritmo 1 – Elaboração do conjunto de equações de regressão logística

Para ilustrar esta atividade, na Figura 11 é apresentado o fluxograma respectivo.

Finalizando esta etapa, pode-se ver na Tabela 3 um esquema das atividades e respectivas saídas produzidas.

Tabela 3 - Entradas e saídas das atividades da etapa 3.4.3

Atividades	Saídas
Identificação de variáveis relevantes	Relação de variáveis selecionadas
Geração de modelos probabilísticos	Equações de regressão
Obtenção de um conjunto de equações	Conjunto de equações

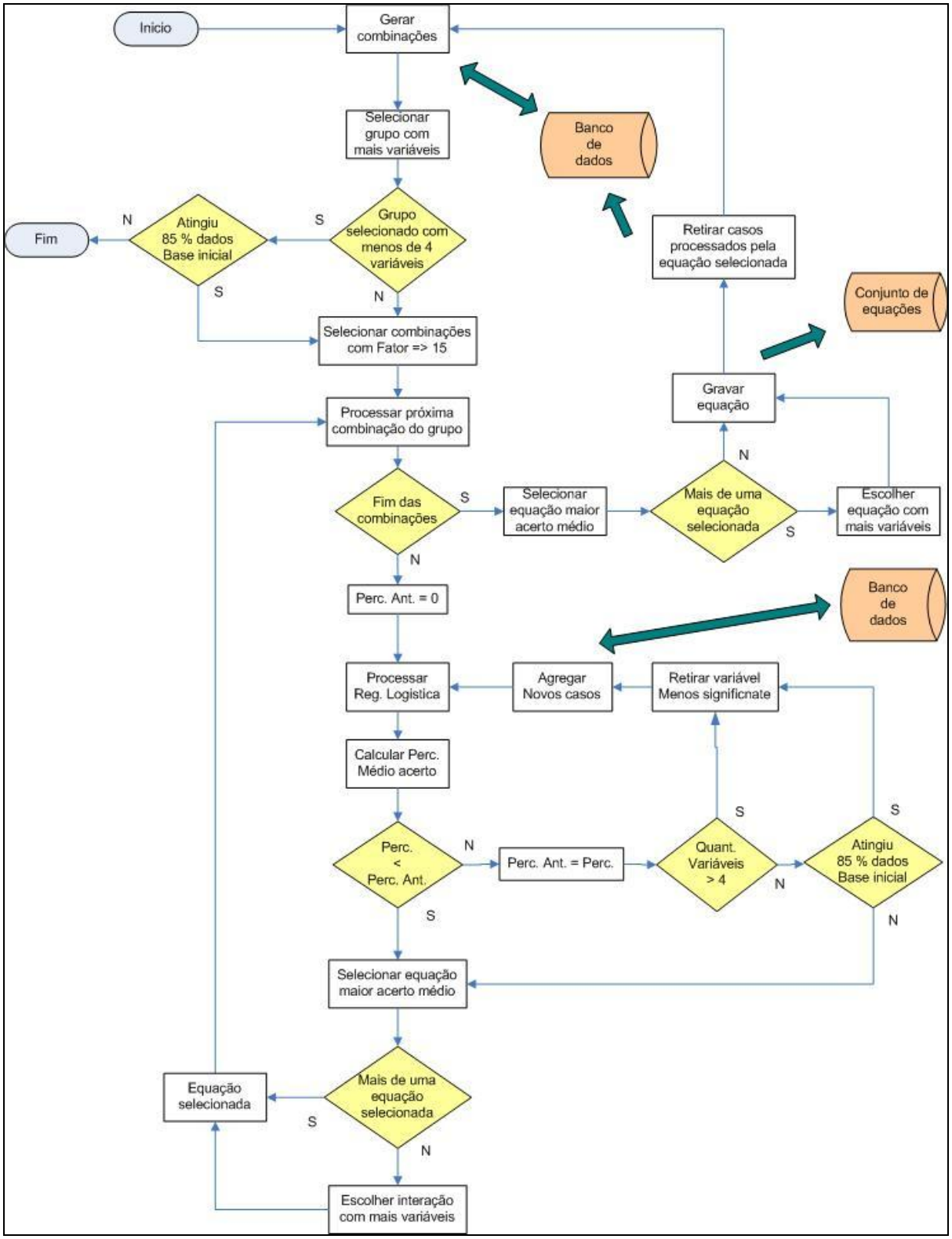

Figura 11 – Fluxograma para obtenção do conjunto de equações de regressão logística

3.4.4 - Classificação de contribuintes

Esta etapa será desenvolvida no decorrer dos exercícios mensais. Após o recebimento das declarações dos contribuintes e demais informações periódicas, procede-se a geração de relação com os maiores indícios de sonegação, com base nos modelos probabilísticos. Caso se utilize de um conjunto de equações o processamento de cada empresa deverá seguir uma ordem de prioridade que privilegie as equações que tenham sido obtidas em primeiro lugar. Se, ao processar uma empresa esta não possuir valores declarados para alguma das variáveis presentes na equação, deve-se passar à próxima equação na seqüência e verificar a possibilidade de se fazer o processamento. Segue-se nesta seqüência até que a empresa a ser processada seja encaixada em uma das equações ou então, no caso de nenhuma das equações poderem ser utilizada para processar a empresa, esta não terá valor de probabilidade calculada para o mês em questão.

Ao final do processamento mensal, teremos uma relação com as probabilidades de haver irregularidade nas informações mensais de cada empresa. Compõem esta etapa as seguintes atividades:

a) **Relacionar maiores indícios:** deve-se elaborar uma relação das empresas com maior probabilidade de sonegação. Para tanto os dados mensais serão processados com o modelo de regressão logística respectivo.

Na Tabela 4 é apresentado um esquema das atividades e respectivas saídas geradas nesta etapa final.

Tabela 4 - Entradas e saídas das atividades da etapa 3.4.4

Atividades	Saídas
Relação dos maiores indícios	Relação de empresas com maiores probabilidades de terem informado dados irregulares nas declarações mensais.

3.5 - RESUMO DO MODELO

Após a seleção dos dados a serem utilizados procede-se a limpeza dos mesmos e seleção das variáveis para geração da matriz de dados. Na seqüência, com as informações anuais de arrecadação, são gerados agrupamentos homogêneos para todas as empresas contribuintes. A partir do resultado das auditorias realizadas e com base nos dados históricos mensais disponíveis é elaborado um conjunto de equações de regressão logística. Em seguida, a cada mês, com os dados de arrecadação mensal é elaborada uma relação com as probabilidades de cada empresa estar com indícios de sonegação fiscal. Com o objetivo de oferecer um resumo seqüencial do modelo proposto, a Figura 12 representa o diagrama de todas as etapas e atividades para implementar o presente modelo.

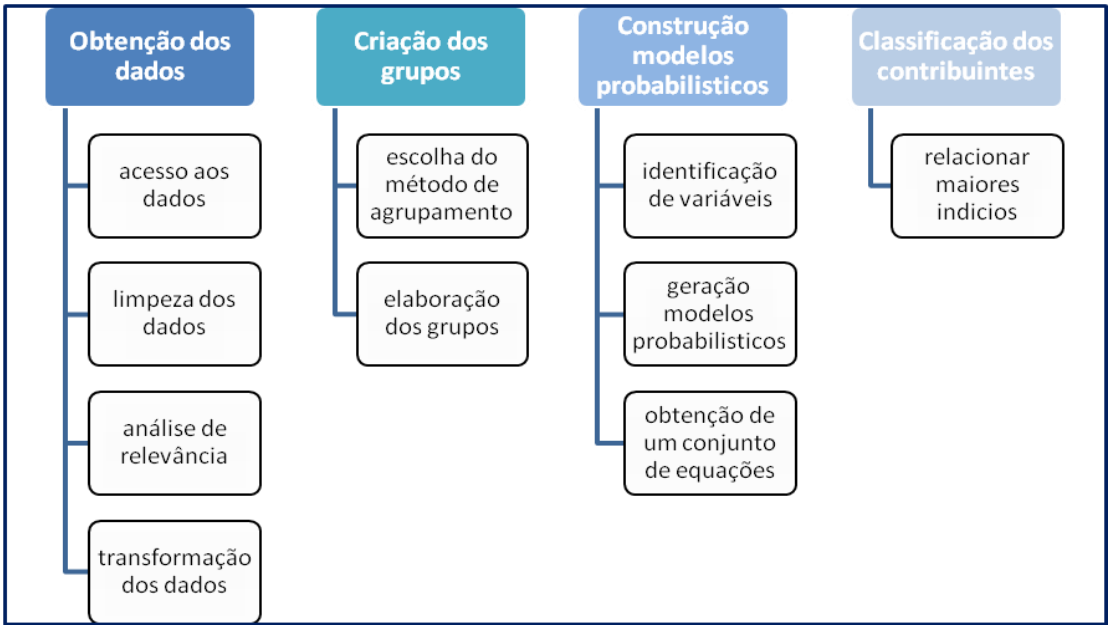

Figura 12 - Etapas e atividades do modelo proposto

3.6 – CONCLUSÕES DO CAPÍTULO

Utilizando-se de um conjunto de técnicas estatísticas e de conceitos de *data mining* bem difundidos, este modelo pretende se tornar um guia para classificar empresas contribuintes junto aos órgãos de fiscalização. A partir dos dados apresentados nas

declarações periódicas e com base nas probabilidades calculadas a classificação servirá de apoio para seleção dos contribuintes a serem fiscalizados.

Este modelo será aplicado em um estudo de caso, no próximo capítulo, para tanto serão analisados dados das empresas contribuintes do ICMS no Estado de Santa Catarina nos anos de 2005 a 2007.

4 - APLICAÇÃO DO MODELO – ESTUDO DE CASO

4.1 – INTRODUÇÃO

Este capítulo apresenta a aplicação do modelo de classificação proposto para validação e ajustes necessários em todo o processo. A pesquisa foi realizada na Secretaria de Estado da Fazenda de Santa Catarina (SEF-SC) entre agosto de 2007 e julho de 2009 junto à GEFIS – Gerência de Fiscalização, nesta aplicação foi analisado o comportamento dos contribuintes do ICMS .

O Imposto sobre Circulação de Mercadorias e Serviços (ICMS) é um imposto de competência dos Estados, incide sobre operações relativas à circulação de mercadorias e prestações de serviços, e também sobre serviços de transporte interestadual e intermunicipal e de comunicação. Conforme Biasoto Jr. *et al.* (1998), na maioria dos estados brasileiros o ICMS responde por mais de 80% da sua arrecadação. No apêndice A é apresentado detalhes da composição deste imposto, aspectos históricos e os procedimentos utilizados para seleção dos contribuintes a serem auditados.

4.2 - DADOS SEF-SC: VISÃO PRELIMINAR

Estão cadastradas no Estado de Santa Catarina 138.136 empresas, conforme dados de dezembro de 2006. Deste total, 35.855 são empresas de porte médio ou grande, que serão objetos de estudo desta aplicação prática do modelo de classificação. Atualmente o Estado conta com 443 fiscais em atividade para as funções de auditoria. Destas informações foram extraídos os dados de classificação conforme os setores econômicos apresentado na Tabela 5.

Tabela 5 - Quantidade de ocorrências agrupadas por setor econômico (dez. 2006)

SETOR	DESCRIÇÃO	QTD.
1	Agropecuária	499
3	Indústria	11.194
7	Comércio atacadista	4.460
8	Comércio varejista	15.172
9	Serviços	4.525

Para organizar e acompanhar as atividades de fiscalização em todo o Estado foram criadas as Unidades Setoriais de Fiscalização da Secretaria da Fazenda de Santa Catarina (USEFI's), de caráter geográfico. A Figura 13 mostra a distribuição das USEFI's no Estado.

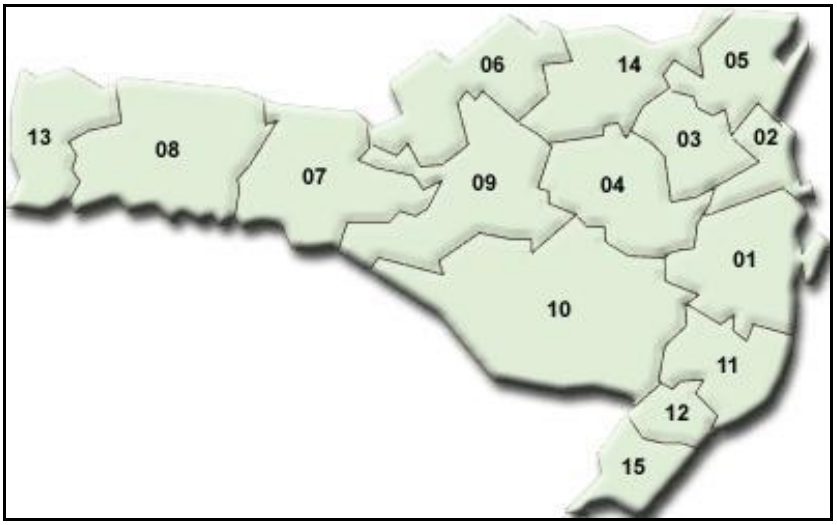

Figura 13 - Distribuição das Regiões Fiscais (USEFI´s) em Santa Catarina

A

Tabela **6** apresenta as 15 regiões fiscais do Estado com a quantidade de contribuintes e faturamento médio para cada região fiscal.

Tabela 6 - Quantidade de ocorrências agrupadas por regiões de fiscalização (USEFI - classificação SEF. Dez. 2006)

USEFI	DESCRIÇÃO	QTD.	FAT. MÉDIO (R$)
1	Florianópolis	4.711	7.125.085,16
2	Itajaí	3.839	11.166.122,08
3	Blumenau	5.063	5.178.180,36
4	Rio do Sul	1.841	3.491.603,76
5	Joinville	5.153	11.754.427,80
6	Porto União	1.058	9.297.969,46
7	Joaçaba	1.689	5.806.669,58
8	Chapecó	3.195	5.651.808,52
9	Curitibanos	1.387	5.144.795,08
10	Lages	1.266	4.786.249,19
11	Tubarão	1.678	4.938.435,18
12	Criciúma	2.031	4.986.490,83
13	São Miguel do Oeste	756	3.610.436,04
14	Mafra	1.416	4.315.118,65
15	Araranguá	770	3.526.773,26

As classificações mostradas acima, quando analisadas individualmente representam uma característica apenas dos dados, e quando são realizadas análises cruzadas entre as tabelas 5 e 6 apresentam uma quantidade alta de grupos (75). Esta constatação dificulta o trabalho de análise das informações e as empresas em cada grupo não possuem comportamento econômico homogêneo devido à grande variância dos valores apresentados por cada uma delas.

4.3 – OBTENÇÃO DOS DADOS

4.3.1 – Acesso aos Dados

Os dados foram coletados diretamente do SAT - Sistema de Arrecadação Tributária da Secretaria Estadual da Fazenda em Santa Catarina. Originalmente no banco de dados Oracle foram desenvolvidos dois programas para captura e montagem da matriz de dados. O processamento das análises descritivas e multivariadas deu-se com o uso do programa SPSS v. 13 para Windows.

Conforme a legislação vigente as empresas devem entregar fichas declaratórias de sua movimentação econômica (DIME) todos os meses com dados de movimentação financeira e anualmente uma declaração com dados mais abrangentes da organização da empresa, especialmente dados do balanço contábil. Para a elaboração das equações de probabilidade de ocorrência de fraude foi utilizada a tabela de AUDITORIAS realizadas pela Gerencia de Fiscalização nos anos de 2005 até 2007. Nesta tabela existe o registro das auditorias feitas nas empresas e o resultado encontrado pode ser a ausência de notificações fazendárias, situação na qual a empresa não viola a legislação vigente, ou no caso de infração é registrado o código da notificação correspondente bem como o valor da autuação.

4.3.2 – Limpeza dos Dados

Neste caso específico os dados encontram-se em base de dados computacionais de sistemas em ambiente estável o que não apresentou problemas na consistência das informações.

4.3.3 – Análise de Relevância

DIME ANUAL - Foram selecionadas 22 variáveis dentre as 253 informações da DIME anual. Esta seleção levou em conta a importância dos dados, quantidade de contribuintes que responderam a cada variável, já que os contribuintes não são obrigados a preencher todos os campos. Um resumo destas 22 variáveis e suas características estatísticas encontra-se na Tabela 7.

AUDITORIA – Os dados das empresas auditadas nos anos 2005 a 2007 foram selecionados para compor a base que será utilizada para elaborar as equações de regressão logística. A base é composta de 7.557 auditorias, onde consta além da identificação da empresa, o período que ocorreu a auditoria, quais os fiscais envolvidos, quanto tempo durou cada auditoria e também o resultado desta atividade. Caso existam irregularidades na documentação contábil e fiscal da empresa, estas devem ser apontadas no relatório da fiscalização e nesta base de dados temos as infrações respectivas.

DIME MENSAL - Foram selecionadas 14 variáveis dentre as 207 informações da DIME mensal. Esta seleção partiu da base de dados de AUDITORA, para todas as 7.557 auditorias do período de análise foram acrescentados os dados da DIME mensal respectiva a cada empresa auditada. Para haver sincronia entre as datas (períodos de informação) utilizou-se os dados da DIME do mês anterior ao período que aconteceu a auditoria. Desta forma para uma auditoria ocorrida em abril de 2006 na empresa X se buscou os dados da DIME mensal do período de março de 2006 da referida empresa.

Na seleção das variáveis representativas da DIME mensal, levou-se em conta a importância dos dados e a quantidade de contribuintes que responderam a cada variável, já que os contribuintes não são obrigados a preencher todos os campos. Na Tabela 8 temos as 14 variáveis escolhidas e suas características estatísticas.

Tabela 7 - Variáveis selecionadas para análise de cluster (DIME Anual)

CL.	VARIÁVEL	Obs.	Média	Desvio Padrão	Valor Mínimo	Valor Máximo
3060	Saídas - valor contábil	35.855	6.987.064,35	67.022.334,41	240.003,20	4.610.260.202,19
3010	Entradas - valor contábil	27.737	8.266.960,83	155.143.751,11	240.005,09	23.642.605.893,10
3050	Entradas - outras operações sem crédito de imposto	20.529	4.778.026,17	30.335.372,57	240.005,09	1.449.518.388,76
3100	Saídas - outras operações sem débito de imposto	20.119	4.081.036,43	25.075.907,68	240.003,20	1.410.089.150,82
6030	Simples - total receita tributável - Emp. Regime simples	12.872	610.862,48	422.788,13	240.003,20	7.596.140,76
6010	Simples - receita tributável do estabelecimento	12.689	595.703,32	412.114,77	240.003,20	7.596.140,76
3020	Entradas - base de cálculo	9.187	10.968.782,17	299.690.251,24	240.026,67	23.080.104.415,57
3090	Saídas - operações isentas ou não tributadas	7.974	7.932.850,84	72.857.925,18	240.049,58	3.967.123.172,41
9080	Saldo - total créditos	6.241	4.019.224,06	29.015.480,83	240.060,20	1.376.846.998,93
9050	Saldo - subtotal créditos	6.215	3.904.970,34	28.721.328,22	240.060,20	1.376.846.998,93
80030	Inventario - receita bruta de vendas e serviços - INF ANUAL	23.214	7.836.755,47	163.508.581,94	240.010,98	17.043.335.000,00
83310	DRE - receita bruta vendas/serviços - INF ANUAL	13.624	15.552.887,16	249.137.150,30	240.010,98	17.043.335.000,00
83320	DRE - receita liquida vendas/serviços - INF ANUAL	13.309	13.214.221,97	203.794.010,09	240.040,18	13.075.080.000,00
82299	Passivo - Total geral do passivo - INF ANUAL	10.530	14.906.267,89	252.587.801,38	240.035,68	15.590.751.000,00
81199	Ativo - total geral do ativo - INF ANUAL	10.530	14.906.267,89	252.587.801,38	240.035,68	15.590.751.000,00
84499	Detalhamento das despesas Total - INF ANUAL	11.928	4.666.885,77	48.980.834,79	240.036,93	2.915.720.731,12
83323	DRE - custo da mercadoria ou prod. vendidos - INF ANUAL	10.034	12.905.494,82	186.166.598,76	240.028,55	10.361.716.000,00
81110	Ativo - circulante - INF ANUAL	8.593	7.830.276,73	98.586.171,10	240.024,22	5.817.885.000,00
80020	Inventario - estoque no fim do exercício - INF ANUAL	7.478	1.925.440,98	9.843.504,70	240.034,49	510.366.451,00
82270	Passivo - patrimônio liquido - INF ANUAL	6.987	9.769.003,21	122.730.030,42	240.102,01	5.504.581.000,00
83330	DRE - lucro bruto - INF ANUAL	7.476	6.154.304,10	69.451.920,49	240.027,00	3.802.248.000,00
80010	Inventario - estoque inicio exercício - INF ANUAL	6.410	2.279.050,10	12.896.120,85	240.000,32	438.058.419,00

Tabela 8 - Variáveis selecionadas para equações logísticas (AUDITORIA e DIME Mensal)

CL.	VARIÁVEL	Obs.	Média	Desvio Padrão	Min.	Max.
3060	Resumo - Saídas - valor contábil	7.303	1.619.635,98	11.736.553,32	0,01	420.445.089,20
3010	Resumo - Entradas - valor contábil	7.002	1.398.269,84	10.097.646,70	0,01	500.929.447,66
9040	Saldo - Total de débitos	6.925	147.107,88	1.750.375,69	0,12	78.274.538,99
3050	Resumo - Entradas - outras operações sem crédito de imposto	6.133	676.966,72	3.833.646,29	0,01	146.936.359,18
9999	Saldo - imposto a recolher	5.014	62.047,64	1.264.002,43	0,26	67.436.796,68
3100	Resumo - Saídas - outras operações sem débito de imposto	5.418	561.740,27	3.211.315,26	0,01	129.267.196,26
3070	Resumo - Saídas - base de cálculo	5.665	980.458,07	7.908.728,49	1,70	354.524.611,01
3080	Resumo - Saídas - imposto debitado	5.574	146.582,43	1.644.910,29	0,01	78.274.538,99
9080	Saldo - total créditos	5.036	263.493,42	3.113.054,47	1,01	143.521.603,65
3030	Resumo - Entradas - imposto creditado	4.339	135.287,03	1.249.626,43	0.06	44.829.957,18
3020	Resumo - Entradas - base de cálculo	4.326	730.343,83	3.965.145,75	0.01	170.166.962,92
3090	Resumo - Saídas - operações isentas ou não tributadas	2.966	1.054.104,28	11.526.391,57	0,01	410.322.720,28
3040	Resumo - Entradas - operações isentas ou não tributadas	3.062	625.518,29	7.023.343,48	0.01	241.738.995,48
9998	Saldo - saldo credor para mês seguinte	2.072	348.963,95	4.136.617,00	0,01	135.101.222,93

4.3.4 – Transformação dos Dados

Conforme visto na seção anterior, diversas questões não são preenchidas pelos contribuintes, portanto faz-se necessário uma análise dos dados faltantes. Para a análise de agrupamento, a partir da base DIME anual, optou-se pelo preenchimento destes dados com valores médios ponderados para cada atividade econômica dentro da respectiva unidade fiscal (região) que o contribuinte pertence. Inicialmente foi criada uma tabela com os valores médios para cada variável, agrupadas dentro do setor e da região (USEFI); como todas as empresas informaram o valor da variável 3060 – Saídas, valor contábil, no processo de preenchimento se utilizou a seguinte expressão:

$$Z\_v = (z\_3060 * M\_v) / m\_3060 \qquad (16)$$

Onde:

Z_v – valor faltante de uma empresa para uma determinada variável v

z_3060 – valor da variável 3060 para a respectiva empresa

M_v – valor médio dentro do setor/região da empresa para a variável v

m_3060 – valor médio dentro do setor/região da empresa para a variável 3060

Empregou-se, também, a transformação dos valores monetários destas variáveis devido à grande disparidade entre os valores apresentados pelos contribuintes que abrangia desde 240 mil reais até cifras acima de 20 bilhões de reais; para tanto as variáveis foram transformados para uma escala com valores entre 0 e 1. Conforme Bussab *et al.* (1990), para variáveis quantitativas pode-se usar a transformação:

$$Z = \frac{z - z\min}{z\max - z\min} \qquad (17)$$

Onde:

Z - número transformado no intervalo 0,1

z – número original

zmin – menor valor original na série de dados

zmax – maior valor original na série de dados

Para esta transformação foi desenvolvido um programa específico, PGM-01 em linguagem *MS Visual Basic* que facilitou o processamento das informações. No apêndice F é transcrita a rotina principal do programa.

Na análise das informações mensais dos contribuintes, também ficou evidente a grande variabilidade nos valores apontados. Neste caso foram testados dois procedimentos de transformação nos valores. A Tabela 9 e a Figura 14 representam o resultado da análise para a variável v3060, sendo v3060_Z os dados após transformação padronizada e v3060_L10 após transformação logarítmica com base 10. A análise das demais variáveis apresentou comportamento semelhante.

Tabela 9 – Análise descritiva variável v3060

		v3060	**v3060_Z**	**v3060_L10**
N	Validos	5816	5816	5816
	Faltantes	0	0	0
Média		1706840,453026	0,00000	5,230316
Mediana		126073,150000	-0,12362	5,100623
Desvio Padrão		12787355,465294	1,00000	0,808311
Assimetria		22,443130	22,44313	0,263568
Erro Padrão de Assimetria		0,032111	0,03211	0,032111
Curtose		622,212119	622,21212	2,993329
Erro Padrão de Curtose		0,064211	0,06421	0,064211

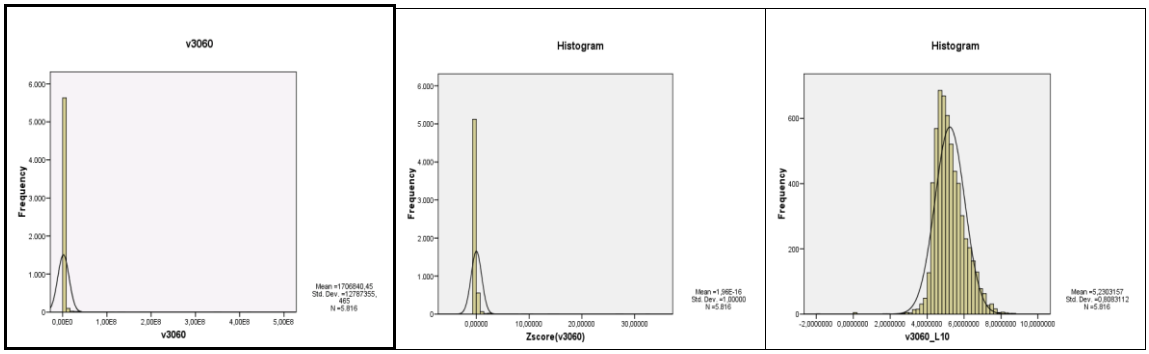

Figura 14 - Gráfico de distribuição de freqüência das variáveis v3060, v3060_T e v3060_L10

Como resultado da análise preliminar das variáveis e conforme apontam os resultados estatísticos e os gráficos de distribuição de freqüência optou-se pelo emprego de logaritmos com base 10 dos valores para esta análise.

4.4 – CRIAÇÃO DE GRUPOS

4.4.1 – Escolha do Método de Agrupamento

Devido à grande quantidade de dados foi escolhida a técnica de análise de cluster *two-step* que além de manipular grandes volumes de dados, também processa tanto variáveis contínuas quanto categóricas.

4.4.2 – Elaboração dos Grupos

Foram analisadas formações de grupos no intervalo entre 3 e 60, para se estabelecer a quantidade de grupos, partiu-se de o critério AIC (Akaike) e da Razão das Distâncias Medidas (*Ratio of Distance Measures*), ambos apresentados pelo SPSS. Podem-se ver na Tabela 10 as opções com melhores resultados para os critérios indicados.

Tabela 10 – Critérios para escolha da quantidade (N) de grupos

N	**AIC**	***Razão das Distâncias Medidas*(c)**
6	-6837838,85	1,458
24	-6957124,82	1,176
29	-6970995,37	1,140

Após diversas análises com especialistas da Secretaria da Fazenda, a quantidade de agrupamentos que apresentou interesse ao estudo foram 24 clusters. Na Tabela 11 apresentam-se as quantidades de casos em cada cluster, bem como características gerais de cada grupo. Deve-se salientar a formação de grupos com empresas que possuem faturamento muito alto ou então com faturamento com valores altos independente da região ou setor econômico. Outra associação forte existe entre empresas da indústria de transformação com empresas de construção civil. Dentro do setor de comércio atacadista as que envolvem atividades ligadas ao setor automobilístico também formaram grupos independentes, merecendo atenção à parte.

Tabela 11 – Quantidade de ocorrências em cada *cluster* (CL.)

CL.	CARACTERÍSTICAS	QTD.	FAT. MÉDIO
1	Faturamento muito alto	45	1.297.109.844,82
2	Faturamento alto	134	255.323.853,23
3	Faturamento alto com emp. Financeiras	274	137.084.454,21
4	Faturamento médio	434	62.113.964,87
5	Com. varejo; usefi 1	2.450	1.889.808,19
6	Com. varejo; usefi 7, 10	1.355	1.746.070,70
7	Com. varejo; usefi 5	1.939	1.964.388,07
8	Serviços; usefi 3, 5, 10	1.101	1.843.961,38
9	Indústria e varejo; usefi 13	632	1.959.418,27
10	Indústria e construção; usefi 4, 7, 10	1.271	2.755.081,87
11	Serviços; usefi 2, 6, 9, 11,12,13, 14	1.569	1.719.594,31
12	Serviços; usefi 1	1.239	2.808.656,60
13	Serviços; usefi 7, 8	913	4.237.391,24
14	Com. atacado; usefi 6, 9	1.084	1.726.843,23
15	Com. varejo; usefi 3	1.531	1.738.476,09
16	Com. varejo; usefi 11, 14, 15	1.632	3.337.450,33
17	Indústria; usefi 6, 14, 15	999	1.749.831,93
18	Com. varejo; usefi 4, 12	1.510	1.840.275,60
19	Com. varejo; usefi 6, 9	1.169	3.290.150,36
20	Com. varejo; usefi 2	1.603	3.615.863,35
21	Agropecuária	950	2.565.770,28
22	Indústria e construção; usefi 2	1.084	3.146.989,66
23	Indústria e construção; usefi 8, 12	1.473	4.501.973,51
24	Com. varejo e serviço; usefi 6,12, 15	726	1.709.026,30

A **Erro! Auto-referência de indicador não válida.** apresenta a composição dos 24 grupos relativa à participação dos setores econômicos, à exceção dos 4 primeiros grupos onde o valor de faturamento foi preponderante pode-se observar a formação de grupos específicos para empresas do setor industrial e também no comércio varejista. Para os demais setores a análise multivariada sugere a junção com empresas de outros setores.

Tabela 12 – Quantidade de ocorrências em cada *cluster* (CL.) para cada setor de atividade econômica

CL.	Agropec.	Indústria	Atacado	Varejo	Serviços
1		18	15	2	10
2		37	33	21	46
3		136	75	33	31
4	4	224	39	176	25
5				2450	
6				1355	
7				1939	
8					2008
9		165	86	1547	
10		1271			
11					1571
12			599		640
13			2249		
14				1531	
15				1632	
16		2066			192
17				1510	
18				1603	
19	116		1364		
20		1473			
21		1639			
22		2413			
23		1752			
24				1373	

A Tabela 13 apresenta a composição dos 24 grupos relativa à participação das regiões geográficas (USEFI's), em contraste com a principal divisão de empresas que existe hoje na Secretaria de Estado da Fazenda que inicialmente separa os contribuintes em suas regiões. A análise de agrupamento indica que pode ser feita a junção de empresas de diferentes regiões visando a melhor compreensão do comportamento dos contribuintes.

Tabela 13 – Quantidade de ocorrências em cada cluster (CL.) para cada região geográfica (USEFI)

CL.	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
1	9	9	1		14	1	2	1		3	5				
2	29	19	21	1	22	1	5	11	2	8	8	5	1	2	2
3	48	54	30	9	46	8	8	18	4	1	17	14	3	9	6
4	53	73	78	23	67	7	16	33	13	10	21	34	9	26	5
5	2450														
6							712			643					
7					1939										
8			419		676		456	457							
9						558			611				629		
10				712			297			262					
11		410				105			243	154	144	218	103	194	
12	1239														
13		568	547		600	103	153		145					133	

CL.	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
14			1531												
15											799			512	321
16		1074				233								508	251
17				785								725			
18		1603													
19				294				527		185	199	296			171
20								734				739			
21	856								298		485				
22			2413												
23					1752										
24								1373							

4.5 – CONSTRUÇÃO DOS MODELOS PROBABILÍSTICOS

Nesta etapa serão gerados modelos estatísticos para uma empresa que tenha apresentado declarações com e sem indícios de sonegação. Foram utilizados dados das auditorias realizadas entre janeiro de 2005 a junho de 2007; a estes dados foram incorporadas informações das declarações mensais (DIME mensal) do mês anterior ao da ocorrência da auditoria na empresa. Estes dados foram então separados em duas partes, sendo a primeira com auditorias dos anos de 2005 e 2006 utilizadas para a construção das equações de regressão logística, e a segunda parte com o primeiro semestre de 2007 que foi utilizado para validação do modelo criado anteriormente. A Figura 15 ilustra este procedimento.

Conforme exposto na seção 4.3.3 foram selecionadas 14 variáveis dentre as informações da DIME mensal para esta etapa. Na tabela 14 pode-se observar a composição das duas bases de auditoria, sendo AUDIT-1 referente às auditorias feitas nos dois primeiros anos desta pesquisa: 2005 e 2006 e na base AUDIT-2 as auditorias do primeiro semestre de 2007. Nesta tabela fica evidenciada a quantidade de empresas que preencheram as informações de cada variável. Foi adicionado a estes dados a codificação do agrupamento respectivo de cada empresa conforme definido na etapa anterior.

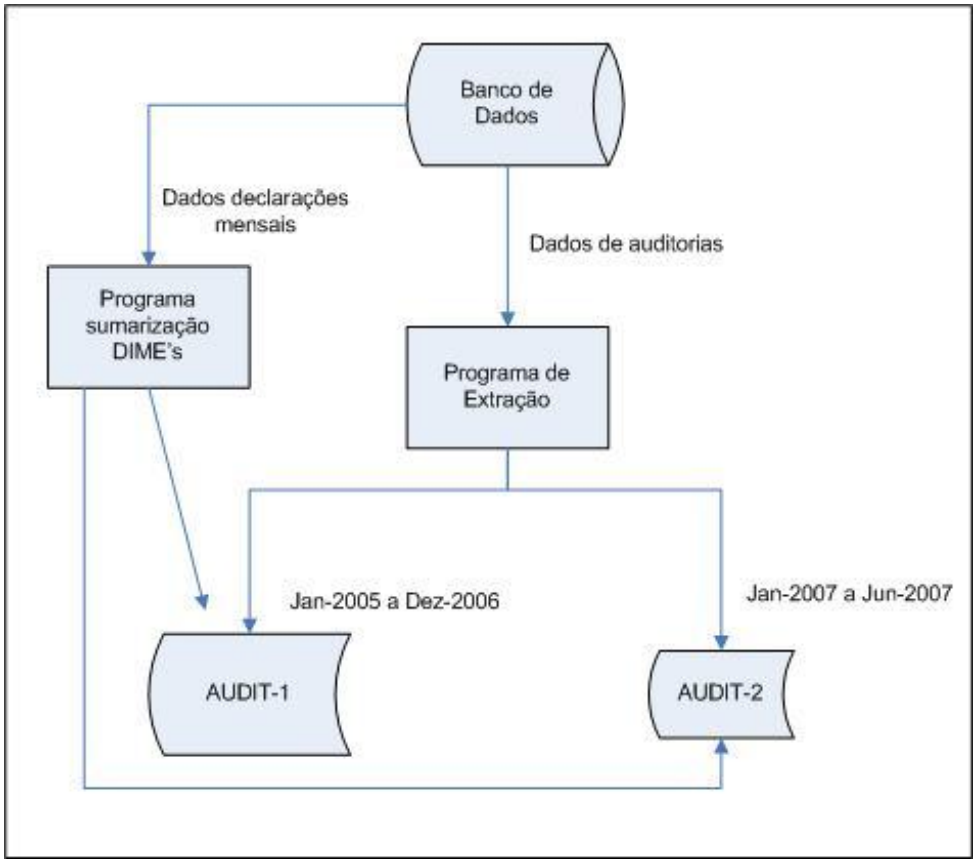

Figura 15 - Diagrama de extração dos dados para aplicar regressão logística

4.5.1 – Identificação das Variáveis

A DIME mensal em Santa Catarina é um formulário com 207 questões que as empresas apresentam à fiscalização. Da mesma forma que ocorreu nos dados da declaração anual nem todos os campos são preenchidos por todas as empresas. Após a análise das questões que tiveram uma quantidade maior de respostas o número de variáveis foi reduzido a 40. Como diversos campos representam operações aritméticas simples entre outros campos uma segunda redução permitiu a utilização de apenas 14 variáveis. Com a incorporação dos dados das auditorias e também a identificação do agrupamento específico de cada empresa, foram criadas as bases AUDIT-1 e AUDIT-2.

Na Tabela 14 são apresentadas as variáveis a serem utilizadas para elaboração dos modelos probabilísticos e a quantidade de ocorrências em cada base de dados.

Tabela 14– Variáveis selecionadas para elaborar modelos probabilísticos e suas freqüências

			AUTID-1		AUTID-2	
			Preenchida	*sem valor*	*Preenchida*	*sem valor*
1	3060	Resumo - Saídas - valor contábil	5.839	193	1.464	61
2	3010	Resumo - Entradas - valor contábil	5.590	442	1.412	113
3	9040	Saldo - Total de débitos	5.508	524	1.417	108
4	3050	Resumo - Entradas - outras operações sem crédito de imposto	4.861	1.171	1.272	253
5	9999	Saldo - imposto a recolher	3.951	2.081	1.063	462
6	3100	Resumo - Saídas - outras operações sem débito de imposto	4.354	1.678	1.064	461
7	3070	Resumo - Saídas - base de calculo	4.496	1.536	1.169	356
8	3080	Resumo - Saídas - imposto debitado	4.428	1.604	1.146	379
9	9080	Saldo - total créditos	4.076	1.956	960	565
10	3030	Resumo - Entradas - imposto creditado	3.493	2.539	846	679
11	3020	Resumo - Entradas - base de cálculo	3.489	2.543	837	688
12	3090	Resumo - Saídas - operações isentas ou não tributadas	2.444	3.588	522	1.003
13	3040	Resumo - Entradas - operações isentas ou não tributadas	2.494	3.538	568	957
14	9998	Saldo - saldo credor para mês seguinte	1.691	4.341	381	1.144

4.5.2 – Combinação de equações e geração dos modelos estatísticos

Para esta etapa foram desenvolvidos programas de apoio na linguagem MS *Visual Basic*. A seguir, descrição sucinta dos mesmos:

PGM-02 - para gerar as combinações de variáveis e quantidade de empresas que informaram valores para todas as variáveis de cada combinação. A relação é classificada pela maior quantidade de variáveis.

Conforme o algoritmo 1 apresentado no capítulo anterior, a partir da base de aprendizado, AUDIT-1 que tem 6.032 casos foram executadas diversas vezes os passos 1 a 7 do referido algoritmo. Na primeira execução, com o auxilio do programa *PGM-02* foi gerada a relação com as combinações das 14 variáveis.

Na Tabela 15 é exibida parcialmente esta relação, destacando-se as combinações que terão interesse imediato. Ao todo foram geradas 16.383 combinações

Tabela 15 – Combinações das 14 variáveis com quantidade de casos respondidos (1ª. execução)

N	Combinação	Qt. Casos	Qt. Variável	Fator
1	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	14	0
2	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13	556	13	15,444
3	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 14	0	13	0
5	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 13 - 14	0	13	0
9	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 12 - 13 - 14	0	13	0
17	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 11 - 12 - 13 - 14	0	13	0
33	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 10 - 11 - 12 - 13 - 14	0	13	0
65	01 - 02 - 03 - 04 - 05 - 06 - 07 - 09 - 10 - 11 - 12 - 13 - 14	0	13	0
129	01 - 02 - 03 - 04 - 05 - 06 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	13	0
257	01 - 02 - 03 - 04 - 05 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	13	0
513	01 - 02 - 03 - 04 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	338	13	9,389
	...	...	...	...
1.025	01 - 02 - 03 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	13	0
	...	...	...	...
2.049	01 - 02 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	13	0
	...	...	...	...
4.097	01 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	13	0
	...	...	...	...
8.193	02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	13	0
	. . .	. . .	. . .	. . .
16.383	14	1691	1	70,458

Pode-se observar na primeira coluna a numeração seqüencial das combinações geradas pelo programa PGM-02, na segunda coluna a composição de cada combinação de variáveis sendo que a codificação 01 representa a primeira variável: 3060-Resumo-saídas-valor contábil, o código 02 a segunda variável: 3090-Resumo-saídas-valor contábil e assim por diante até a variável 14. Na terceira coluna tem-se a quantidade de casos que informaram valores em todas as variáveis da combinação, a próxima coluna mostra a quantidade de variáveis na combinação e a ultima coluna exibe o fator resultante da operação:

$$Fator = \frac{Qt\_casos}{(Qt\_\text{var} + Qt\_cluster - 1)} \qquad (18)$$

Onde,

Fator – quantidade de casos por variável independente

Qt_casos – quantidade de casos numa combinação de variáveis em particular

Qt_var – quantidade de variáveis na combinação

Qt_cluster – quantidade de grupos que serão incorporados ao modelo como variáveis dummy.

Conforme o algoritmo 1, no passo 2 deve-se separar as combinações em grupos com a mesma quantidade de variáveis. Tal procedimento é apresentado na Tabela 16, onde são informados para cada grupo as quantidades de combinações possíveis. Observa-se na segunda coluna o total de combinações; na terceira coluna o total de combinações que possuem valores informados para todas as variáveis da combinação; e, na última coluna o total de combinações que informaram valores em todas as variáveis da combinação e possuam quantidade de casos suficientes para se obter o fator igual ou superior a 15 casos por variável dependente.

Tabela 16 – Composição dos grupos de variáveis na 1ª. Execução

Grupo	Qt. combinações	Qt. Comb. Casos > 0	Qt. Comb. Fator => 15
14 variáveis	1	0	0
13 variáveis	14	2	1
12 variáveis	91	-	-
11 variáveis	364	-	-
10 variáveis	1.001	-	-
09 variáveis	2.002	-	-
08 variáveis	3.003	-	-
07 variáveis	3.432	-	-
06 variáveis	3.003	-	-
05 variáveis	2.002	-	-
04 variáveis	1.001	-	-
03 variáveis	364	-	-
02 variáveis	91	-	-
01 variável	14	-	-

Na Tabela 16 pode-se verificar que existe apenas uma combinação com 14 variáveis e nenhuma empresa apresentou valores para todas estas variáveis. Na seqüência é analisado o conjunto de combinações com 13 variáveis onde existem 14 combinações possíveis, mas apenas duas delas com quantidade de casos maior que zero, e somente uma combinação com fator igual ou superior a 15 casos por variável independente. De acordo com o passo 3 do algoritmo 1 deve-se selecionar este grupo para a seqüência do processamento.

Conforme especificado no passo 4 do algoritmo, deve-se gerar uma base de dados para cada combinação de variáveis do grupo selecionado. Neste caso, apenas uma

combinação (N=2 da Tabela 15) participa do grupo selecionado e seus 556 casos foram armazenados na base.

No passo 5 do algoritmo indica que os dados devem ser processados por meio de análise de regressão logística; como variável dependente utilizou-se o resultado das auditorias que na solução informatizada foi nomeada de *op_Notif,* esta variável assume o valor 0 quando, ao término da auditoria, não houve notificação de infração e, o valor 1 quando a auditoria resultou numa notificação. Também a variável *cl_24* foi utilizada para indicar de qual agrupamento pertence a empresa que foi auditada. Esta informação do agrupamento participa do modelo como uma variável categórica e para tanto é representada por uma seqüência de 23 variáveis *dummy*, onde o valor 1 indica que a empresa pertence ao grupo em questão e 0 ela não participa deste grupo.

Com auxilio do SPSS através do módulo de regressão logística foram analisados os 556 casos iniciais, e apresentaram um grau de acerto na ordem de 74,6% conforme quadro 1.

Quadro 1 – Classificação (a) para a base AUDIT-1 (1ª. Iteração)

		Previsto		
		op_Notif		%
Observado		0	1	Corretos
op_Notif	0	38	121	23,9
	1	20	377	95,0
% média				74,6

(a) Valor de corte é 0,50

Na avaliação do grau de acerto do modelo foi utilizada a tabela de classificação gerada pelo SPSS que consta no Quadro 1 onde mostra que das 556 auditorias realizadas, 159 não apresentaram notificação e 397 tiveram notificação. Com um valor de corte igual a 0,50 o modelo ajustou corretamente 38 das 159 observações não notificadas, perfazendo um acerto de 23,9%. No caso das auditorias que tiveram notificações o modelo apresentou um acerto de 95,0%, ajustando corretamente 377 das 397 observações notificadas. Para esta etapa do processo de análise de regressão o acerto médio foi de 74,6%.

No relatório de análise do SPSS, a partir dos valores da estatística de Wald, sugere-se que a variável 06: 3100-Resumo - Saídas - outras operações sem débito de imposto, teve a menor significância estatística dentre as 13 variáveis presentes no momento. Conforme o passo 5.1 do algoritmo 1, a variável 06 foi retirada para ser feita nova análise de regressão e verificar se nesta nova formação com 12 variáveis o grau de acerto do modelo é melhor que o

valor obtido na primeira iteração (74,6%). Nesta nova formação existe um acréscimo na quantidade de casos, já que mais empresas apresentaram valores em todas as 12 variáveis que serão analisadas, as novas informações foram adicionadas passando de 556 para 655 casos. O Quadro 2 apresenta o grau de acerto desta segunda iteração.

Quadro 2 – Classificação (a) para a base AUDIT-1 (2ª. Iteração)

		Previsto		
		op_Notif		%
Observado		0	1	Corretos
op_Notif	0	38	138	21,1
	1	23	456	95,2
% média				75,4

(a) Valor de corte é 0,50

Como houve uma melhoria no grau de acerto médio, uma nova iteração foi executada, neste caso, eliminou-se a variável 08: 3080-Resumo - Saídas - imposto debitado. Desta vez não ocorreu acréscimo na quantidade de casos e o acerto médio foi o mesmo da segunda iteração. Da mesma forma, o procedimento foi repetido até que não se obteve ganho no grau de acerto, o que aconteceu na quarta iteração. A Tabela 17 mostra as iterações realizadas com a composição das combinações de variáveis e grau de acerto obtido em cada iteração.

Tabela 17 – Seqüência de reduções de variáveis para estimar modelo logístico na 1ª. Execução.

Variáveis na combinação	Casos	% acerto	Remover
01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13	556	74,6	06
01 - 02 - 03 - 04 - 05 - 07 - 08 - 09 - 10 - 11 - 12 - 13	655	75,4	08
01 - 02 - 03 - 04 - 05 - 07 - 09 - 10 - 11 - 12 - 13	655	75,4	13
01 - 02 - 03 - 04 - 05 - 07 - 09 - 10 - 11 - 12	820	76,4	01
02 - 03 - 04 - 05 - 07 - 09 - 10 - 11 - 12	1.478	72,1	-

A partir da quinta iteração o grau de acerto decaiu em relação à iteração anterior, seguindo o passo 5.2 do algoritmo 1, a equação resultante da quarta iteração foi selecionada. Como existia apenas uma combinação no grupo de 13 variáveis, conforme o passo 6 do algoritmo, a equação selecionada passou a compor o conjunto de equações.

De acordo com o passo 7 do algoritmo, foi criada uma nova base de dados de auditorias retirando-se os 820 casos que participaram da 1ª execução. A nova base de dados passou a ter 5.212 casos e foi reiniciado o processo no passo 1 do algoritmo com a geração das combinações de variáveis por intermédio do programa *PGM-02.*

Tabela 18 – Combinações das 14 variáveis com quantidade de casos respondidos (2ª. execução)

N	Combinação	Qt. Casos	Qt. Variável	Fator
1	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	14	0
2	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13	0	13	0
	...	...	...	...
8	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11	673	11	20,029
	...	...	...	...
516	01 - 02 - 03 - 04 - 06 - 07 - 08 - 09 - 10 - 11 - 12	548	11	16,117
	...	...	...	...
518	01 - 02 - 03 - 04 - 06 - 07 - 08 - 09 - 10 - 11 – 13	770	11	22,647
519	01 - 02 - 03 - 04 - 06 - 07 - 08 - 09 - 10 - 11 – 14	921	11	27,088
	...	...	...	...
771	01 - 02 - 03 - 04 - 07 - 08 - 09 - 10 - 11 - 12 – 14	552	11	16,235
	...	...	...	...
773	01 - 02 - 03 - 04 - 07 - 08 - 09 - 10 - 11 - 13 – 14	564	11	16,588
	...	...	...	...
1.793	01 - 02 - 03 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	510	11	15,000
	. . .	. . .	. . .	. . .
16.383	14	1691	1	70,458

Conforme apresentado na Tabela 18, nenhuma combinação com 14, 13 ou 12 variáveis apresentou quantidade de casos com fator igual ou superior a 15, Com 11 variáveis existem 364 combinações possíveis, sendo que em 7 delas a quantidade de casos foi suficiente para se proceder a análise de regressão. Nesta situação deve-se conduzir a análise com redução de variáveis para cada uma das 7 combinações separando-se em cada análise a equação resultante da iteração que obteve o melhor grau de acerto médio. A Tabela 19 apresenta um resumo com as equações selecionadas para cada combinação processada na segunda execução.

Tabela 19 – Equações selecionadas em cada combinação de variáveis na 2ª. Execução.

N	Variáveis na equação	Qt. Iterações	Casos	% acerto
8	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11	1	673	80,1
516	01 - 02 - 03 - 04 - 06 - 07 - 08 - 09 - 10 - 11 - 12	1	548	73,1
518	01 - 02 - 03 - 05 - 07 - 09 - 10 - 11 - 12 - 13	3	837	73,0
519	01 - 02 - 04 - 06 - 07 - 08 - 09 - 10 - 11 – 14	3	922	73,9
771	01 - 04 - 10 - 11 - 14	7	1.145	72,3
773	01 - 02 - 03 - 04 - 07 - 08 - 09 - 10 - 11 - 13 – 14	1	564	68,5
1.793	01 - 02 - 03 - 08 - 10 - 11 - 12 - 13 - 14	3	511	68,0

Na Tabela 19 pode-se observar o seqüencial (N) da combinação que foi processada. Na segunda coluna, quais variáveis permaneceram após o processo de redução em cada combinação. A terceira coluna indica quantas iterações foram necessárias para se atingir o melhor grau de acerto médio no processo de redução das variáveis. Na quarta coluna quantos casos foram contemplados com a equação selecionada, e na última coluna o grau de acerto médio obtido pela equação selecionada.

Seguindo o passo 6 do algoritmo, deve-se escolher a equação que alcançou o melhor grau de acerto, com isto a equação da primeira combinação (N=8), com grau de acerto médio de 80,1%, foi selecionada para compor o conjunto de equações. Na seqüência, de acordo com o passo 7, foi criada uma nova base de dados de auditorias retirando-se os 673 casos contemplados nesta análise. A nova base de dados passou a ter 4.539 casos, neste momento a quantidade de casos contemplados pelas equações das duas primeiras execuções atingiu 24,75% do total de casos da base de aprendizado. O processo foi reiniciado e em cada execução uma equação foi selecionada.

Após a oitava execução se alcançou um total 85,15% dos casos da base original, de acordo com o passo 5.3 do algoritmo 1 pode-se processar a última equação sem a necessidade de se observar o número mínimo de 4 variáveis na combinação. Na Tabela 20 são listadas todas as equações selecionadas em cada passo; se observa que apenas 334 dos casos não foram processados pelas equações perfazendo 5,54% das auditorias presentes na base de aprendizado (AUDIT-1).

No processamento das bases com declarações mensais dos contribuintes, as equações devem ser executadas na ordem em que foram selecionadas, conforme exibido na Tabela 20. Desta forma, num determinado período, cada declaração da base deverá ser processada pela equação selecionada na primeira execução, se por acaso nem todas as variáveis desta equação estiverem informadas na declaração deve-se passar para a equação selecionada na segunda execução e assim sucessivamente. Este procedimento se repete até que os dados do contribuinte sejam processados por uma das equações ou então esta declaração não terá probabilidade de irregularidade calculada neste período.

Tabela 20 – Equações selecionadas em cada execução.

Execução	Variáveis na equação	% acerto	Casos	% total	% acum.
1	01 - 02 - 03 - 04 - 05 - 07 - 09 - 10 - 11 - 12	76,4	820	13,59	-
2	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11	80,1	673	11,16	24,75
3	01 - 02 - 04 - 06 - 07 – 08 - 10 - 11 – 14	73,9	922	15,29	40,04
4	02 - 07 - 08 - 09 - 11 – 13	72,7	544	9,02	49,06
5	01 - 02 - 03 - 04 - 05 - 06 - 07 – 08	79,3	524	8,69	57,74
6	01 - 02 - 05 – 08	78,8	630	10,44	68,19
7	02 - 04 - 05 – 06	76,3	760	12,60	80,79
8	01 - 03 - 07 - 08	73,0	263	4,36	85,15
9	01 - 02	74,5	562	9,32	94,46
	Casos não alocados nas equações	-	334	5,54	100,00
	Total		6.032		

Outro programa foi desenvolvido em *MS Visual Basic* para processar as declarações mensais atendendo a regra de prioridade nas equações. Este programa denominado *PGM-03*, calcula as probabilidades e apresenta quadros sumarizados para análises posteriores. O programa também oferece totalização pelo grau de acerto médio no caso de ser processada uma base que contenha resultados de auditorias para o período a ser analisado. Com auxilio deste programa, a base completa AUDIT-1 com 6.032 casos foi processada para se obter o grau de acerto médio final na base de aprendizado. O Quadro 3 exibe a classificação dos casos onde se obteve 74,1% de acerto médio nesta etapa.

Quadro 3 – Classificação (a) para a base AUDIT-1 (final)

		Previsto		
		op_Notif		%
Observado		0	1	Corretos
op_Notif	0	832	1.056	44,1
	1	421	3.389	89,0
% média				74,1

(a) Valor de corte é 0,50

O valor obtido com o conjunto das nove equações de 74,1% no grau de acerto médio é considerado satisfatório e atende aos propósitos deste trabalho. Com um valor de corte igual a 0,500 o modelo ajustou corretamente 832 das 1.888 observações não notificadas perfazendo um acerto de 44,1%. No caso das auditorias que tiveram notificações o modelo apresentou um acerto de 89,0%, ajustando corretamente 3.389 das 3.810 observações notificadas. Outra informação pertinente é a distribuição de casos em cada equação e também

a quantidade de casos que não foram atendidos nas equações, na Tabela 21 esta informação é apresentada:

Tabela 21 – Distribuição de casos em cada equação na base AUDIT-1

Equações	01	02	03	04	05	06	07	08	09	sem	Total
Casos	820	673	922	544	524	630	760	263	562	334	6.032

O percentual de empresas que não tiveram probabilidades calculadas foi de 334 perfazendo 5,54% dos casos. Para verificar a consistência dos resultados obtidos na base de treinamento, foram processados com auxilio do *PGM-03* os 1.525 casos da base AUDIT-2 com as auditorias relativas ao primeiro semestre de 2007. No Quadro 4, se apresentam os resultados com o grau de acerto médio obtido e pode-se observar que o resultado final teve comportamento semelhante ao obtido na base de aprendizado, demonstrando consistência no procedimento.

Quadro 4 – Classificação (a) para a base AUDIT-2 (final)

		Previsto		
		op_Notif		%
Observado		0	1	Corretos
op_Notif	0	56	302	15,6
	1	112	956	89,5
% média				71,0

(a) Valor de corte é 0,50

Com relação à distribuição dos casos, apresentados na Tabela 22, nota-se que apenas 99 casos não tiveram probabilidades calculadas.

Tabela 22 – Distribuição de casos em cada equação na base AUDIT-2

Equações	01	02	03	04	05	06	07	08	09	sem	Total
Casos	183	185	201	129	181	187	202	68	90	99	1.525

4.6 – CLASSIFICAÇÃO DOS CONTRIBUINTES

A partir do conjunto de equações selecionadas e com auxilio do programa *PGM-03* foi gerada a relação com a classificação dos contribuintes. Para isto, basta ser selecionado no programa o mês/ano de referência desejado na base das declarações mensais. Nesta aplicação foram processados todos os meses do ano de 2006 e também o primeiro semestre de 2007. A Tabela 23 apresenta a distribuição dos casos pelas equações em cada período de referencia. Pode-se observar que a quantidade de declarações onde não foram calculadas probabilidades ficou pouco acima de 10%.

Tabela 23 – Distribuição de casos em cada equação na base mensal DIME

	Equações										
Ref.	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**sem**	**Total**
2007-06	2.401	1.944	3.294	2.594	4.284	3.169	5.900	1.800	5.484	4.969	35.839
2007-05	2.440	2.017	3.213	2.725	4.312	3.071	5.854	1.689	5.745	4.773	35.839
2007-04	2.409	1.985	3.324	2.729	4.304	3.129	5.907	1.772	5.837	4.443	35.839
2007-03	2.436	1.884	3.440	2.723	4.209	3.153	5.977	1.758	5.806	4.453	35.839
2007-02	2.290	1.856	3.508	2.735	4.080	3.255	5.893	1.862	6.066	4.294	35.839
2007-01	2.359	1.815	3.511	2.757	3.990	3.340	5.914	1.904	6.087	4.162	35.839
2006-12	2.401	1.945	3.276	2.770	4.049	3.419	5.990	1.930	5.976	4.083	35.839
2006-11	2.390	1.918	3.477	2.831	4.145	3.326	6.036	1.812	6.235	3.669	35.839
2006-10	2.391	1.969	3.472	2.816	4.157	3.203	6.035	1.906	6.321	3.569	35.839
2006-09	2.362	1.948	3.465	2.829	4.120	3.295	6.091	1.883	6.340	3.506	35.839
2006-08	2.339	1.949	3.477	2.844	4.142	3.247	6.065	1.840	6.361	3.575	35.839
2006-07	2.339	1.904	3.567	2.853	4.041	3.224	6.120	1.933	6.524	3.334	35.839
2006-06	2.251	1.955	3.363	2.777	4.086	3.274	6.003	2.007	6.111	4.012	35.839
2006-05	2.368	2.015	3.464	2.824	4.003	3.302	6.043	1.888	6.360	3.572	35.839
2006-04	2.287	1.961	3.298	2.849	3.745	3.342	5.899	1.965	6.316	4.177	35.839
2006-03	2.274	1.887	3.538	2.816	3.862	3.241	5.888	1.831	6.205	4.297	35.839
2006-02	2.162	1.796	3.585	2.846	3.636	3.342	5.868	1.933	6.307	4.364	35.839
2006-01	2.207	1.783	3.520	2.811	3.487	3.307	5.839	1.827	6.354	4.704	35.839

Foram geradas probabilidades de irregularidade para cada declaração processada e o programa exibe a classificação ordenada pelos valores de maiores probabilidades. Na Tabela 24 pode-se observar a relação das 25 empresas com maior probabilidade de apresentarem irregularidades na declaração mensal do período de junho de 2007. É apresentado na ultima coluna qual equação gerou a probabilidade para cada empresa.

Tabela 24 - Classificação das 25 empresas com maior probabilidade de incorrerem em notificação período: 2007-06

Código	RUC	Prob. (%)	Eq.
23197	aaaa	98,7805	9
33141	bbbb	98,6086	2
13147	ccccc	98,5449	9
17820	dddd	98,4791	2
24026	eeee	98,4741	2
22431	fffff	98,4647	2
2348	gggg	98,4533	2
17820	hhhh	98,4189	9
26872	iiiiiiii	98,4023	5
9419	jjjjjj	98,3157	6
21228	kkkkk	98,2899	6
5619	lllllll	98,2727	2
19206	mmmm	98,2689	2
14536	nnnnn	98,2537	2
31966	ooooo	98,2535	1
30406	pppppp	98,2424	1
5090	qqqqq	98,2418	7
16830	rrrrrrrr	98,2271	7
8361	sssssss	98,1661	2
29139	tttttttt	98,1394	9
15879	uuuuuu	98,1286	4
7486	vvvvvv	98,1067	2
10278	xxxxxx	98,1002	1
4678	yyyyyy	98,0875	2
27453	wwwww	98,0647	3

O programa *PGM-03* possui diversos filtros para exibir as probabilidades calculadas permitindo visualizar a classificação para uma região geográfica especifica, ou então uma determinada atividade econômica ou grupo setorial. O apêndice J ilustra algumas opções do programa e no apêndice I, o modelo de dados utilizado para a solução informatizada.

4.7 – CONSIDERAÇÕES

Após diversos testes, a utilização do agrupamento elaborada na segunda etapa do modelo de classificação mostrou ser um instrumento importante para obtenção de modelos de probabilidade. A disponibilização dos programas de apoio *PGM-01, PGM-02 e PGM-03* facilitaram a obtenção de resultados rápidos e poderá ser de muita importância para a análise e seleção de contribuintes a serem auditados.

Deve-se destacar que apesar da baixa quantidade de casos que não são processados, isto é, que não apresentam valores a todas as variáveis das nove equações desta solução, é importante monitorar quais são estes casos. Posteriormente, a partir do monitoramento pode-se verificar a possibilidade de mudanças no algoritmo proposto para atender a um maior número de casos.

Importante salientar que a dinamicidade da economia moderna e mudanças na legislação correlata, afetam as características gerais que deram ensejo ao agrupamento desenvolvido nesta pesquisa, bem como às equações de probabilidade. Faz-se necessário novamente acompanhar os resultados futuros nas próximas auditorias para se estabelecer o momento oportuno de se re-estimar os modelos.

5 - CONCLUSÕES

Devido ao grande volume de dados apresentados mensalmente pelas empresas contribuintes ao Fisco Estadual, a atividade de auditoria dispõe de poucas ferramentas para selecionar quais empresas possuem maiores indícios de irregularidades nas suas declarações. Os métodos existentes para seleção de empresas a serem auditadas geralmente são baseados no volume de faturamento mensal, bem como indicadores setoriais e/ou regionais.

Nesta tese foi desenvolvido um modelo para classificação dos contribuintes baseado no CRISP-DM e em recentes aplicações de *data mining*. Após uma breve revisão dos conceitos e técnicas empregadas no *data mining*, a solução é descrita no capítulo 4. O modelo sugerido é composto de quatro etapas: obtenção dos dados, criação dos agrupamentos, elaboração das probabilidades e, classificação dos contribuintes.

O primeiro desafio foi a escolha das técnicas a serem utilizadas tanto na etapa de criação dos agrupamentos quanto para elaboração das probabilidades em cada declaração mensal apresentada. Para a segunda etapa, foi adotada a análise de cluster *two-step* pela facilidade para manipular tanto variáveis contínuas quanto categóricas, bem como processar grandes volumes de dados.

O emprego de análise multivariada possibilitou a definição mais precisa dos perfis de contribuintes no Estado de Santa Catarina, que atualmente se baseia unicamente no binômio setor econômico versus região geográfica, para agrupar as empresas contribuintes. Na solução proposta neste trabalho, além da redução significativa de grupos de 75 para 24, notou-se juntamente com os especialistas da Secretaria da Fazenda uma uniformidade maior no desempenho econômico das empresas em cada um dos 24 grupos. Será possível a construção de perfis mais representativos para o processo de auditoria e seleção de empresas com indícios de fraude ou sonegação de tributos.

Na terceira etapa, elaboração das probabilidades, a maior dificuldade encontrada ficou por conta da técnica escolhida, regressão logística, que processa conjunto de dados onde todas as variáveis tenham valores informados. Tal pressuposto não é encontrado nas declarações mensais, com isto desenvolveu-se um algoritmo que atendesse ao maior número possível de declarações. Este procedimento se baseia num conjunto de equações de regressão logística que são formuladas a partir de sucessivas iterações na base de dados original. Após cada iteração uma nova equação é selecionada e os casos processados por esta equação são

retirados da base original. Nesta aplicação prática nove equações foram selecionadas e apenas 5,54 % dos casos da base original não foram processados por nenhuma das equações.

Para averiguação do comportamento do conjunto de equações selecionadas, os dados das auditorias foram divididos em duas bases: uma de aprendizado e outra para validação da solução. Os resultados obtidos foram considerados relevantes na avaliação dos auditores fiscais, com um acerto médio de 74,1% dos dados de aprendizado e 71,0% na base de validação. A etapa seguinte foi a elaboração do relatório com a classificação das empresas utilizando-se o conjunto de equações. Foram processadas as declarações de 35.839 empresas contribuintes para os meses de janeiro de 2006 até junho de 2007.

Os gestores da atividade de fiscalização podem utilizar a classificação resultante deste trabalho para direcionar os esforços de auditoria. Nas análises realizadas com os auditores fiscais o retorno obtido foi plenamente satisfatório com relação aos indícios de irregularidades calculado para as empresas. Pode-se observar também um grande interesse na implementação desta solução por meio de um aplicativo informatizado.

5.1 – PRINCIPAIS RESULTADOS

Os resultados obtidos na aplicação do estudo de caso no capítulo 4 mostraram a consistência do modelo de classificação proposto e permite sugerir que a mesma seja utilizada nos demais Órgãos Fazendários. A execução da etapa de criação dos grupos permitiu uma classificação das empresas de forma mais homogênea e contribuiu para a obtenção de melhores resultados na elaboração das equações de regressão logística.

Devido à característica particular das declarações mensais, que não obriga o preenchimento de todas as informações por todas as empresas contribuintes, fez-se necessário o desenvolvimento de um conjunto de equações que permitissem o cálculo de indícios de probabilidade. A elaboração do algoritmo 1 no capítulo quatro permitiu a geração de probabilidades para 94,46% das empresas na fase inicial com os dados das auditorias realizadas. Na seqüência foram processadas as declarações de dezoito meses com informações de todas as empresas contribuintes e geradas probabilidades para 90% das declarações.

Três programas específicos foram desenvolvidos e disponibilizados para a Secretaria da Fazenda de Santa Catarina:

- Preenchimento de dados faltantes nas DIME's anuais para geração dos agrupamentos;
- Criação do quadro com todas as combinações de variáveis e quantidades de informantes em cada combinação;
- Processamento da base mensal de declarações e cálculo das probabilidades após seleção da melhor equação para cada empresa contribuinte.

Estes programas executam procedimentos de apoio que facilitam a utilização do modelo de classificação sugerido e os trechos relevantes dos códigos fonte estão listados nos apêndices F, G e H deste trabalho.

6.2 – LIMITAÇÕES DO TRABALHO

Diversas dificuldades foram encontradas no transcorrer da pesquisa, desde a mudança na legislação tributária até a indisponibilidade de fontes de dados que inicialmente seriam utilizadas. A validação do conjunto de equações ficou restrita aos dados de trinta meses considerados na pesquisa, mas é um item que deverá ser melhor avaliado em futuros experimentos.

A complexidade da tarefa se reflete no volume de dados envolvidos, foram desenvolvidos diversos programas para extração das informações do sistema operacional em ambiente de grande porte. Na análise de cluster para a construção dos agrupamentos fez-se necessário a divisão da base de dados devido à quantidade de informações a serem processadas.

5.3 – RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Outras técnicas podem ser utilizadas para se obter as probabilidades de irregularidade nas declarações financeiras, entre elas árvores de decisão e redes neurais. Sugere-se também a inclusão de novas fontes de dados das empresas, como volume de energia consumido, valores

disponibilizados pelas operadoras de cartão de crédito e também dados específicos da mão de obra de cada empresa contribuinte.

Dois procedimentos merecem atenção especial nesta metodologia, a criação de grupos e o cálculo das probabilidades, que foram executados a partir das informações existentes na base de dados. Como a atividade econômica é dinâmica e novas informações são acrescentadas ao banco de dados todos os meses, é conveniente se dispor de procedimentos de monitoramento que possam indicar o momento mais adequado para se ajustar o modelo. Neste caso, tanto a criação dos agrupamentos quanto o conjunto de equações utilizadas para elaborar as probabilidades de irregularidades, podem ser acompanhados e criados indicadores que dêem ensejo ao procedimento de atualização dos mesmos.

Outra proposição é o desenvolvimento de um modelo de previsão para os valores a serem notificados no caso das empresas auditadas estarem cometendo irregularidades. Este procedimento pode ser implantado após a etapa três do modelo proposto, cálculo das probabilidades, e possibilitará um melhorar direcionamento das equipes de auditoria. A afirmação anterior se justifica uma vez que o possível retorno, no caso de haver notificação, pode ser usado juntamente com as probabilidades calculadas para se maximizar o resultado dos processos de auditoria.

REFERÊNCIAS BIBLIOGRÁFICAS

ARENS, A. A.; ELDER, R. J.; BEASLEY, M. S. **Auditing and Assurance Services:** An Integrated Approach. 9. ed. New Jersey: Prentice Hall, 2002.

ATALIBA, Geraldo. Reforma Tributária, Não Constitucional, In: GUTIERRES, Lourdes *et al.* (orgs.), **Aspectos da Questão Tributária no Brasil**, São Paulo: Fundação Getúlio Vargas e Unafisco, 1995.

BACHER, J.; WENZIG, K.; VOGLER, M. SPSS Two Step Cluster – A first evaluation. **Universitat Erlangen**. Nurnberg, 2004.

BARRETO, Alexandre S. **Previsão de Comportamento e Classificação de Contribuintes Tributários:** Uma Abordagem por Modelos Lineares Generalizados Hierárquicos. 2005.Tese (Doutorado em Engenharia de Produção) - Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis, 2005.

BASTOS, Rui M. P. da C. **Auditoria Tributária:** uma abordagem Conceitual. Disponível em: <http://www.infocontab.com.pt/download/revInfocontab/2006/10/auditoria.pdf>. Acesso em: 28 mai. 2008.

BERRY, Michael J. A.; LINOFF, Gordon. **Data Mining Techniques for Marketing, Sales and Customer Support**. 2. ed. New York: John Wiley & Sons, 2004.

BEZERRA, Francisco A. Análise fatorial. In: Corrar, Luiz J.; Paulo, Edílson; Dias Filho, José M. (Org.). **Análise multivariada para os cursos de Administração, Ciências contábeis e Economia**. São Paulo: Atlas, 2007.

BIASOTO JR., G. *et al.*.O ICMS Hoje**:** avanços e questões em aberto sobre a tributação do consumo no Brasil. In: XXVI Brazilian National Meeting of Economics, Vitória, 1998.. **Anais...** , vol. 2. p.891-900. Vitória: ANPEC, 1998.

BOMFIM, Gustavo A. **PPD-PP Conclusão.** Xerox de material de aula. Rio de Janeiro: PUC-RJ, 1999.

BONCHI, F. *et al.* **Using data mining techniques in fiscal fraud detection.** Disponível em: <www.citeseer.ist.psu.edu/bonchi99using.html>. Acesso em: 07 jun. 2008.

BORBA, Cláudio. **Direito tributário:** teoria e 1000 questões. Atualizado até a emenda n. 42/2003 e pelo novo código civil. Série Provas e Concursos. 15. ed. Rio de Janeiro: Impetus, 2004.

BORDIN, L. C. V. **A origem dos tributos:** estudos econômico-fiscais. Governo do Estado do Rio Grande do Sul, Secretaria da Fazenda, Departamento da Receita Pública Estadual, Divisão de Estudos Econômico-Tributários, ano 8, n. 9. nov. 2002.

BOTTOU, L.; BENGIO, Y. **Convergence properties of the k-means algorithms.** In: Advances in neural information processing systems – Editores: Tesauro, G.; Touretzky, D.; Leen, T. vol. 7. p. 585-592. The MIT Press. 1995.

BUSSAB, W. O., MIAZAKI, E.S.; ANDRADE, D.F.. **Introdução à Análise de Agrupamentos** IN: Anais do 9º. Simpósio Nacional de Probabilidade e Estatístico. São Paulo: Associação Brasileira de Estatística /ABE, julho, 1990.

CABENA, Peter *et al.* **Discovering Data Mining:** From Concept to Implementation. New Jersey: Prentice Hall, 1998.

CASTANHEIRA, Nuno M. C. **Auditoria fiscal:** conceito e âmbito de aplicação. Disponível em: <http://www.infocontab.com.pt/download/revInfocontab/2007/21/auditoriafiscal.pdf>. Acesso em: 28 mai. 2008.

CECIL, H. Wayne. Assuring individual taxpayer compliance: Audit rates, selection methods, and electronic auditing. **The CPA Journal**, Nova Iorque, dez.1998.

CHIU, T. *et al.***A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment**. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 263-268. XXXX, 2001.

CORVALÃO, Eder D. **Previsão da Arrecadação do Imposto sobre Circulação de Mercadorias e Serviços em Santa Catarina: Aplicação da Abordagem Geral para Específico em Modelos Dinâmicos**. 2002. Dissertação (Mestrado em Engenharia de Produção) - Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis, 2002.

DEHER, Jorge A. Tecnologia para o Contribuinte. **Serpro – Tema,** edição 187, set-out 2006. Entrevista. Disponível em: <http://www.serpro.gov.br/imprensa/publicacoes/Tema/tema_187/materias/entrevista/>. Acesso em: 08 jun. 2008

DEMPSTER, A.P. Logicist statistics I. Models and modeling. **Statistical Science**, 13, n. 3, p. 248-276. 1998.

DIAS FILHO, J. M.; CORRAR, L.J. Regressão Logística. In: **Análise multivariada para os cursos de Administração, Ciências contábeis e Economia**. FIPECAFI – Coordenadores: Luiz J. Corrar, Edilson Paulo e José Maria Dias Filho. Cap. 5. São Paulo: Atlas, 2007.

DILLY, R.**Data Mining:** an Introduction. Belfast: Parallel Computer Centre, Queens University, 1999.

FAYYAD, Usama *et al.* (editores). **Advances in Knowledge Discovery and Data Mining**. AAAI/MIT Press, 1996.

FUTEMA, F. Sonegação fiscal cresce e atinge quase 30% das empresas, diz IBPT. **Folha de São Paulo**, São Paulo, 18 ago. 2005. Caderno dinheiro.

GIUDICI, P. **Applied Data Mining: Statistical methods for business and industry.** John Willey & Sons; London, 2003.

GUPTA, Manish; NAGADEVARA, Vishnuprasad. **Audit Selection Strategy for Improving Tax Compliance:** Application of Data Mining Techniques. In: INTERNATIONAL CONFERENCE ON E-GOVERNANCE, 5., 2007, Hyderabad. Foundations on E-government. Hyderabad: Ashok Agarwal

And V Venkata Ramana, 2007. p. 378 - 387. Disponível em: <http://www.iceg.net/2007/books/1/39_354.pdf>. Acesso em: 23 jun. 2008.

HAND, David J. Data Mining: Statistics and more? **The American Statistician**, 52. n. 2, p.112-118. mai. 1998.

HAND, David J. Mining the past to determine the future: Problems and possibilities. **International Journal of Forecasting**, 25. n. 3, p.441-451. 2009.

HARDLE, Wofgang; SIMAR, Léopold. **Applied Multivariate Statistical Analysis.** 2. ed. Heidelberg: Springer, 2007.

HERDEIRO, Roberto F. Escalonamento Multidimensional. In: **Análise multivariada para os cursos de Administração, Ciências contábeis e Economia.** FIPECAFI – Coordenadores: Luiz J. Corrar, Edilson Paulo e José Maria Dias Filho. Cap. 7. São Paulo: Atlas, 2007.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression.** 2. ed. New York: Wiley, 1989.

IME**. Curso de Inteligência Tecnológica**, 2005. Disponível em: <http://www.de9.ime.eb.br/~intec/Data%20Mining/Artigos%20de%20Suporte/Overview%20Data%20Mining.pdf>. Acesso em: 10 jun. 2008.

INMON, W.H.; HACKATHORN, R. D. **Como usar o Data Warehouse.** Rio de Janeiro: Infobook, 1997**.**

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. New Jersey: Prentice-Hall, 1992.

KLEINBAUM, D. G.; KLEIN, M. **Logistic Regression. A self-learning text.** 2. ed. New York: Springer-Verlag, 2002.

KOTSIANTIS, Sotiris *et al*. Forecasting Fraudulent Financial Statements Using Data Mining. **Interntaional Journal of Computational Intelligence.** v. 3(2); p. 104-110. Heidelberg: Springer, 2006.

KUMAR, Anuj; NAGADEVARA, Vishnuprasad. **Development of Hybrid Classification Methodology for Mining Skewed Data Sets** – A Case Study of Indian Customs Data. 4th ACS/IEEE International Conference on Computer Systems and Applications, Sharjah, UAE. 2006.

LAKATOS, E. M.; MARCONI, M. A. **Metodologia científica**. São Paulo: Atlas, 1982.

LAGEMANN, Eugenio; KIELING, Edgar Germano; GUADAGNIN, Luís Alberto. **Fiscalização Setorial: Reflexões acerca de uma experiência.** Disponível em: <http://www.via-rs.net/pessoais/luisguadagnin/artigosetorial.htm>. Acesso em: 27 jun. 2008.

LAROSE, Daniel. **Data Mining Methods and Models.** New Jersey: John Wiley & Sons, 2006.

LIMA, Gilmário M. *et al*. **Planejamento de Fiscalização:** Fiscalização Vertical em Estabelecimentos Comerciais. Salvador: Unifacs, 2002. 13 p. Disponível em: <http://intranet.sefaz.ba.gov.br/gestao/rh/treinamento/monografia_gilmario_luiz_maria_mauricio.pdf> . Acesso em: 27 jun. 2008.

LOPES, Maria I. V. de. **Pesquisa em comunicação.** 6 ed. São Paulo: Loyola, 2001.

MICCI-BARRECA, Daniele; RAMACHANDRAN, Satheesh. **Improving Tax Administration with Data Mining.** Analytics Elite, Disponível em: <http://www.spss.com>. Acesso em: 20 Jun. 2008.

MURRAY, Matthew N. Sales Tax Compliance and Audit Selection. **National Tax Journal**, p. 515-530. 01 dez. 1995.

NETER, J. *et al.* **Applied Linear Regression Models**. 4 ed. Chicago:Irwin, 1996.

NISBET, R.; ELDER, J.; MINER, G. **The Handbook of Statistical Analysis & Data Mining applications**. Academic Press, 2009.

OLSON, David L.; DELEN, Dursun. **Advanced Data Mining Techniques.** Berlim: Springer, 2008.

PARK, Hea-Sook; JUNG, Eun-kyung; BAIK, Doo-Kwon. **Client-class based admission control algorithm in web server using cluster analysis.** Proceedings of Third International Conference on Software Engineering Research, Management and Aplications (SERA`05). Washington:IEEE Computer Society, 2005

PASSINI, Sílvia R. R.; TOLEDO, Carlos M. T. Mineração de Dados para Detecção de Fraudes em Ligações de Água. **XI SEMINCO - Seminário de Computação**, p. 229-242. Blumenau, 2002.

PEDUZZI, P. *et al.* A simulation of the number of events per variable in logistic regression analysis. **Journal of Clinical Epidemiology**. v. 49, n. 12, p. 1373-1379. Elsevier Inc. 1996.

PHUA, Clifton *et al.* **A comprehensive survey of data mining-based fraud detection research**. 2005. Disponível em: < http://www.bsys.monash.edu.au/people/cphua/>. Acesso em: 23 jun. 2008.

POHLMANN, Marcelo C. Análise de Conglomerados In: **Análise multivariada para os cursos de Administração, Ciências contábeis e Economia.** FIPECAFI – Coordenadores: Luiz J. Corrar, Edilson Paulo e José Maria Dias Filho. Cap. 6. São Paulo: Atlas. 2007

QUEIROGA, Rodrigo M. **Uso de técnicas de data mining para detecção de fraudes em energia elétrica**. 2005. Dissertação (Mestrado em Informática) – Universidade Federal do Espírito Santo, UFES, Vitória, 2005.

RENCHER, A.C. **Methods of Multivariate Analysis**. 2 ed. New York: John Wiley and Sons, 2002.

ROGER, R. J.; GEATZ, M. W. **Data Mining: A Tutorial-Based primer.** Boston: Addilson Wesley, 2003**.**

RUBIN, R. **[DW Sefaz] Doses de Eficiência.** Disponível em: <http://www.celedo.com.br/portal/modules.php?name=News&file=article&sid=30>. Acesso em: 27 jun. 2008.

RUD, Olivia P. **Data mining cookbook:** modeling data for marketing, risk, and customer relationship management. New York: John Wiley & Sons, 2001.

SÁ, Antônio L. **Fraudes Contábeis**, 2. ed. Rio de Janeiro: Tecnoprint, 1982.

SAYEG, Roberto. Sonegação Tributária e Complexidade. In: **RAE-eletrônica**, São Paulo, v. 2, n. 1, jan./jun. 2003. FGV. Disponível em: <http://www.rae.com.br/eletronica/index.cfm?FuseAction=Artigo&ID=1359&Secao=PÚBLICA&Volume=2&Numero=1&Ano=2003>. Acesso em: 26 jun. 2008.

SFERRA, Heloisa H.; CORRÊA, Ângela M. C. J. Conceitos e Aplicações de Data Mining. **Revista de Ciência & Tecnologia**, v.1, n. 22, p. 19-34. jul./dez. 2003.

SHARMA, S. **Applied Multivariate Techniques**. New York: John Wiley & Sons, 1996.

SIQUEIRA, Marcelo L.; RAMOS, Francisco S. A economia da sonegação: teorias e evidências empíricas. **Revista de Economia Contemporânea**. Rio de Janeiro. v. 9, n. 3, p. 555-581, set./dez. 2005.

SOUZA, Carlos R. S. **Secretaria da Fazenda do estado da Bahia:** um case de transformação organizacional através da Tecnologia da Informação. In: CONGRESSO INTERNACIONAL DEL CLAD SOBRE LA REFORMA DE ESTADO Y DE LA ADMINISTRACIÓN PÚBLICA, 7. 8-11 out. Lisboa, 2002.

SPATHIS, Charalambos T. Detecting false financial statements using published data: some evidence from Greece, **Managerial Auditing Journal,** Vol. 17, n. 4, p. 179-191, 2002.

SPSS. Statistical Package for Social Science. **SPSS for windows: standard version**, release 13,0. Chicago, 1999.

SUMATHI, S.; SIVANANDAM, S.N. **Introduction to Data Mining and its Applications.** Berlim: Springer, 2006.

VALDERRAMA, J.L. **Teoria y Prática de la Auditoria**. Madrid: CDN, Ciencias de la Direccion, 1996.

VIANNA NETTO, Matteus. **ICMS: a lei complementar n. 87/96 interpretada**. São Paulo: Editora de Direito Ltda., 1997.

VIGLIONI, Giovanni M. C. **Metodologia para previsão de demanda ferroviária utilizando data mining**. 2007. Dissertação (Mestrado em Engenharia de Transportes) - Instituto Militar de Engenharia, IME. Rio de Janeiro, 2007.

VIRDHAGRISWARAN, Sankar; DAKIN, Gordon..O ICMS Hoje: avanços e questões em aberto sobre a tributação do consumo no Brasil. In: 12th. ACM SIGKDD International conference on Knowledge discovery and Data mining, Philadelphia-PA, USA Vitória, 1998.. **Anais**..., vol. 2. p.941-947. New York: ACM, 2006.

ZHANG, T.; RAMAKRISHNON, R.; LIVNY, M. **BIRCH: An Efficient Data Clustering Method for Very Large Databases**. Proceedings of the ACM SIGMOD Conference on Management of Data, p. 103-114. XXXX, 1996.

ZORNETZER, S. *et al.* (editores). **An Introduction to Neural and Electronic Networks**. 2a. Ed. London: Academic Press, 1994.

APÊNDICE A – O ICMS E A ADMINISTRAÇÃO TRIBUTÁRIA

A.1 – ASPECTOS HISTÓRICOS

O processo de tributação remonta à idade antiga, no Egito qualquer mercadoria em trânsito era tributada, Bordin (2002) nos mostra a origem e evolução dos tributos, onde se destaca:

- Roma - 09 d.C., o Imperador Augusto instituiu a centésima *rerum venalium*, que incidia sobre o giro dos negócios, destinando-se a financiar os gastos militares;
- França – 1292, o Rei Felipe impôs alíquotas de 5 a 12% sobre todas as compras e vendas;
- Brasil – 1923, a União instituiu o "imposto sobre vendas mercantis" (alíquota de 0,30%) e em 1934 passou a se chamar "imposto sobre venda e Consignações" (alíquota de 5% a 7%);
- França – 1952, IVA Francês (imposto sobre valor agregado).

Para um melhor entendimento do tema desenvolvido neste trabalho os seguintes conceitos merecem destaque:

- **Tributos:**

Formam a receita da União, Estados e municípios e abrangem impostos, taxas, contribuições e empréstimos, devendo ser disciplinado por normas do Direito Público, que constituem o Direito Tributário.

- **Imposto:**

É um tributo não vinculado a uma ação governamental, sendo utilizado normalmente para o financiamento de serviços universais, como saúde, educação e segurança, visando promover o bem comum. Podem incidir sobre o patrimônio, renda e o consumo. (Art. 16 CTN).

- **Taxa:**

É o tributo vinculado a um serviço público específico prestado ao contribuinte ou posto à sua disposição, conforme define o art. 18, inc. II do Código Tributário Nacional. "Taxa é o tributo vinculado, cuja hipótese de incidência consiste numa atuação estatal direta e imediatamente referida ao obrigado". (ATALIBA, 1995).

- **Contribuições**:

Dividem-se em dois grupos: de melhoria, cobradas em uma situação que representa um benefício ao contribuinte; e, especiais quando há uma destinação específica para um determinado grupo, como o PIS (Programa de Integração Social) que são direcionados a um fundo de trabalhadores do setor privado.

- **Empréstimos Compulsórios:**

Só podem ser instituídos pela União e servem para atender a situações excepcionais. O governo pode utilizar deste tributo em situações de emergência; calamidade pública; guerra externa, ou a sua iminência. Sua regulamentação encontra-se no art. 148 da atual Constituição e também no art. 14 do CTN.

A.2 - ICMS - IMPOSTO SOBRE CIRCULAÇÃO DE MERCADORIAS E SERVIÇOS

O ICMS, Imposto sobre Circulação de Mercadorias e Serviços, é o principal instrumento tributário dos Estados e Distrito Federal. As alíquotas desse imposto são variáveis dentro do território nacional, pois são fixadas de forma independente pelas legislações estaduais e do Distrito Federal. O ICMS apresenta-se no Brasil como um tributo compulsório sobre operações relativas à circulação de mercadorias e prestação de serviço.

Operações relativas à circulação de mercadorias são os atos ou negócios que implicam na circulação jurídica de mercadoria e que acarretam mudança de propriedade sobre a mercadoria feita dentro da circulação econômica, pois assim será levada da fonte até o consumidor. O imposto também incide sobre serviços de transporte interestadual e intermunicipal, de comunicações, de energia elétrica, de entrada de mercadorias importadas e aqueles serviços prestados no exterior. O ICMS é regulamentado pela Lei Complementar 87/1996, também conhecida como "Lei Kandir".

A origem deste imposto vem do início do século XX com a mudança na estrutura da tributação, que até então era constituída quase que em sua totalidade por impostos incidentes sobre o patrimônio. Corvalão (2002) apresenta o histórico do ICMS até a Constituição Federal de 1988.

O imposto incide no momento da saída da mercadoria do estabelecimento ou no ato da prestação do serviço, sendo não cumulativo, isto é, podendo ser compensado através do crédito obtido quando da entrada de bens de produção e/ou produtos destinados à comercialização. Apresenta como característica a não-cumulatividade, onde para efeito de apuração do tributo devido, deduz-se do imposto incidente sobre a saída de mercadorias o imposto já cobrado nas operações anteriores relativamente à circulação daquelas mesmas mercadorias ou às matérias-primas necessárias à sua industrialização.

A base de cálculo é o valor da operação com mercadorias, incluindo importâncias acessórias e excluídos os descontos incondicionais. As alíquotas apresentam-se como o percentual de carga específica que se lançará sobre o valor (base de cálculo) que tem relação com o ato/fato que gerou a obrigação tributária.

A parcela do imposto a ser paga tem como base o valor da operação comercial ou o preço da prestação do serviço. Conforme Vianna Neto (1997): A base de cálculo do imposto corresponde à totalidade dos elementos econômicos ínsitos à operação de circulação de mercadoria ou prestação de serviços e representáveis em moeda.

A alíquota é de:

17 % - para operações internas e prestação de serviços de comunicação realizadas no Estado, energia elétrica, urbana e rural;

13% - nas operações ou prestações que destinem mercadorias ao exterior;

12% - nas operações interestaduais, transportes interestaduais e intermunicipais, energia elétrica a produtor rural e outros, e nas operações com produtos da cesta básica;

25% - Operações com energia elétrica, prestação de serviço de comunicação, operações com gasolina automotiva e álcool carburante.

O ICMS é um imposto não cumulativo, compensando-se o valor devido em cada operação ou prestação com o montante cobrado anteriormente. Em cada etapa da circulação de mercadorias e em toda prestação de serviço sujeita ao ICMS, deve haver emissão da nota fiscal ou cupom fiscal. Esses documentos serão escriturados nos livros fiscais para que o imposto possa ser calculado pelo contribuinte, por ele lançado e recolhido ao Estado.

A.3 - PROCESSO DE FISCALIZAÇÃO E AUDITORIA

A fiscalização é uma atividade desenvolvida pelos órgãos públicos para controlar as obrigações dos sujeitos passivos dos seus impostos e taxas. Deste modo a fiscalização tem por objetivo lançar os créditos tributários não declarados ou declarados incorretamente e/ou sancionar o descumprimento de outras obrigações formais, que possam se constituir em fraude e/ou sonegação. Neste sentido fraude é toda ação ou omissão dolosa tendente a impedir ou retardar, total ou parcialmente, a ocorrência do fato gerador da obrigação tributária principal, ou a excluir ou modificar as suas características essenciais, de modo a reduzir o montante de imposto devido.

Para Sá (1982): "Fraude é um erro cometido propositadamente com a finalidade de prejudicar alguém". No entendimento de Borba (2004) o conceito de fraude seria uma "tentativa dolosa de impedir ou retardar a ocorrência do fato gerador".

De acordo com Lima *et al.* (2002), uma administração tributária eficaz deve ter como características, também: sistematicidade, simplicidade das normas, serviços de atenção aos contribuintes, controle efetivo quanto ao cumprimento das obrigações tributárias, desenvolvimento efetivo de ações visando recuperação de obrigações não cumpridas voluntariamente, entre outras. Ainda segundo este autor, a função de fiscalização para ser desenvolvida, é mister que:

- O fenômeno da evasão fiscal seja conhecido amplamente: causas, manifestações, dimensões;
- disponha-se de recursos legais, materiais, financeiros, tecnológicos, de informação, humanos, etc.;
- haja domínio de ferramentas de análise que permitam conhecer o contribuinte, seu negócio e seu padrão de relacionamento com o fisco;
- tenha sido tomada a decisão em nível estratégico de enfrentar a evasão fiscal e que essa decisão esteja refletida claramente na explicitação da política de fiscalização e em uma adequada programação e controle de gestão.

Para Arens, Elder e Besley (2002) a auditoria pode ser definida como um processo de acumulação e avaliação de evidências sobre informação, de forma a se determinar e reportar o grau de correspondência entre a informação e os critérios estabelecidos na sua elaboração.

> O processo de auditoria baseia-se em evidências apuradas pelo auditor com o objetivo de apurar a correlação existente entre a informação auditada e os critérios utilizados a que ela se reporta. O auditor irá definir, de forma independente, um montante acumulado de evidências suficientes que o habilitem a relatar a conclusão apropriada. (BASTOS, 2006).

O termo auditoria fiscal é normalmente utilizado para definir as ações de fiscalização tributária desenvolvida pela administração fiscal. Contudo, o seu conceito poderá apresentar conteúdos diferentes em função da posição do sujeito que executa a auditoria.

A auditoria fiscal tem por objetivo fazer um exame da situação fiscal da empresa, tendo em vista o controle da sua regularidade fiscal. Este conceito pode ser ampliado:

> A auditoria da área fiscal persegue um objetivo duplo, em primeiro lugar o de comprovar que a entidade cumpriu, adequadamente, as obrigações tributárias, se provisionou corretamente os riscos derivados de possíveis contingências fiscais e, em segundo lugar, se procedeu ao pagamento efetivo do tributo de acordo com os prazos e formalismos legais (VALDERRAMA, 1996).

A fiscalização tributária é levada a cabo pela Administração Fiscal e visa, essencialmente, verificar se o contribuinte, pessoa singular ou coletiva, cumpriu as suas obrigações fiscais de uma forma correta. Neste contexto, a Administração fiscal procura minimizar a diferença existente entre o imposto definido por Lei e o imposto declarado pelos contribuintes, com o objetivo amplo de combate à fuga e evasão fiscal (CASTANHEIRA, 2007).

A Administração Tributária possui como escopo fundamental o de minimizar a diferença existente entre o imposto declarado pelos contribuintes e o imposto potencialmente definido pela Lei, com o objetivo amplo de combate à fuga e evasão fiscal. Neste sentido, o auditor tributário não se limita a efetuar um relatório sobre a existência, sendo caso disso, de créditos contingentes, mas a propor liquidações tributárias (que poderão ser liquidações adicionais ou liquidações oficiosas), que se irão constituir em dívidas tributárias originando, eventualmente, as respectivas penalidades (BASTOS, *op. cit.*).

A falta de cumprimento das obrigações fiscais implica um risco fiscal que pode provocar a aplicação de penalidades, além da exigência do cumprimento da própria obrigação

fiscal, situações que podem ter conseqüências graves na situação financeira das empresas e na imagem verdadeira e apropriada das demonstrações financeiras (CASTANHEIRA, *op. cit.*).

A.4 - DEFINIÇÃO DE POLÍTICAS PARA AUDITORIA

Planejar a fiscalização significa escolher quais contribuintes devem ser auditados, de forma a assegurar o maior efeito global sobre a percepção de risco no descumprimento das obrigações tributárias e, assim, proporcionar a potencialização dos níveis de cumprimento voluntário. Como a quantidade de auditores fiscais disponível para a execução das auditorias é inferior ao universo de contribuintes, a seleção dos contribuintes merece destaque. Esta seleção deve ser baseada em estudos econômico-fiscais, que procure atingir as principais manifestações de sonegação, através de programas pontuais de fiscalização. Segundo Lima *et al.* (*op. cit.*):

> Os tipos de ações a serem executadas, devem ser eleitos, a partir da classificação dos contribuintes pelo padrão de comportamento em relação ao cumprimento de suas obrigações tributárias e por outras características, quais sejam: o porte econômico, especificidades do negócio, ciclo de vida, abrangência espacial das atividades e modalidades das obrigações a que estão sujeitos.

Visando atingir uma maior área de abrangência as Secretarias Estaduais da Fazenda iniciaram com divisões regionais, sendo criadas unidades específicas de fiscalização. Obteve-se com isto, uma aproximação junto aos contribuintes, o que possibilitou além do conhecimento da realidade regional por parte do fisco estadual uma ação mais imediata quanto necessário.

Constatou-se que a maior parte da arrecadação do ICMS era proveniente de um grupo reduzido de empresas com faturamento muito alto. Esta informação direcionou a atividade de seleção dos contribuintes a serem auditados, com isto, as grandes empresas se tornaram o foco prioritário da análise e descoberta de indícios de fraude. De acordo com Lagemann, Kieling e Guadagnin (2008):

> O foco está centrado no volume de recursos. Assim, os grandes contribuintes trabalham com a certeza de que serão acompanhados de perto pelo Fisco, enquanto as pequenas e microempresas ficam praticamente livres de auditorias, à exceção daquelas possíveis de serem realizadas mediante meios eletrônicos.

Concomitantemente outras ações foram desenvolvidas no processo de seleção para evitar o caráter puramente determinístico da "seleção prioritária" que levaria às empresas de médio e pequeno porte a sensação de impunidade, para tanto dentro de determinados setores algumas empresas eram escolhidas aleatoriamente. Do resultado das auditorias realizadas, as empresas que apresentaram ocorrência de fraude passaram a integrar o grupo de empresas com prioridade de seleção.

A diversidade de atividades econômicas e a excessiva complexidade da legislação pertinente a cada uma destas atividades abriram espaço para outros modelos de planejamento, todos apontando para a especialização dos entes envolvidos na fiscalização.

A.5 - GRUPOS SETORIAIS

A partir da década de 1990 diversas experiências foram realizadas no sentido de buscar alternativas aos modelos existentes. No relato de Sayeg (2003) as novas propostas vão ao encontro de uma visão por sistemas, onde o processo de fiscalização deve ser entendido como um "todo" com lógica e finalidades próprias. Esta visão encontra sua expressão maior na "Fiscalização Setorial" que busca atuar sobre os contribuintes individuais levando em conta o desempenho fiscal do seu respectivo setor econômico.

As equipes de Fiscais assumem a missão de orientar e acompanhar todos os contribuintes do Estado que operam com determinado produto ou desenvolvem certa atividade econômica. Ainda, conforme Lagemann, Kieling e Guadagnin (2008) a fiscalização setorial constitui, essencialmente, uma estratégia de fiscalização por produto ou atividade econômica. Os contribuintes que operam com esse produto ou se dedicam a essa atividade econômica formam o setor. Desta experiência do RS iniciada em 1999, os autores relacionam como funções do grupo setorial:

a) Revisar e atualizar as informações cadastrais, a fim de conhecer com precisão os contribuintes e a sua atividade econômica;
b) conhecer as normas tributárias e não-tributárias que regem o setor;
c) elaborar diagnóstico da situação do setor, inclusive apurando a relevância dele e de cada um dos contribuintes na arrecadação do tributo; e

d) examinar se existem inconsistências no que tange ao enquadramento.

Em Santa Catarina foram criados Grupos Especialistas Setoriais (GES) de abrangência estadual. Conforme o endereço eletrônico da Secretaria da Fazenda do Estado de Santa Catarina[7]: criados a partir de 2003, os principais objetivos dos Grupos de Especialistas Setoriais (GES) são:

a) Acompanhar permanentemente cada setor;
b) obter cognição sobre a organização do setor, através de levantamento dos aspectos técnicos, jurídicos, comerciais, contábeis e tributários;
c) disponibilizar ao Grupo Fisco conhecimento e informações adquiridas sobre o setor;
d) contatar as entidades organizadas representativas do segmento, buscando eliminar a concorrência desleal originada em evasões e elisões fiscais;
e) identificar as empresas que compõem cada segmento do setor;
f) orientar quanto ao cumprimento das obrigações tributárias e estimular a regularização espontânea de débitos;
g) relatar as formas de operação utilizadas pelos contribuintes;
h) verificar as formas de tributação, arrecadação e fiscalização do setor nas demais Unidades da Federação que afetem nosso Estado.

Para criação dos "grupos setoriais" não existe um padrão universal, "A criatividade surge como um ingrediente fundamental para a delimitação dos 'recortes' que irão constituir as 'Setoriais', podendo-se, talvez, antever que a eficiência/eficácia da ação do poder fiscalizador deverá estar relacionada com uma habilidade que pode ser entendida como uma questão de arte" (SAYEG, *op. cit.*).

A Figura 16 apresenta um resumo com as principais divisões setoriais em vigência nos Estados do Rio Grande do Sul, Santa Catarina e São Paulo.

[7] ESTADO DE SANTA CATARINA. SECRETARIA DE ESTADO DA FAZENDA. DIRETORIA DE ADMINISTRAÇÃO TRIBUTÁRIA. **Resultados da DIAT em 2005**. Disponível em: <http://www.cee.sc.gov.br/upload_admin/noticias/sef/RESULTADOSDIATEM2005.pdf>. Acesso em: 01/06/2008.

RS, a partir de 1999	SC, a partir de 2003	SP, a partir de 1997
Metal mecânico	Metalurgia e metal-mecânico	Metalúrgicos
Bebidas	Bebidas	Bebidas
Medicamentos e cosméticos	Farmacêuticos e medicamentos	Farmacêuticos e perfumaria
Lojas de departamentos	Redes de estabelecimentos	Redes de estabelecimentos
Veículos	Automotores	Automotivos
Combustíveis e lubrificantes	Combustíveis e lubrificantes	Químicos e petroquímicos
Energia elétrica	Energia	Energia elétrica
Comunicações	Comunicações	Comunicações
Supermercados	Supermercados	Alimentícios
Agronegócios	Fumo	Madeira
Couro e calçados	Laticínios	Moveis e papel
Atacadistas e alimentos	Transportes	Eletroeletrônicos
Petroquímico	Têxtil	Plásticos e borracha
Moveleiro	Comércio exterior	
	Mercadorias ou bens via correios	
	Materiais de construção	

Figura 16 - Classificação de atividades por grupos setoriais nos Estados do Rio Grande do Sul, Santa Catarina e São Paulo.

A.6 - EXPERIÊNCIAS NO PROCESSO DE SELEÇÃO DE CONTRIBUINTES

Na impossibilidade de serem fiscalizadas todas as empresas que estão sujeitas a apresentação das declarações mensais, os setores de administração tributária intentaram diversas alternativas para seleção dos contribuintes que deveriam ser auditados, partindo-se de escolhas puramente aleatórias, passando em seguida a direcionar o esforço de fiscalização para as empresas com maior faturamento. Um esforço mais avançado se constitui na seleção baseada em informações econômicas da empresa ou do grupo setorial, baseado na área de atividade da empresa.

Para Gupta e Nagadevara (2007), enquanto que a seleção aleatória dá um tratamento igual para os contribuintes honestos e desonestos, já que a probabilidade de seleção é a mesma para ambos; o critério de seleção baseado em informações pressupõem certos

sintomas de não conformidade, que podem representar outros sintomas como mudança na situação econômica num setor em particular.

Na evolução constante dos métodos de seleção, cada vez mais os agentes fiscalizadores procuram agregar novas informações, apoiando-se nos dados disponíveis em diversos sistemas operacionais implantados nas ultimas décadas. Neste sentido o atual Secretário da Receita Federal enfatiza:

> Na administração tributária moderna, intensifica-se a cada dia o uso de sistemas para seleção de contribuintes, cruzando-se múltiplas fontes de informação, para auditoria fiscal e para controle e cobrança dos créditos tributários. Em conseqüência, eleva-se a percepção de risco do contribuinte faltoso, estimulando o cumprimento voluntário das obrigações tributárias. (DEHER, 2006)

Encontram-se disponíveis na literatura diversas experiências de seleção de contribuintes para efeitos de auditoria, destacando-se entre elas o trabalho de Murray (1995) que analisa as características dos tributos não apresentados em conformidade com a legislação e também as técnicas disponíveis para criação de regras de seleção. Também merecem registro as experiências dos Estados do Texas e do Tennessee que utilizam técnicas de regressão (Fisher *Apud* Gupta e Nagadevara, *op. cit.*).

Cecil (1998) relata diversos procedimentos não aleatórios utilizados nos Estados Unidos pelo IRS - *Internal Revenue Service* (equivalente à Receita Federal no Brasil). Entre estes procedimentos destacam-se a contabilização de itens não permitidos e a análise com função de discriminante baseada em dados estatísticos. Semelhante procedimento com função de discriminante foi implementado na Índia, de acordo com Kumar e Nagadevara (2006).

Micci-Barreca e Ramachandran (2006) descrevem a implantação de um sistema empregando técnicas de *data mining* no Estado do Texas para seleção de auditoria. Anteriormente as duas estratégias utilizadas eram:

- *Priority One*: O Estado do Texas classifica todas as atividades econômicas que contribuem para os primeiros 65 por cento dos valores tributados como contribuintes prioritários. O Estado audita estes contribuintes aproximadamente a cada quatro anos;
- *Prior Productive*: Com base no resultado de auditorias anteriores, os contribuintes com ajustamento de tributos com valores maiores que $10.000. são selecionados automaticamente para serem auditados.

No sistema descrito por Micci-Barreca e Ramachandran (*op. cit.*) são utilizadas cinco fontes de dados para a criação de padrões para os contribuintes. Para o processo de seleção o sistema emprega regressão linear e modelos de redes neurais. Como parte integrante do sistema os autores descrevem uma metodologia com seis fases para a sua perfeita implementação: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação.

Técnicas de classificação são utilizadas no trabalho de Bonchi *et al.* (1999) que sugere o uso de árvore de decisão e apresenta uma metodologia para construção de perfis para comportamento fraudulento que abrange: identificação das fontes de dados disponíveis, identificação do custo do modelo, preparação dos dados para análise, construção do modelo, e avaliação.

No Brasil a maioria dos Estados tem privilegiado a seleção baseando-se em critérios setoriais, no caso do Estado da Bahia inicialmente são selecionadas empresas que representam 80% da arrecadação e:

> [...] trabalha também com indicadores de conjuntura – macro e microeconômicos – que permitem estudar o setor em que está inserido o contribuinte e fazer ilações entre eles assim como com o mesmo setor em outros estados da federação. Isso possibilita mapear o comportamento do contribuinte e otimizar o processo fiscalizador. (SOUZA, 2002).

No Estado do Pará é elaborada uma relação de contribuintes a serem fiscalizados, sendo utilizado um critério de pontuação que privilegia o porte da empresa e o volume de transações financeiras a partir dos campos saídas e entradas de mercadorias bem como os valores das operações mercantis em fronteira (realizadas fora do Estado). Este critério prioriza os grandes contribuintes para serem selecionados.

Em Minas Gerais, com a utilização do GECIN (Gerenciamento de Carteiras de Índices) é implementada uma função de *score* total, baseado no sistema que as instituições financeiras usam para avaliar seus clientes, através de pontos obtidos em diversos critérios. Diversos índices são calculados para a criação do *score*. O GECIN utiliza faixas de pontuação que são atribuídas pelos auditores fiscais para agrupar as empresas que apresentam indícios de sonegação e, portanto, devem ser selecionadas prioritariamente.

Ainda com procedimentos direcionados ao acompanhamento mais próximo dos grandes contribuintes, o Estado de Santa Catarina, após implantar os grupos setoriais, tem

acompanhado o comportamento das atividades econômicas e se utiliza de indicadores setoriais para prever o comportamento das empresas pertencentes à atividade específica.

APÊNDICE B – DESCRIÇÃO DADOS DIME ANUAL, ANÁLISE SETORIAL

Tabela 25 – Descrição dados DIME anual – análise setorial

Setor	Freqüência	Fat. Total 2006	Fat. Médio 2006
AGROPECUARIA	499	3.364.107.737,17	6.741.698,87
INDUSTRIA	11.194	112.616.847.794,64	10.060.465,23
COM_ATACADISTA	4.460	56.381.289.830,24	12.641.544,81
COM_VAREJISTA	15.172	44.951.912.100,21	2.962.820,47
SERVICOS	4.525	33.193.653.511,35	7.335.614,04
TOTAL	35.850	250.507.810.973,61	6.987.665,58

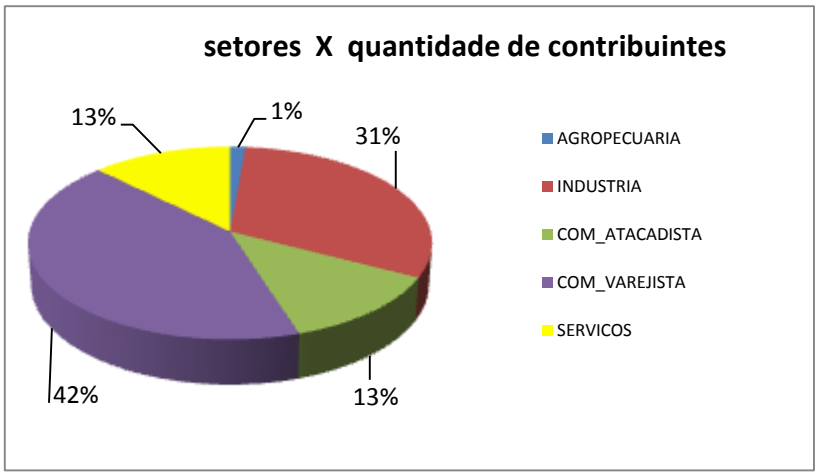

Figura 17 - Setores X Quantidade de contribuintes

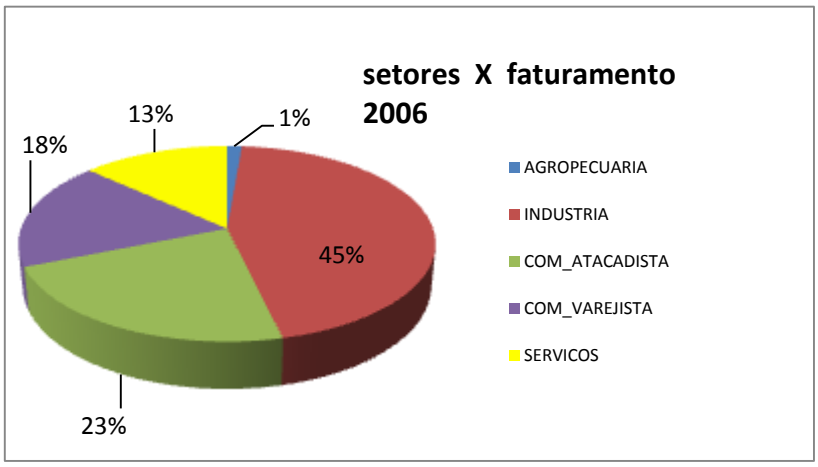

Figura 18 – Setores X Faturamento

APÊNDICE C – DESCRIÇÃO DADOS DIME ANUAL, ANÁLISE REGIONAL

Tabela 26 - Descrição dados DIME anual – análise regional

	Usefi	Freqüência	Fat. Total 2006	Fat. Médio 2006
1	FLORIANOPOLIS	4.711	33.566.276.177,93	7.125.085,16
2	ITAJAI	3.839	42.866.742.676,23	11.166.122,08
3	BLUMENAU	5.063	26.217.127.145,71	5.178.180,36
4	RIO_DO_SUL	1.841	6.428.042.520,01	3.491.603,76
5	JOINVILLE	5.153	60.570.566.466,33	11.754.427,80
6	PORTO_UNIAO	1.058	9.837.251.689,82	9.297.969,46
7	JOACABA	1.689	9.807.464.926,98	5.806.669,58
8	CHAPECO	3.195	18.057.528.209,62	5.651.808,52
9	CURITIBANOS	1.387	7.135.830.781,02	5.144.795,08
10	LAGES	1.266	6.059.391.479,38	4.786.249,19
11	TUBARAO	1.678	8.286.694.240,23	4.938.435,18
12	CRICIUMA	2.031	10.127.562.877,79	4.986.490,83
13	S. MIGUEL DO OESTE	756	2.729.489.644,43	3.610.436,04
14	MAFRA	1.416	6.110.208.007,86	4.315.118,65
15	ARARANGUA	770	2.715.615.413,02	3.526.773,26
	OUTROS	2	3.372.534,94	1.686.267,47
	TOTAL	35.855	250.519.164.791,30	6.987.007,80

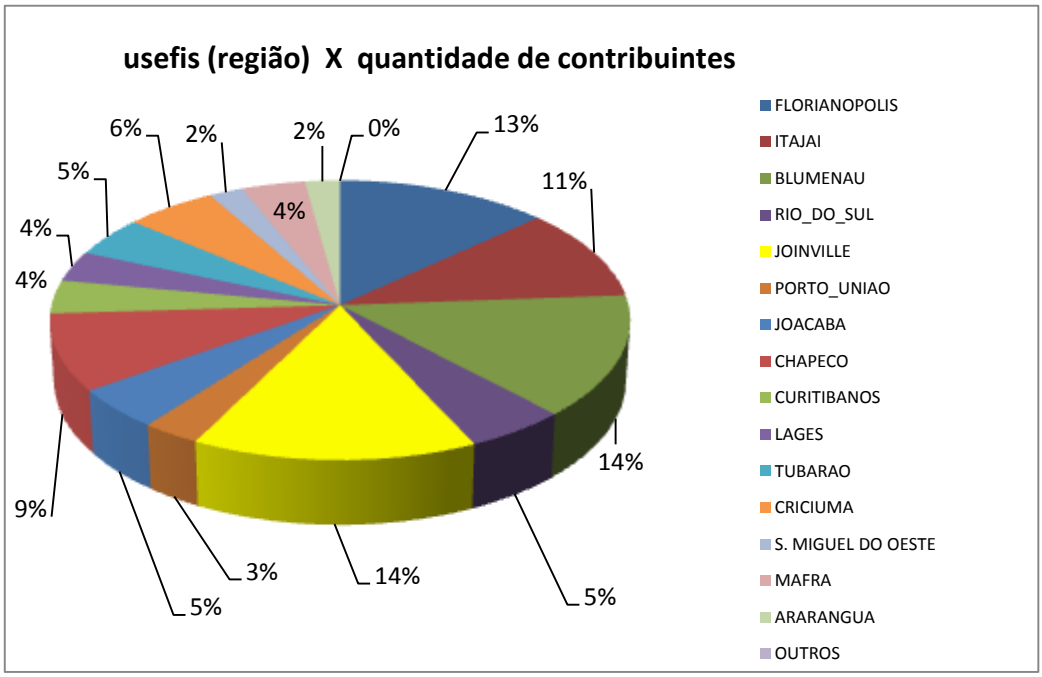

Figura 19 – Usefis X Quantidade de contribuintes

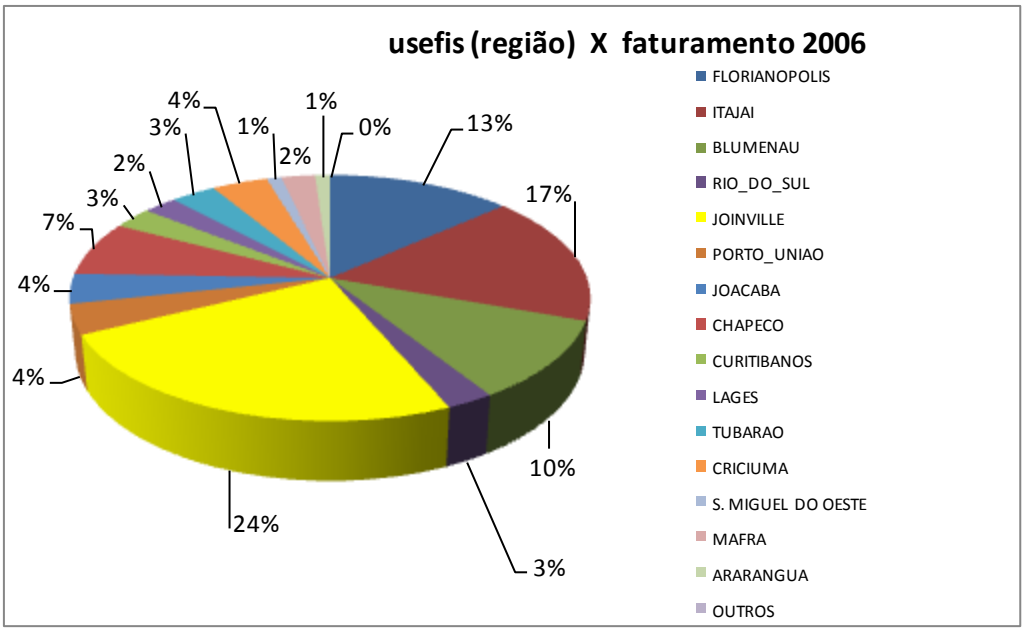

Figura 20 – Usefis X Faturamento

APÊNDICE D – DESCRIÇÃO DADOS DIME ANUAL, CRUZAMENTO DADOS SETORIAL X REGIONAL

Tabela 27 – Cruzamento dados setorial x regional – Por USEFI (Região Fiscal)

	AGROPECUARIA		INDÚSTRIA		COM_ATACADISTA		COM_VAREJISTA		SERVICOS		
	Qtd	Fat. Tot	Qtd	Fat. Tot	Qtd	Fat. Tot	Qtd	Fat. Tot	Qtd	aFat. Tot	
FLORIANOPOLIS	0,57%	0,16%	18,55%	11,26%	13,27%	21,80%	53,34%	27,46%	14,26%	39,32%	100%
ITAJAI	0,70%	0,37%	29,63%	20,84%	15,92%	48,63%	42,61%	12,48%	11,13%	17,67%	100%
BLUMENAU	0,45%	0,15%	49,16%	71,01%	11,06%	8,72%	30,83%	16,61%	8,49%	3,51%	100%
RIO_DO_SUL	0,92%	3,22%	39,82%	56,21%	8,75%	13,58%	42,91%	23,38%	7,60%	3,61%	100%
JOINVILLE	0,70%	0,12%	35,64%	60,59%	11,94%	21,46%	38,35%	9,94%	13,37%	7,88%	100%
PORTO_UNIAO	3,79%	1,53%	23,30%	34,46%	9,94%	5,86%	52,94%	55,40%	10,04%	2,75%	100%
JOACABA	2,37%	2,13%	18,47%	58,74%	9,30%	9,86%	42,45%	14,78%	27,41%	14,48%	100%
CHAPECO	1,28%	3,25%	23,79%	52,47%	17,03%	20,79%	43,44%	15,74%	14,46%	7,75%	100%
CURITIBANOS	5,12%	7,82%	22,28%	48,73%	10,67%	21,54%	44,34%	17,90%	17,59%	4,01%	100%
LAGES	3,71%	3,49%	21,33%	57,05%	11,06%	9,53%	51,42%	24,22%	12,48%	5,71%	100%
TUBARAO	1,87%	5,29%	37,62%	57,31%	13,29%	12,51%	36,29%	19,42%	10,93%	5,47%	100%
CRICIUMA	1,87%	5,29%	37,62%	57,31%	13,29%	12,51%	36,29%	19,42%	10,93%	5,47%	100%
S. MIG. OESTE	1,46%	1,77%	22,62%	48,32%	11,64%	21,25%	50,40%	22,00%	13,89%	6,66%	100%
MAFRA	2,26%	2,31%	37,57%	68,05%	9,46%	6,74%	36,65%	17,99%	14,05%	4,92%	100%
ARARANGUA	1,82%	3,30%	33,64%	44,85%	15,45%	17,93%	42,08%	29,06%	7,01%	4,86%	100%

Tabela 28 – Cruzamento dados setorial x regional – Por Setor economico

	AGROPECUARIA		INDUSTRIA		COM_ATACADISTA		COM_VAREJISTA		SERVICOS	
	Qtd	Fat. Tot	Qtd	Fat. Tot	Qtd	Fat. Tot	Qtd	Fat. Tot	Qtd	Fat Tot
FLORIANOPOLIS	5,41%	1,60%	7,81%	3,36%	14,01%	12,978%	16,56%	20,50%	14,85%	39,77%
ITAJAI	5,41%	4,75%	10,16%	7,93%	13,70%	36,973%	10,78%	11,90%	9,44%	22,82%
BLUMENAU	4,61%	1,19%	22,24%	16,53%	12,56%	4,056%	10,29%	9,69%	9,50%	2,77%
RIO_DO_SUL	3,41%	6,16%	6,55%	3,21%	3,61%	1,548%	5,21%	3,34%	3,09%	0,70%
JOINVILLE	7,21%	2,11%	16,40%	32,59%	13,79%	23,058%	13,02%	13,40%	15,23%	14,38%
PORTO_UNIAO	8,02%	4,48%	2,20%	3,01%	2,35%	1,021%	3,68%	12,11%	2,34%	0,81%
JOACABA	8,02%	6,21%	2,79%	5,12%	3,52%	1,715%	4,73%	3,23%	10,23%	4,28%
CHAPECO	8,22%	17,45%	6,79%	8,41%	12,20%	6,657%	9,15%	6,32%	10,21%	4,22%
CURITIBANOS	14,23%	16,58%	2,76%	3,09%	3,32%	2,726%	4,05%	2,84%	5,39%	0,86%
LAGES	9,42%	6,29%	2,41%	3,07%	3,14%	1,024%	4,29%	3,26%	3,49%	1,04%
TUBARAO	7,01%	8,95%	4,48%	2,59%	4,10%	3,374%	5,31%	3,48%	3,36%	4,83%
CRICIUMA	7,62%	15,94%	6,83%	5,15%	6,05%	2,247%	4,86%	4,38%	4,91%	1,67%
S. MIG. OESTE	2,20%	1,43%	1,53%	1,17%	1,97%	1,029%	2,51%	1,34%	2,32%	0,55%
MAFRA	6,41%	4,19%	4,75%	3,69%	3,00%	0,730%	3,42%	2,45%	4,40%	0,91%
ARARANGUA	2,81%	2,67%	2,31%	1,08%	2,67%	0,863%	2,14%	1,76%	1,19%	0,40%
	100%		100%		100%		100%		100%	

APÊNDICE E – RESULTADOS DA APLICAÇÃO E OBTENÇÃO DO CONJUNTO DE EQUAÇÕES

Este apêndice apresenta os resultados de todas as execuções realizadas para elaboração do conjunto de equações. Consiste de quatro tabelas para cada execução, onde são informados os resultados obtidos na análise

1. Combinações das 14 variáveis com quantidade de casos respondidos – relação gerada pelo programa PGM-02, lista todas as combinações possíveis a partir da base de dados.
2. Composição dos grupos de variáveis – relação classificada por grupos de variáveis informa quantas combinações existem para cada grupo e também quantas combinações obtiveram fator igual ou superior a 15 casos por variável dependente.
3. Equações selecionadas em cada combinação – informa todas as combinações que foram processadas para obtenção da equação de regressão logística e o resultado da melhor iteração no processo de redução de variáveis. Esta tabela destaca também, qual equação foi selecionada pelo critério de maior percentual de acerto médio.
4. Variáveis na equação - resultados da equação de regressão logística, após as reduções. Valores obtidos com a utilização do SPSS.

Na primeira tabela observa-se:

N – número de identificação da combinação, gerado pelo PGM-02

Combinação – seqüência de variáveis em cada combinação, a codificação referente a cada variável encontra-se na coluna VAR da tabela 29.

Qt. Casos – quantidade de declarações apresentadas pelos contribuintes que informaram valores a todas as variáveis da combinação.

Qt. Variável – quantidade de variáveis na combinação.

Fator – indica a quantidade de casos por variável dependente.

Na segunda tabela observa-se:

Grupo – identificação com a quantidade de variáveis em cada grupo analisado.

Qt. Comb. – quantidade de combinações possíveis com as variáveis de cada grupo.

Casos >0 – quantidade de combinações que possuem declarações dos contribuintes que informaram valores a todas as variáveis da combinação.

Fator =>15 – quantidade de combinações que possuem declarações dos contribuintes que informaram valores a todas as variáveis da combinação em quantidade suficiente de casos que atenda o fator de 15 casos por variável dependente.

Na terceira tabela observa-se:

N – número de identificação da combinação, gerado pelo PGM-02

Variáveis na combinação – seqüência de variáveis em cada combinação,

Casos – quantidade de declarações apresentadas pelos contribuintes que informaram valores a todas as variáveis da combinação.

Int. – quantidade de iterações que foram realizadas no processo de redução de variáveis.

Variáveis após redução – indica seqüência de variáveis em cada combinação após processo de redução.

Casos – quantidade de declarações apresentadas pelos contribuintes que informaram valores a todas as variáveis da combinação resultante ao final do processo de redução de variáveis.

% – valor do percentual médio de acerto obtido com a equação selecionada em cada combinação.

A codificação referente a cada variável das colunas dois e quatro desta tabela encontra-se na coluna VAR da tabela 29.

Na quarta tabela observa-se na primeira coluna a identificação das variáveis e na seqüência:

B – coeficiente de regressão logística da variável incluída no modelo.

S.E. – erro padrão da variável associado à variável,

Wald – valor da estatística de Wald.

df. – graus de liberdade.

Sig. – significância estatística.

Exp(B) – razão de chance de cada variável.

Na Tabela 29 são descritas as variáveis utilizadas neste apêndice.

Tabela 29 – Descrição das variáveis

Var.	Cod. Var.	Descrição
01	v3060_L	Resumo - Saídas - valor contábil
02	v3010_L	Resumo - Entradas - valor contábil
03	v9040_L	Saldo - Total de débitos
04	v3050_L	Resumo - Entradas - outras operações sem crédito de imposto
05	v9999_L	Saldo - imposto a recolher
06	v3100_L	Resumo - Saídas - outras operações sem débito de imposto
07	v3070_L	Resumo - Saídas - base de calculo
08	v3080_L	Resumo - Saídas - imposto debitado
09	v9080_L	Saldo - total créditos
10	v3030_L	Resumo - Entradas - imposto creditado
11	v3020_L	Resumo - Entradas - base de cálculo
12	v3090_L	Resumo - Saídas - operações isentas ou não tributadas
13	v3040_L	Resumo - Entradas - operações isentas ou não tributadas
14	v9998_L	Saldo - saldo credor para mês seguinte

Além destas variáveis, na tabela 4 é utilizada a codificação CL_24 que indica o código do grupo que pertence a empresa a ser processada. Esta variável categórica participa nos modelos de regressão logística como variáveis *dummy*, assumindo valores 0 e 1 para cada um dos 24 grupos.

1ª. EXECUÇÃO - (6.032 casos na base)

Tabela 30 – Combinações das 14 variáveis com quantidade de casos respondidos

N	Combinação	Qt. Casos	Qt. Variável	Fator
1	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	14	0
2	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13	556	13	15,444
3	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 14	0	13	0
5	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 13 - 14	0	13	0
9	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 12 - 13 - 14	0	13	0
17	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 11 - 12 - 13 - 14	0	13	0
33	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 10 - 11 - 12 - 13 - 14	0	13	0
65	01 - 02 - 03 - 04 - 05 - 06 - 07 - 09 - 10 - 11 - 12 - 13 - 14	0	13	0
129	01 - 02 - 03 - 04 - 05 - 06 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	13	0
257	01 - 02 - 03 - 04 - 05 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	13	0
513	01 - 02 - 03 - 04 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	338	13	9,389
	...	...	...	...
8.193	02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	13	0
	. . .	. . .	. . .	. . .
16.383	14	1.691	1	70,458

Tabela 31 – Composição dos grupos de variáveis

Grupo	Qt. comb	Casos >0	Fator => 15
14 variáveis	1	0	0
13 variáveis	14	2	1
12 variáveis	91	-	-
11 variáveis	364	-	-
10 variáveis	1.001	-	-
09 variáveis	2.002	-	-
08 variáveis	3.003	-	-
07 variáveis	3.432	-	-
06 variáveis	3.003	-	-
05 variáveis	2.002	-	-
04 variáveis	1.001	-	-
03 variáveis	364	-	-
02 variáveis	91	-	-
01 variável	14	-	-

Tabela 32 – Equações selecionadas em cada combinação

N	Variáveis na combinação	Casos	Int	Variáveis após redução	Casos	%
2	01-02-03-04-05-06-07-08-09-10-11-12-13	556	4	01-02-03-04-05-07-09-10-11-12	820	76,4

Tabela 33 – Variáveis na equação – 1ª. Execução

	B	S.E.	Wald	df	Sig.	Exp(B)
cl_24			61,86398	23	0,00002	
cl_24(1)	1,098372	1,183054	0,8619666	1	0,35319	2,99928
cl_24(2)	0,157414	0,821603	0,0367083	1	0,84806	1,17048
cl_24(3)	-0,04234	0,632839	0,0044757	1	0,94666	0,95855
cl_24(4)	0,20573	0,531112	0,1500463	1	0,69849	1,22842
cl_24(5)	1,084608	0,503289	4,6441917	1	0,03116	2,95828
cl_24(6)	1,473407	0,882685	2,7863431	1	0,09507	4,36408
cl_24(7)	-0,09831	0,495063	0,0394328	1	0,84259	0,90637
cl_24(8)	0,795488	0,709456	1,2572361	1	0,26217	2,21552
cl_24(9)	0,372422	0,56418	0,4357468	1	0,50918	1,45124
cl_24(10)	1,396572	0,696859	4,0164019	1	0,04506	4,04132
cl_24(11)	38,15037	78258655	2,376E-13	1	1,00000	3,7E+16
cl_24(12)	1,071442	0,760093	1,9870257	1	0,15865	2,91959
cl_24(13)	0,544122	0,614176	0,7848855	1	0,37565	1,72309
cl_24(14)	1,678215	0,678312	6,121197	1	0,01336	5,35599
cl_24(15)	-0,72501	0,537413	1,8199824	1	0,17732	0,48432
cl_24(16)	0,616833	0,614024	1,0091697	1	0,31510	1,85305
cl_24(17)	-0,42294	0,580714	0,5304296	1	0,46643	0,65512
cl_24(18)	-0,31458	0,788736	0,1590698	1	0,69001	0,7301
cl_24(19)	1,698723	0,756749	5,0389621	1	0,02478	5,46696
cl_24(20)	0,337259	0,553145	0,3717481	1	0,54205	1,4011
cl_24(21)	0,839779	0,557234	2,2712012	1	0,13180	2,31586
cl_24(22)	1,774724	0,673476	6,9441197	1	0,00841	5,89865
cl_24(23)	1,345013	0,49435	7,4025937	1	0,00651	3,83824
v3060_L	0,019118	0,508962	0,001411	1	0,97004	1,0193
v3010_L	-0,12117	0,490763	0,0609628	1	0,80498	0,88588
v9040_L	0,259944	0,783802	0,1099887	1	0,74016	1,29686
v3050_L	-0,19415	0,138502	1,9650307	1	0,16098	0,82353
v9999_L	0,084626	0,279736	0,0915189	1	0,76225	1,08831
v3070_L	-0,32364	0,506853	0,4077184	1	0,52313	0,72351
v9080_L	0,653059	0,677366	0,9295184	1	0,33499	1,92141
v3030_L	-1,69839	0,826517	4,2225126	1	0,03989	0,18298
v3020_L	1,530721	0,732035	4,3724905	1	0,03652	4,62151
v3090_L	-0,10072	0,086097	1,3684877	1	0,24207	0,90419
Constant	-0,85557	1,338303	0,408697	1	0,52263	0,42504

2ª. EXECUÇÃO - (5.203 casos na base)

Tabela 34 – Combinações das 14 variáveis com quantidade de casos respondidos

N	Combinação	Qt. Casos	Qt. Variável	Fator
1	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	14	0
2	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13	0	13	0
8	01 - 02 - 03 - 04 – 05 - 06 - 07 - 08 - 09 - 10 - 11	673	11	20,029
516	01 - 02 - 03 - 04 - 06 - 07 - 08 - 09 - 10 - 11 – 12	548	11	16,117
	...	...	...	...
518	01 - 02 - 03 - 04 - 06 - 07 - 08 - 09 - 10 - 11 - 13	770	11	22.647
519	01 - 02 - 03 - 04 - 06 - 07 - 08 - 09 - 10 - 11 - 14	921	11	27,088
	...	...	...	...
771	01 – 02 - 03 - 04 - 07 - 08 - 09 - 10 - 11 - 12 - 14	552	11	16,235
	...	...	...	...
773	01 – 02 - 03 - 04 - 07 - 08 - 09 - 10 - 11 - 13 - 14	564	11	16,588
	...	...	...	...
1.793	01 - 02 - 03 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	510	11	15,000
	. . .	. . .	. . .	. . .
16.383	14	1.691	1	70,458

Tabela 35 – Composição dos grupos de variáveis

Grupo	Qt. comb	Casos >0	Fator => 15
14 variáveis	1	0	0
13 variáveis	14	1	0
12 variáveis	91	16	0
11 variáveis	364	112	7
10 variáveis	1.001	-	-
09 variáveis	2.002	-	-
08 variáveis	3.003	-	-
07 variáveis	3.432	-	-
06 variáveis	3.003	-	-
05 variáveis	2.002	-	-
04 variáveis	1.001	-	-
03 variáveis	364	-	-
02 variáveis	91	-	-
01 variável	14	-	-

Tabela 36 – Equações selecionadas em cada combinação

N	Variáveis na combinação	Casos	Int	Variáveis após redução	Casos	%
8	01-02-03-04-05-06-07-08-09-10-11	673	1	01-02-03-04-06-07-08-09-10-11-12	673	80,1
516	01-02-03-04-06-07-08-09-10-11-12	548	1	01-02-03-04-06-07-08-09-10-11-12	548	73,1
518	01-02-03-04-06-07-08-09-10-11-13	770	3	01-02-03-06-07-08-09-11-13	837	73,0
519	01-02-03-04-06-07-08-09-10-11-14	921	3	01-02-04-06-07-08-10-11-14	922	73,9
771	01-02-03-04-07-08-09-10-11-12-14	552	7	01-04-10-11-14	1.145	72,3
773	01-02-03-04-07-08-09-10-11-13-14	564	1	01-02-03-04-07-08-09-10-11-13-14	564	68,5
1.793	01-02-03-07-08-09-10-11-12-13-14	510	3	01-02-03-08-10-11-12-13-14	511	68,0

Tabela 37 – Variáveis na equação – 2ª. Execução

	B	S.E.	Wald	df	Sig.	Exp(B)
cl_24			25,88706	23	0,30623	
cl_24(1)	0,177624	1,309406	0,018401	1	0,89210	1,1944
cl_24(2)	-50,9383	8,96E+10	3,23E-19	1	1,00000	8E-23
cl_24(3)	1,348943	1,377688	0,958707	1	0,32751	3,8534
cl_24(4)	1,403817	0,99159	2,004272	1	0,15686	4,0707
cl_24(5)	1,186224	0,753242	2,480071	1	0,11530	3,2747
cl_24(6)	0,482099	0,961507	0,251401	1	0,61609	1,6195
cl_24(7)	1,101705	0,766774	2,064411	1	0,15077	3,0093
cl_24(8)	0,85776	0,837955	1,047828	1	0,30601	2,3579
cl_24(9)	-0,14508	0,947748	0,023435	1	0,87833	0,8649
cl_24(10)	0,621273	0,804733	0,596019	1	0,44010	1,8613
cl_24(11)	38,00549	83119750	2,09E-13	1	1,00000	3E+16
cl_24(12)	1,58171	0,937499	2,846506	1	0,09157	4,8633
cl_24(13)	0,466666	0,760448	0,376594	1	0,53943	1,5947
cl_24(14)	1,2702	0,864793	2,157346	1	0,14189	3,5616
cl_24(15)	0,073244	0,886375	0,006828	1	0,93414	1,076
cl_24(16)	0,393435	0,793557	0,245804	1	0,62005	1,4821
cl_24(17)	0,390402	0,949891	0,168918	1	0,68108	1,4776
cl_24(18)	1,162262	0,909469	1,633173	1	0,20126	3,1972
cl_24(19)	1,150842	0,923157	1,554102	1	0,21253	3,1609
cl_24(20)	0,792585	0,782884	1,024936	1	0,31135	2,2091
cl_24(21)	0,191005	0,732043	0,06808	1	0,79415	1,2105
cl_24(22)	3,011283	0,993743	9,182372	1	0,00244	20,313
cl_24(23)	0,817727	0,724789	1,272897	1	0,25922	2,2653
v3060_L	-1,73843	0,622867	7,78979	1	0,00525	0,1758
v3010_L	-0,34055	0,57783	0,347341	1	0,55562	0,7114
v9040_L	0,834124	0,934994	0,795873	1	0,37233	2,3028
v3050_L	0,165212	0,198829	0,690435	1	0,40602	1,1796
v9999_L	0,417983	0,332092	1,584168	1	0,20816	1,5189
v3100_L	0,07987	0,13899	0,330219	1	0,56553	1,0831
v3070_L	-0,72781	1,245254	0,3416	1	0,55891	0,483
v3080_L	1,41116	1,251027	1,272386	1	0,25932	4,1007
v9080_L	-0,34281	0,750861	0,208443	1	0,64799	0,7098
v3030_L	-0,75212	0,895037	0,706145	1	0,40073	0,4714
v3020_L	1,015992	0,766285	1,757927	1	0,18488	2,7621
Constant	2,471233	1,703772	2,103801	1	0,14693	11,837

3ª. EXECUÇÃO - (4.539 casos na base)

Tabela 38 – Combinações das 14 variáveis com quantidade de casos respondidos

N	Combinação	Qt. Casos	Qt. Variável	Fator
1	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	14	0
2	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13	0	13	0
	...	...	...	...
516	01 - 02 - 03 - 04 - 06 - 07 - 08 - 09 - 10 - 11 - 12	548	11	16,117
	...	...	...	...
518	01 - 02 - 03 - 04 - 06 - 07 - 08 - 09 - 10 - 11 - 13	568	11	16.705
519	01 - 02 - 03 - 04 - 06 - 07 - 08 - 09 - 10 - 11 - 14	921	11	27,088
	...	...	...	...
771	01 – 02 - 03 - 04 - 07 - 08 - 09 - 10 - 11 - 12 - 14	552	11	16,235
	...	...	...	...
773	01 – 02 - 03 - 04 - 07 - 08 - 09 - 10 - 11 - 13 - 14	564	11	16,588
	...	...	...	...
1.793	01 - 02 - 03 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	510	11	15,000
	. . .	. . .	. . .	. . .
16.383	14	1691	1	70,458

Tabela 39 – Composição dos grupos de variáveis

Grupo	Qt. comb	Casos >0	Fator => 15
14 variáveis	1	0	0
13 variáveis	14	1	0
12 variáveis	91	15	0
11 variáveis	364	105	6
10 variáveis	1.001	-	-
09 variáveis	2.002	-	-
08 variáveis	3.003	-	-
07 variáveis	3.432	-	-
06 variáveis	3.003	-	-
05 variáveis	2.002	-	-
04 variáveis	1.001	-	-
03 variáveis	364	-	-
02 variáveis	91	-	-
01 variável	14	-	-

Tabela 40 – Equações selecionadas em cada combinação

N	Variáveis na combinação	Casos	Int	Variáveis após redução	Casos	%
516	01-02-03-04-06-07-08-09-10-11-12	548	1	01-02-03-04-06-07-08-09-10-11-12	548	73,1
518	01-02-03-04-06-07-08-09-10-11-13	568	5	01-02-06-08-09-10- 13	640	72,8
519	01-02-03-04-06-07-08-09-10-11-14	921	3	01-02-04-06-07-08-10-11-14	922	73,9
771	01-02-03-04-07-08-09-10-11-12-14	552	6	01-03-04-10-11-14	1.091	71,3
773	01-02-03-04-07-08-09-10-11-13-14	564	1	01-02-03-04-07-08-09-10-11-13-14	564	68,5
1.793	01-02-03-07-08-09-10-11-12-13-14	510	3	01-02-03-08-10-11-12-13-14	511	68,0

Tabela 41 – Variáveis na equação – 3ª. Execução

	B	S.E.	Wald	df	Sig.	Exp(B)
cl_24			79,55323	23	3,75E-08	
cl_24(1)	1,629906	0,989096	2,715494	1	0,09938	5,1034
cl_24(2)	1,676615	0,690207	5,900753	1	0,01513	5,3474
cl_24(3)	1,38667	0,727375	3,634369	1	0,05660	4,0015
cl_24(4)	1,242873	0,574783	4,67568	1	0,03059	3,4656
cl_24(5)	2,123141	0,555906	14,58665	1	0,00013	8,3573
cl_24(6)	0,618148	0,592195	1,089573	1	0,29657	1,8555
cl_24(7)	0,398767	0,543326	0,538664	1	0,46299	1,4900
cl_24(8)	1,491634	0,629368	5,617141	1	0,01779	4,4444
cl_24(9)	0,496827	0,556349	0,797471	1	0,37185	1,6435
cl_24(10)	2,915222	0,667265	19,08741	1	0,00001	18,4529
cl_24(11)	1,28366	0,870764	2,173195	1	0,14043	3,6098
cl_24(12)	2,054191	0,731904	7,877222	1	0,00501	7,8005
cl_24(13)	1,914108	0,599494	10,19444	1	0,00141	6,7809
cl_24(14)	1,229661	0,656945	3,503584	1	0,06124	3,4201
cl_24(15)	-0,02112	0,634084	0,001109	1	0,97343	0,9791
cl_24(16)	1,581847	0,519705	9,264373	1	0,00234	4,8639
cl_24(17)	0,47747	0,755151	0,399783	1	0,52720	1,6120
cl_24(18)	-0,77041	0,881882	0,76317	1	0,38234	0,4628
cl_24(19)	2,185553	0,591749	13,64102	1	0,00022	8,8956
cl_24(20)	1,843506	0,538442	11,72226	1	0,00062	6,3186
cl_24(21)	1,358208	0,532995	6,493612	1	0,01083	3,8892
cl_24(22)	2,4821	0,584236	18,04937	1	0,00002	11,9664
cl_24(23)	2,199458	0,516221	18,15346	1	0,00002	9,0201
v3060_L	-0,97478	0,388111	6,30811	1	0,01202	0,3773
v3010_L	0,417277	0,390633	1,141066	1	0,28543	1,5178
v3050_L	-0,24003	0,148105	2,626656	1	0,10508	0,7866
v3100_L	0,094045	0,116633	0,65017	1	0,42005	1,0986
v3070_L	0,407863	0,513415	0,631092	1	0,42696	1,5036
v3080_L	-0,20494	0,535777	0,146315	1	0,70208	0,8147
v3030_L	-0,93545	0,602557	2,410156	1	0,12055	0,3924
v3020_L	1,571167	0,571809	7,549937	1	0,00600	4,8123
v9998_L	-0,21568	0,101722	4,495651	1	0,03398	0,8060
Constant	-1,24817	1,079907	1,335899	1	0,24776	0,2870

4ª. EXECUÇÃO - (3.617 casos na base)

Tabela 42 – Combinações das 14 variáveis com quantidade de casos respondidos

N	Combinação	Qt. Casos	Qt. Variável	Fator
1	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	14	0
2	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13	0	13	0
	...	...	...	...
1.798	01 - 02 - 03 - 07 - 08 - 09 - 10 - 11 - 13	544	9	16,750
	. . .	. . .	. . .	. . .
16.383	14	769	1	32,041

Tabela 43 – Composição dos grupos de variáveis

Grupo	Qt. comb	Casos >0	Fator => 15
14 variáveis	1	0	0
13 variáveis	14	0	0
12 variáveis	91	6	0
11 variáveis	364	71	0
10 variáveis	1.001	356	0
09 variáveis	2.002	1.029	1
08 variáveis	3.003	-	-
07 variáveis	3.432	-	-
06 variáveis	3.003	-	-
05 variáveis	2.002	-	-
04 variáveis	1.001	-	-
03 variáveis	364	-	-
02 variáveis	91	-	-
01 variável	14	-	-

Tabela 44 – Equações selecionadas em cada combinação

N	Variáveis na combinação	Casos	Int	Variáveis após redução	Casos	%
1.798	01-02-03-07-08-09-10-11-13	544	4	02-07-08-09-11-13	544	72,7

Tabela 45 – Variáveis na equação – 4ª. Execução

	B	S.E.	Wald	df	Sig.	Exp(B)
cl_24			38,57414	23	0,02208	
cl_24(1)	22,07107	28368,79	6,05E-07	1	0,99938	4E+09
cl_24(2)	2,78781	0,967325	8,3058	1	0,00395	16,25
cl_24(3)	1,674932	0,740351	5,118228	1	0,02368	5,338
cl_24(4)	0,606016	0,608371	0,992275	1	0,31919	1,833
cl_24(5)	0,409494	0,510396	0,643694	1	0,42238	1,506
cl_24(6)	-0,30904	0,708613	0,190205	1	0,66275	0,734
cl_24(7)	-0,40775	0,704035	0,335427	1	0,56248	0,665
cl_24(8)	-1,64202	0,709528	5,355715	1	0,02065	0,194
cl_24(9)	-0,48588	0,673324	0,520734	1	0,47053	0,615
cl_24(10)	1,734744	0,856872	4,098629	1	0,04292	5,667
cl_24(11)	0,028605	0,777636	0,001353	1	0,97066	1,029
cl_24(12)	1,626638	1,138887	2,039954	1	0,15321	5,087
cl_24(13)	0,307786	0,576289	0,285244	1	0,59328	1,36
cl_24(14)	0,045375	1,161298	0,001527	1	0,96883	1,046
cl_24(15)	-0,14628	0,578537	0,063927	1	0,80039	0,864
cl_24(16)	1,051383	0,57607	3,330971	1	0,06799	2,862
cl_24(17)	-0,07703	0,631807	0,014864	1	0,90296	0,926
cl_24(18)	-1,90711	1,159795	2,703879	1	0,10010	0,149
cl_24(19)	0,920421	0,545841	2,843421	1	0,09175	2,51
cl_24(20)	0,422851	0,478966	0,779409	1	0,37732	1,526
cl_24(21)	0,124737	0,503873	0,061285	1	0,80448	1,133
cl_24(22)	21,4235	10569,51	4,11E-06	1	0,99838	2E+09
cl_24(23)	0,436695	0,603077	0,524338	1	0,46900	1,548
v3010_L	-1,13199	0,274099	17,05585	1	0,00004	0,322
v3070_L	0,2379	0,497316	0,228836	1	0,63239	1,269
v3080_L	-0,56623	0,53751	1,109724	1	0,29214	0,568
v9080_L	-0,57299	0,202155	8,033775	1	0,00459	0,564
v3020_L	1,52647	0,331262	21,2341	1	0,00000	4,602
v3040_L	0,133788	0,112654	1,410383	1	0,23499	1,143
Constant	1,844656	1,037485	3,161307	1	0,07540	6,326

5ª. EXECUÇÃO - (3.073 casos na base)

Tabela 46 – Combinações das 14 variáveis com quantidade de casos respondidos

N	Combinação	Qt. Casos	Qt. Variável	Fator
1	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	14	0
2	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13	0	13	0
	...	...	...	...
64	01 – 02 - 03 - 04 – 05 – 06 - 07 – 08	524	8	17,032
	...	...	...	...
16.383	14	561	1	23,375

Tabela 47 – Composição dos grupos de variáveis

Grupo	Qt. comb	Casos >0	Fator => 15
14 variáveis	1	0	0
13 variáveis	14	0	0
12 variáveis	91	2	0
11 variáveis	364	36	0
10 variáveis	1.001	229	0
09 variáveis	2.002	781	0
08 variáveis	3.003	1.665	1
07 variáveis	3.432	-	-
06 variáveis	3.003	-	-
05 variáveis	2.002	-	-
04 variáveis	1.001	-	-
03 variáveis	364	-	-
02 variáveis	91	-	-
01 variável	14	-	-

Tabela 48 – Equações selecionadas em cada combinação

N	Variáveis na combinação	Casos	Int	Variáveis após redução	Casos	%
64	01-02-03-04-05-06-07-08	524	1	01-02-03-04-05-06-07-08	524	79,3

Tabela 49 – Variáveis na equação – 5ª. Execução

	B	S.E.	Wald	df	Sig.	Exp(B)
cl_24			43,48972	20	0,00176	
cl_24(1)	-46,6123	1,31E+11	1,26E-19	1	1,00000	6E-21
cl_24(2)	1,098842	0,856953	1,644209	1	0,19975	3,001
cl_24(3)	0,626014	1,104946	0,320986	1	0,57102	1,87
cl_24(4)	0,5168	0,818761	0,39841	1	0,52791	1,677
cl_24(5)	0,841787	0,897126	0,880436	1	0,34808	2,321
cl_24(6)	0,885005	1,063764	0,692152	1	0,40543	2,423
cl_24(7)	39,97144	1E+08	1,59E-13	1	1,00000	2E+17
cl_24(8)	1,321197	1,032447	1,637569	1	0,20066	3,748
cl_24(9)	2,029002	1,103552	3,380486	1	0,06597	7,606
cl_24(10)	1,848608	0,890797	4,30657	1	0,03797	6,351
cl_24(11)	2,86689	1,004906	8,139	1	0,00433	17,58
cl_24(12)	-1,65288	1,326302	1,553089	1	0,21268	0,191
cl_24(13)	1,301278	0,92987	1,958376	1	0,16169	3,674
cl_24(14)	1,065535	0,917972	1,347336	1	0,24574	2,902
cl_24(15)	-0,01199	1,063131	0,000127	1	0,99100	0,988
cl_24(16)	1,003588	1,057624	0,900426	1	0,34267	2,728
cl_24(17)	1,64262	1,026502	2,560675	1	0,10955	5,169
cl_24(18)	1,293641	0,91007	2,020588	1	0,15518	3,646
cl_24(19)	2,702222	0,919304	8,640194	1	0,00329	14,91
cl_24(20)	1,851692	0,849052	4,756297	1	0,02919	6,371
v3060_L	-1,49492	0,483592	9,556003	1	0,00199	0,224
v3010_L	0,464576	0,496227	0,876503	1	0,34916	1,591
v9040_L	-0,58954	0,990901	0,353975	1	0,55187	0,555
v3050_L	-0,60029	0,432313	1,928108	1	0,16497	0,549
v9999_L	1,408616	0,994112	2,007773	1	0,15649	4,09
v3100_L	0,358063	0,147922	5,859366	1	0,01549	1,431
v3070_L	2,284047	1,281548	3,176438	1	0,07471	9,816
v3080_L	-2,01162	1,269363	2,511412	1	0,11302	0,134
Constant	0,829153	1,804201	0,211203	1	0,64583	2,291

6ª. EXECUÇÃO - (2.549 casos na base)

Tabela 50 – Combinações das 14 variáveis com quantidade de casos respondidos

N	Combinação	Qt. Casos	Qt. Variável	Fator
1	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	14	0
2	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13	0	13	0
	...	...	...	...
256	01 - 02 - 03 - 04 – 05 – 06	761	6	26,241
	...	...	...	...
832	01 - 02 - 03 - 04 - 07 – 08	489	6	16.862
	...	...	...	...
1.344	01 - 02 - 03 – 05 - 07 - 08	630	6	15,551
	. . .	. . .	. . .	. . .
16.383	14	561	1	23,375

Tabela 51 – Composição dos grupos de variáveis

Grupo	Qt. comb	Casos >0	Fator => 15
14 variáveis	1	0	0
13 variáveis	14	0	0
12 variáveis	91	1	0
11 variáveis	364	22	0
10 variáveis	1.001	162	0
09 variáveis	2.002	621	0
08 variáveis	3.003	1.452	0
07 variáveis	3.432	2.235	0
06 variáveis	3.003	2.362	3
05 variáveis	2.002	-	-
04 variáveis	1.001	-	-
03 variáveis	364	-	-
02 variáveis	91	-	-
01 variável	14	-	-

Tabela 52 – Equações selecionadas em cada combinação

N	Variáveis na combinação	Casos	Int	Variáveis após redução	Casos	%
256	01-02-03-04-05-06	761	3	02-04-05-06	761	76,3
832	01-02-03-04-07-08	489	3	03-04-07-08	489	74,2
1.344	01-02-03-05-07-08	630	3	01-05-07-08	630	78,8

Tabela 53 – Variáveis na equação – 6ª. Execução

	B	S.E.	Wald	df	Sig.	Exp(B)
cl_24			40,94946	21	0,00569	
cl_24(1)	-21,2918	40192,97	2,81E-07	1	0,99958	6E-10
cl_24(2)	21,27206	40192,97	2,8E-07	1	0,99958	2E+09
cl_24(3)	3,168102	1,156997	7,497808	1	0,00618	23,76
cl_24(4)	1,860316	1,158581	2,578226	1	0,10834	6,426
cl_24(5)	0,835941	1,008764	0,686707	1	0,40729	2,307
cl_24(6)	0,869231	0,902893	0,926826	1	0,33569	2,385
cl_24(7)	0,26398	1,281762	0,042416	1	0,83683	1,302
cl_24(8)	1,957822	1,131396	2,994449	1	0,08355	7,084
cl_24(9)	-0,20616	0,927216	0,049436	1	0,82405	0,814
cl_24(10)	0,864017	0,943159	0,839219	1	0,35962	2,373
cl_24(11)	0,68015	0,908634	0,560313	1	0,45413	1,974
cl_24(12)	21,22931	10315,59	4,24E-06	1	0,99836	2E+09
cl_24(13)	0,785107	1,00269	0,613089	1	0,43363	2,193
cl_24(14)	1,403118	1,064849	1,736251	1	0,18761	4,068
cl_24(15)	1,031393	1,366473	0,5697	1	0,45038	2,805
cl_24(16)	0,917394	1,23939	0,547893	1	0,45918	2,503
cl_24(17)	1,56034	1,012515	2,374849	1	0,12330	4,76
cl_24(18)	2,021136	0,998227	4,099517	1	0,04290	7,547
cl_24(19)	0,751962	0,913771	0,6772	1	0,41055	2,121
cl_24(20)	3,529719	1,320808	7,141686	1	0,00753	34,11
cl_24(21)	2,822978	1,113982	6,421823	1	0,01127	16,83
v3060_L	-0,62053	0,295323	4,415051	1	0,03562	0,538
v9999_L	0,154841	0,26984	0,329277	1	0,56609	1,167
v3070_L	-0,70402	0,778139	0,818562	1	0,36560	0,495
v3080_L	1,292599	0,767261	2,83819	1	0,09205	3,642
Constant	0,819509	1,601631	0,261807	1	0,60888	2,269

7ª. EXECUÇÃO - (1.919 casos na base)

Tabela 54 – Combinações das 14 variáveis com quantidade de casos respondidos

N	Combinação	Qt. Casos	Qt. Variável	Fator
1	01 - 02 - 03 - 04 - 05 - 06 - 07 – 08 - 09 - 10 - 11 - 12 - 13 - 14	0	14	0
2	01 - 02 - 03 - 04 - 05 - 06 - 07 – 08 - 09 - 10 - 11 - 12 - 13	0	13	0
	...	...	...	...
256	01 - 02 - 03 - 04 - 05 – 06	760	6	26,241
	...	...	...	...
16.383	14	561	1	23,375

Tabela 55 – Composição dos grupos de variáveis

Grupo	Qt. comb	Casos >0	Fator => 15
14 variáveis	1	0	0
13 variáveis	14	0	0
12 variáveis	91	1	0
11 variáveis	364	19	6
10 variáveis	1.001	130	0
09 variáveis	2.002	483	0
08 variáveis	3.003	1.135	0
07 variáveis	3.432	1.812	0
06 variáveis	3.003	2.030	1
05 variáveis	2.002	-	-
04 variáveis	1.001	-	-
03 variáveis	364	-	-
02 variáveis	91	-	-
01 variável	14	-	-

Tabela 56 – Equações selecionadas em cada combinação

N	Variáveis na combinação	Casos	Int	Variáveis após redução	Casos	%
256	01-02-03-04-05-06	760	3	02-04-05-06	760	76,3

Tabela 57 – Variáveis na equação – 7ª. Execução

	B	S.E.	Wald	df	Sig.	Exp(B)
cl_24			83,97288	21	1,7E-09	
cl_24(1)	35,71649	1,42E+08	6,32E-14	1	1,00000	3E+15
cl_24(2)	35,93053	1,42E+08	6,4E-14	1	1,00000	4E+15
cl_24(3)	-0,02014	0,696958	0,000835	1	0,97695	0,98
cl_24(4)	-0,57279	0,758666	0,570027	1	0,45025	0,564
cl_24(5)	-1,40938	0,694236	4,121346	1	0,04235	0,244
cl_24(6)	-1,4705	0,741362	3,934325	1	0,04731	0,23
cl_24(7)	-1,14256	0,743165	2,363667	1	0,12419	0,319
cl_24(8)	1,017953	0,993872	1,049046	1	0,30573	2,768
cl_24(9)	-1,80051	0,799215	5,075341	1	0,02427	0,165
cl_24(10)	0,034871	0,751072	0,002156	1	0,96297	1,035
cl_24(11)	-0,20929	0,7569	0,076461	1	0,78215	0,811
cl_24(12)	0,586939	0,995186	0,347838	1	0,55534	1,798
cl_24(13)	-1,35519	0,759926	3,180211	1	0,07454	0,258
cl_24(14)	0,850938	0,988998	0,740297	1	0,38957	2,342
cl_24(15)	-1,0178	0,745138	1,865745	1	0,17196	0,361
cl_24(16)	-1,86654	0,749732	6,198129	1	0,01279	0,155
cl_24(17)	-0,85246	0,890538	0,916316	1	0,33844	0,426
cl_24(18)	-0,1959	1,029098	0,036237	1	0,84903	0,822
cl_24(19)	0,201864	0,758505	0,070827	1	0,79014	1,224
cl_24(20)	0,754685	0,840302	0,806603	1	0,36913	2,127
cl_24(21)	1,05349	0,888909	1,40458	1	0,23596	2,868
v3010_L	-0,64076	0,386064	2,754687	1	0,09697	0,527
v3050_L	0,226155	0,360984	0,392496	1	0,53099	1,254
v9999_L	0,470781	0,159783	8,681135	1	0,00322	1,601
v3100_L	0,086936	0,201841	0,185516	1	0,66668	1,091
Constant	1,746102	1,124384	2,411625	1	0,12044	5,732

8ª. EXECUÇÃO - (1.159 casos na base)

Tabela 58 – Combinações das 14 variáveis com quantidade de casos respondidos

N	Combinação	Qt. Casos	Qt. Variável	Fator
1	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	14	0
2	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13	0	13	0
	...	...	...	...
5.952	01 - 03 - 07 – 08	263	4	15,888
	. . .	. . .	. . .	. . .
16.383	14	561	1	23,375

Tabela 59 – Composição dos grupos de variáveis

Grupo	Qt. comb	Casos >0	Fator => 15
14 variáveis	1	0	0
13 variáveis	14	0	0
12 variáveis	91	1	0
11 variáveis	364	18	6
10 variáveis	1.001	119	0
09 variáveis	2.002	430	0
08 variáveis	3.003	994	0
07 variáveis	3.432	1.589	0
06 variáveis	3.003	1.818	0
05 variáveis	2.002	1.496	0
04 variáveis	1.001	871	1
03 variáveis	364	-	-
02 variáveis	91	-	-
01 variável	14	-	-

Tabela 60 – Equações selecionadas em cada combinação

N	Variáveis na combinação	Casos	Int	Variáveis após redução	Casos	%
5.952	01-03-07-08	263	1	01-03-07-08	263	73,0

Tabela 61 – Variáveis na equação – 8ª. Execução

	B	S.E.	Wald	df	Sig.	Exp(B)
cl_24			57,58269	22	5,1E-05	
cl_24(1)	2,942685	1,453895	4,096578	1	0,04297	18,97
cl_24(2)	1,449657	1,64482	0,776771	1	0,37813	4,262
cl_24(3)	2,428728	1,250531	3,771977	1	0,05212	11,34
cl_24(4)	2,971313	0,898614	10,93327	1	0,00094	19,52
cl_24(5)	0,721603	1,036602	0,484588	1	0,48635	2,058
cl_24(6)	0,419914	1,035057	0,164586	1	0,68497	1,522
cl_24(7)	2,436454	0,835865	8,496576	1	0,00356	11,43
cl_24(8)	3,035146	1,059462	8,207077	1	0,00417	20,8
cl_24(9)	23,97153	26904,3	7,94E-07	1	0,99929	3E+10
cl_24(10)	1,948215	0,88114	4,888589	1	0,02703	7,016
cl_24(11)	3,231269	1,013848	10,15783	1	0,00144	25,31
cl_24(12)	2,410793	0,944065	6,521034	1	0,01066	11,14
cl_24(13)	2,382717	1,074503	4,91733	1	0,02659	10,83
cl_24(14)	0,951503	0,995262	0,913998	1	0,33906	2,59
cl_24(15)	2,697178	0,966909	7,781227	1	0,00528	14,84
cl_24(16)	3,442836	1,055185	10,64573	1	0,00110	31,28
cl_24(17)	1,419743	1,000102	2,015258	1	0,15572	4,136
cl_24(18)	1,072981	0,889285	1,455803	1	0,22760	2,924
cl_24(19)	2,227564	1,014584	4,820409	1	0,02812	9,277
cl_24(20)	1,319508	0,954325	1,91175	1	0,16677	3,742
cl_24(21)	3,686901	1,093765	11,36253	1	0,00075	39,92
cl_24(22)	3,62175	1,090804	11,02411	1	0,00090	37,4
v3060_L	-0,76483	0,257938	8,792144	1	0,00303	0,465
v9040_L	-0,15632	0,323699	0,233211	1	0,62915	0,855
v3070_L	-0,57581	0,750608	0,588489	1	0,44300	0,562
v3080_L	1,510312	0,816324	3,423009	1	0,06429	4,528
Constant	-0,3922	1,328827	0,087112	1	0,76788	0,676

9ª. EXECUÇÃO - **(896 casos na base)**

Tabela 62 – Combinações das 14 variáveis com quantidade de casos respondidos

N	Combinação	Qt. Casos	Qt. Variável	Fator
1	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13 - 14	0	14	0
2	01 - 02 - 03 - 04 - 05 - 06 - 07 - 08 - 09 - 10 - 11 - 12 - 13	0	13	0
	...	...	...	...
4.096	01 – 02	562	2	22,440
	...	...	...	...
7.936	01 – 06	411	2	16.440
	...	...	...	...
11.264	02 - 04	409	2	16,360
	...	...	...	...
16.383	14	355	1	14,790

Tabela 63 – Composição dos grupos de variáveis

Grupo	Qt. comb	Casos >0	Fator => 15
14 variáveis	1	0	0
13 variáveis	14	0	0
12 variáveis	91	1	0
11 variáveis	364	12	6
10 variáveis	1.001	68	0
09 variáveis	2.002	241	0
08 variáveis	3.003	599	0
07 variáveis	3.432	1.086	0
06 variáveis	3.003	1.425	0
05 variáveis	2.002	1.316	0
04 variáveis	1.001	828	0
03 variáveis	364	343	0
02 variáveis	91	90	3
01 variável	14	-	-

Tabela 64 – Equações selecionadas em cada combinação

N	Variáveis na combinação	Casos	Int	Variáveis após redução	Casos	%
4.096	01-02	562	1	01-02	562	74,5
7.936	01-06	411	1	01-06	411	72,7
11.264	02-04	409	1	02-04	409	72,3

Tabela 65 – Variáveis na equação – 9ª. Execução

	B	S.E.	Wald	df	Sig.	Exp(B)
cl_24			111,6713	22	5,5E-14	
cl_24(1)	23,41943	40192,97	3,4E-07	1	0,99954	1E+10
cl_24(2)	-19,1342	40192,97	2,27E-07	1	0,99962	5E-09
cl_24(3)	1,806484	1,474177	1,50165	1	0,22042	6,089
cl_24(4)	1,912071	0,440274	18,86085	1	0,00001	6,767
cl_24(5)	0,373701	0,629637	0,352263	1	0,55284	1,453
cl_24(6)	0,789116	0,547646	2,076262	1	0,14961	2,201
cl_24(7)	0,359994	0,714686	0,253723	1	0,61447	1,433
cl_24(8)	1,262171	0,492198	6,575922	1	0,01034	3,533
cl_24(9)	4,637881	1,0902	18,09784	1	0,00002	103,3
cl_24(10)	2,430721	0,696927	12,16455	1	0,00049	11,37
cl_24(11)	3,194093	0,578543	30,48067	1	0,00000	24,39
cl_24(12)	2,815761	0,646328	18,97955	1	0,00001	16,71
cl_24(13)	2,88189	0,750939	14,72805	1	0,00012	17,85
cl_24(14)	0,765815	0,566012	1,830616	1	0,17605	2,151
cl_24(15)	2,734883	0,755323	13,11028	1	0,00029	15,41
cl_24(16)	0,888583	0,560262	2,515441	1	0,11274	2,432
cl_24(17)	-0,49887	0,722562	0,476686	1	0,48993	0,607
cl_24(18)	1,905522	0,55184	11,92343	1	0,00055	6,723
cl_24(19)	2,808537	0,685323	16,7946	1	0,00004	16,59
cl_24(20)	2,107911	0,546589	14,87244	1	0,00012	8,231
cl_24(21)	4,111455	0,824321	24,877	1	0,00000	61,04
cl_24(22)	2,78407	0,630991	19,46768	1	0,00001	16,18
v3060_L	0,267716	0,270566	0,979047	1	0,32243	1,307
v3010_L	-0,52398	0,245628	4,550722	1	0,03290	0,592
Constant	-0,35285	0,829903	0,180766	1	0,67072	0,703

APÊNDICE F – PROGRAMA PARA TRANSFORMAÇÃO VARIÁVEIS NO INTERVALO 0 – 1 (PGM-01)

O programa PGM-01 desenvolvido em *MS Visual Basic* calcula valores máximos e mínimos para cada série e realiza transformação dos dados para o intervalo 0 – 1.

```
********** Rotina para transformação dos valores para o intervalo 0 -1
*************************************************************
   Dim wMin_Mp(22) As Double
   Dim wMax_Mp(22) As Double
   Dim wMin_M(22) As Double
   Dim wMax_M(22) As Double

   With rs
      .CursorLocation = adUseClient
      .CursorType = adOpenStatic
      .LockType = adLockReadOnly
      sql = "select * from VARIAVEIS"
      .Open sql, ocn
      While Not .EOF
         wMin_Mp(!id_Variavel) = !vl_Min_Mp
         wMax_Mp(!id_Variavel) = !vl_Max_Mp
         wMin_M(!id_Variavel) = !vl_Min_M
         wMax_M(!id_Variavel) = !vl_Max_M
         .MoveNext
      Wend
      .Close
   End With

   Dim rsUpd As New ADODB.Recordset
   With rsUpd
      .CursorLocation = adUseClient
      .CursorType = adOpenStatic
      .LockType = adLockPessimistic
      sql = "select * from VALORES_T"
      .Open sql, ocn
      While Not .EOF
         If Not IsNull(!v_3060_Tmp) Then  !v_3060_Tmp = (!v_3060_Tmp –
                  wMin_Mp(1)) / (wMax_Mp(1) - wMin_Mp(1))
         If Not IsNull(!v_3010_Tmp) Then  !v_3010_Tmp = (!v_3010_Tmp -  wMin_Mp(2)) /
                  (wMax_Mp(2) - wMin_Mp(2))
         If Not IsNull(!v_3050_Tmp) Then  !v_3050_Tmp = (!v_3050_Tmp - wMin_Mp(3)) /
                  (wMax_Mp(3) - wMin_Mp(3))
         If Not IsNull(!v_3100_Tmp) Then  !v_3100_Tmp = (!v_3100_Tmp - wMin_Mp(4)) /
                  (wMax_Mp(4) - wMin_Mp(4))
         If Not IsNull(!v_6030_Tmp) Then   !v_6030_Tmp = (!v_6030_Tmp - wMin_Mp(5)) /
                  (wMax_Mp(5) - wMin_Mp(5))
         If Not IsNull(!v_6010_Tmp) Then !v_6010_Tmp = (!v_6010_Tmp - wMin_Mp(6)) /
                  (wMax_Mp(6) - wMin_Mp(6))
```

```
        If Not IsNull(!v_3020_Tmp) Then !v_3020_Tmp = (!v_3020_Tmp - wMin_Mp(7)) /
                (wMax_Mp(7) - wMin_Mp(7))
        If Not IsNull(!v_3090_Tmp) Then !v_3090_Tmp = (!v_3090_Tmp - wMin_Mp(8)) /
                (wMax_Mp(8) - wMin_Mp(8))
        If Not IsNull(!v_9080_Tmp) Then !v_9080_Tmp = (!v_9080_Tmp - wMin_Mp(9)) /
                (wMax_Mp(9) - wMin_Mp(9))
        If Not IsNull(!v_9050_Tmp) Then !v_9050_Tmp = (!v_9050_Tmp - wMin_Mp(10)) /
                (wMax_Mp(10) - wMin_Mp(10))
        If Not IsNull(!v_80030_Tmp) Then !v_80030_Tmp = (!v_80030_Tmp - wMin_Mp(11)) /
                (wMax_Mp(11) - wMin_Mp(11))
        If Not IsNull(!v_83310_Tmp) Then !v_83310_Tmp = (!v_83310_Tmp - wMin_Mp(12)) /
                (wMax_Mp(12) - wMin_Mp(12))
        If Not IsNull(!v_83320_Tmp) Then!v_83320_Tmp = (!v_83320_Tmp - wMin_Mp(13)) /
                (wMax_Mp(13) - wMin_Mp(13))
        If Not IsNull(!v_82299_Tmp) Then!v_82299_Tmp = (!v_82299_Tmp - wMin_Mp(14)) /
                (wMax_Mp(14) - wMin_Mp(14))
        If Not IsNull(!v_81199_Tmp) Then !v_81199_Tmp = (!v_81199_Tmp - wMin_Mp(15)) /
                (wMax_Mp(15) - wMin_Mp(15))
        If Not IsNull(!v_84499_Tmp) Then !v_84499_Tmp = (!v_84499_Tmp - wMin_Mp(16)) /
                (wMax_Mp(16) - wMin_Mp(16))
        If Not IsNull(!v_83323_Tmp) Then !v_83323_Tmp = (!v_83323_Tmp - wMin_Mp(17)) /
                (wMax_Mp(17) - wMin_Mp(17))
        If Not IsNull(!v_81110_Tmp) Then !v_81110_Tmp = (!v_81110_Tmp - wMin_Mp(18)) /
                (wMax_Mp(18) - wMin_Mp(18))
        If Not IsNull(!v_80020_Tmp) Then !v_80020_Tmp = (!v_80020_Tmp - wMin_Mp(19)) /
                (wMax_Mp(19) - wMin_Mp(19))
        If Not IsNull(!v_82270_Tmp) Then !v_82270_Tmp = (!v_82270_Tmp - wMin_Mp(20)) /
                (wMax_Mp(20) - wMin_Mp(20))
        If Not IsNull(!v_83330_Tmp) Then !v_83330_Tmp = (!v_83330_Tmp - wMin_Mp(21)) /
                (wMax_Mp(21) - wMin_Mp(21))
        If Not IsNull(!v_80010_Tmp) Then !v_80010_Tmp = (!v_80010_Tmp - wMin_Mp(22)) /
                (wMax_Mp(22) - wMin_Mp(22))
      '
      .Update
      n = n + 1
      '
      .MoveNext
    Wend
    .Close
End With

MousePointer = 0
MsgBox "Atualização ok - " & n & " registros alterados"
```

APÊNDICE G – PROGRAMA PARA GERAR AS COMBINAÇÕES DE VARIÁVEIS (PGM-02)

O programa PGM-02 desenvolvido em *MS Visual Basic* faz a leitura de todas as declarações mensais dos contribuintes da base de dados e gera uma tabela com todas as combinações possíveis de variáveis, neste caso com 14 variáveis formam-se 16.383 combinações. Para cada combinação é totalizada a quantidade de declarações que atendem a todas as variáveis d combinação.

```
**********  Rotina para gerar combinações de 14 variáveis
*************************************************************
   Private sub Gera_Combinacoes()
      grd.clear
      r = 0
      For a = 1 To 14
        For b = a + 1 To 14
          For c = b + 1 To 14
            For d = c + 1 To 14
              For e = d + 1 To 14
                For f = e + 1 To 14
                  For g = f + 1 To 14
                    For h = g + 1 To 14
                      For i = h + 1 To 14
                        For j = i + 1 To 14
                          For k = j + 1 To 14
                            For l = k + 1 To 14
                              For m = l + 1 To 14
                                For n = m + 1 To 14
                                    r = r + 1
                                    call Preenche_Matriz()
                                Next
                              Next
                            Next
                          Next
                        Next
                      Next
                    Next
                  Next
                Next
              Next
            Next
          Next
        Next
   End sub


Private sub Preenche_Matriz()
    Limpa_Campos
```

```
    wC(1) = a: wC(2) = b: wC(3) = c: wC(4) = d: wC(5) = e
    wC(6) = f: wC(7) = g: wC(8) = h: wC(9) = i: wC(10) = j
    wC(11) = k: wC(12) = l: wC(13) = m: wC(14) = n
    with grd
       .TextMatrix(r, 0) = r
       .TextMatrix(r, 1) = Format(a, "00 - ") & Format(b, "00 - ") & Format(c, "00 - ") _
            & Format(d, "00 - ") & Format(e, "00 - ") & Format(f, "00 - ") & Format(g, "00 - ") _
            & Format(h, "00 - ") & Format(i, "00 - ") & Format(j, "00 - ") & Format(k, "00 - ") _
            & Format(l, "00 - ") & Format(m, "00 - ") & Format(n, "00")
        Call Tot_Audit(wC())
        .TextMatrix(r, 2) = w_1
        .TextMatrix(r, 3) = w_2
    End with
End sub


Private Sub Tot_Audit(Campos() As Integer)
   w_1 = 0
   w_2 = 0
   wsql = ""
   For n = 1 To 14
      If Campos(n) = 0 Then Exit For
      '
      wsql = wsql & "v" & wTab(Campos(n)) & " > 0 and "
   Next
   If Right(wsql, 4) = "and " Then
      wsql = Left(wsql, Len(wsql) - 5)
   End If
   Dim rs As New ADODB.Recordset
   With rs
      .CursorLocation = adUseClient
      .CursorType = adOpenStatic
      .LockType = adLockReadOnly
      sql = "select count(*) as tot from [" & cboTabelas.List(cboTabelas.ListIndex) & "] "
      If wsql <> "" Then
         sql = sql & " where " & wsql
      End If
      .Open sql, ocn
      If Not .EOF Then
         w_1 = !tot
      End If
      .Close

      t = 0
      For n = 1 To Len(wsql)
         If Mid(wsql, n, 1) = ">" Then
            t = t + 1
         End If
      Next
      w_2 = t

   End With
End Sub
```

APÊNDICE H – PROGRAMA PARA PROCESSAR DADOS MENSAIS (PGM-03)

O programa PGM-03 desenvolvido em *MS Visual Basic* permite cadastrar as equações geradas no SPSS e processa os dados das bases de auditoria e declarações mensais, selecionando qual a equação mais conveniente para caso processado e gera as probabilidades de irregularidades.

```
**********  Rotina para processar dados mensais
************************************************************

   wTot = 0
   Dim rstProc As New ADODB.Recordset
   Dim wVl As Double
   With rstProc
      .CursorType = adOpenForwardOnly
      .LockType = adLockOptimistic
      .CursorLocation = adUseClient
      de_Sql = "select T.*, E.cl_24 as CL " _
            & "  from EMPRESAS as E, [" & cmbTab.List(cmbTab.ListIndex) & "] as T" _
            & " where E.TED_RGE_RUC = T.cd_Ruc "
      If ind = 0 Then
         de_Sql = de_Sql & " and dt_Refe = '" & lstRefer.List(lstRefer.ListIndex) & "'"
      End If
      .Open de_Sql, Globais.db, , , adCmdText
      While Not .EOF
         weq = 0
         wEqTemp = 0    '... para o caso de uma das eq. dar probabilidade 100%, tentar a proxima
         For E = 1 To wQtEq
            wEqOk = True
            wVl = 0
            For v = 1 To 14
               If tbCoef(E, v) <> 0 Then
                  wTemVar = False
                  For n = 1 To lvwVar.ListItems.Count
                     If lvwVar.ListItems(n).SubItems(2) = v Then
                        wvar = lvwVar.ListItems(n).SubItems(1)
                        If .Fields(wvar) > 0 Then
                           wVl = wVl + (tbCoef(E, v) * (Log(.Fields(wvar)) / Log(10)))
                           wTemVar = True
                           Exit For
                        End If
                     End If
                  Next
                  '
                  If wTemVar = False Then
                     wEqOk = False
                     Exit For
                  End If
```

```
            End If
         Next
         '
         If wEqOk = True Then
            wVl = wVl + tbConst(E)
            If !cl < 24 Then wVl = wVl + tbCoefCl(E, !cl)
            '
            wVl = Exp(wVl) / (1 + Exp(wVl))
            '
            If wVl = 1 Then
               wEqTemp = E
            Else
               weq = E
               Exit For
            End If
         End If
      Next
      If weq = 0 And wEqTemp > 0 Then
         wVl = 1
         weq = wEqTemp
      End If

'

lvwResultados.ListItems.Clear
If wTot > 0 Then
   For E = 1 To wQtEq
      Set itmX = lvwResultados.ListItems.Add(, , tbDeEq(E))
      itmX.SubItems(1) = Format(tbQt(E), "#,000")
      itmX.SubItems(2) = Format((tbQt(E) / wTot), "0.00%")
   Next
   Set itmX = lvwResultados.ListItems.Add(, , " - ")
   itmX.SubItems(1) = Format(tbQt(0), "#,000")
   itmX.SubItems(2) = Format((tbQt(0) / wTot), "0.00%")
   Set itmX = lvwResultados.ListItems.Add(, , "Total")
   itmX.SubItems(1) = Format(wTot, "#,000")
   MsgBox "Processamento Ok"
End If
```

APÊNDICE I – MODELO DE DADOS DA SOLUÇÃO INFORMATIZADA

A Figura 21 apresenta o modelo relacional da base de dados da solução informatizada.

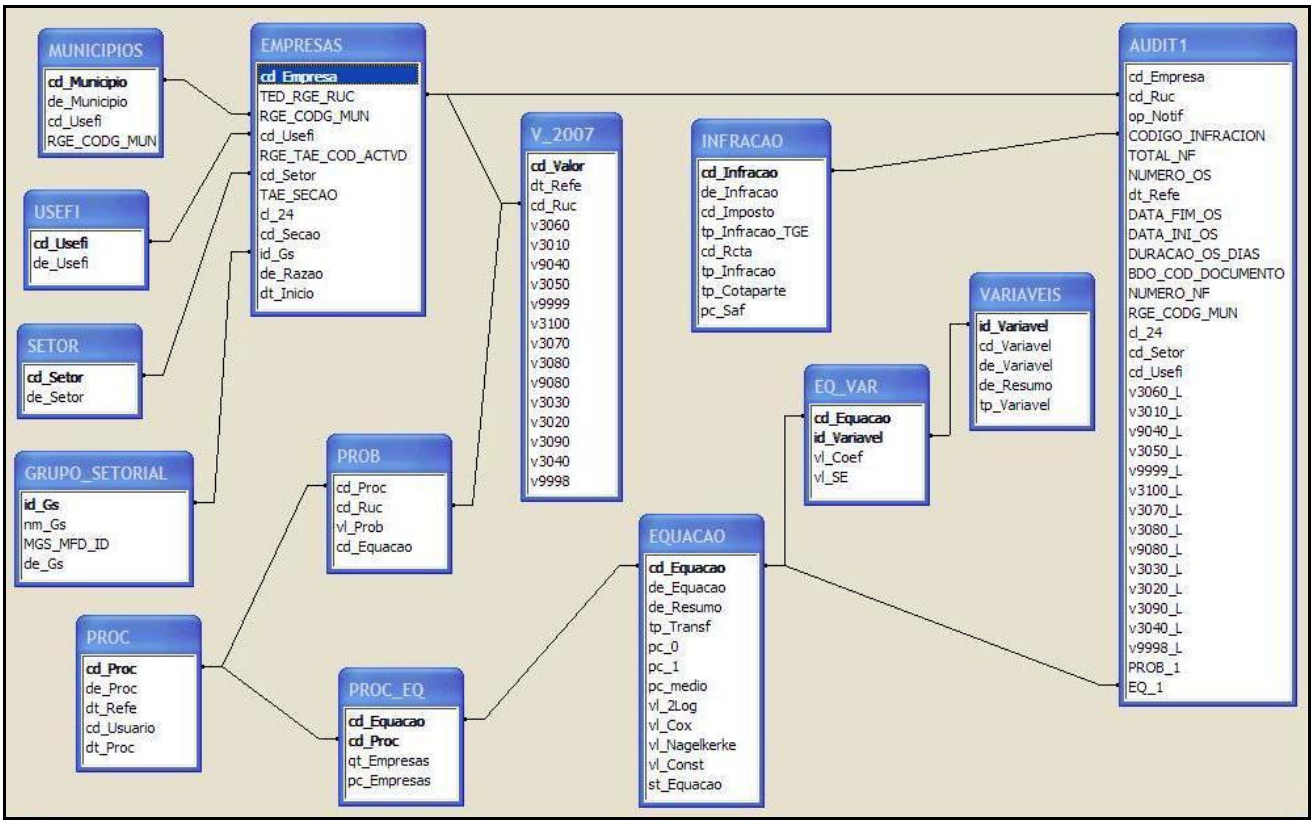

Figura 21 – Modelo de dados solução informatizada

APÊNDICE J – EXEMPLOS DE OPÇÕES DISPONÍVEIS NA SOLUÇÃO INFORMATIZADA

A Figura 22 – Tela de cadastro de equaçõesilustra a opção de cadastro de equações onde o operador cadastra as equações selecionadas na etapa de construção do conjunto de equações da presente metodologia. Pode-se observar detalhe com as 9 equações selecionadas na aplicação prática.

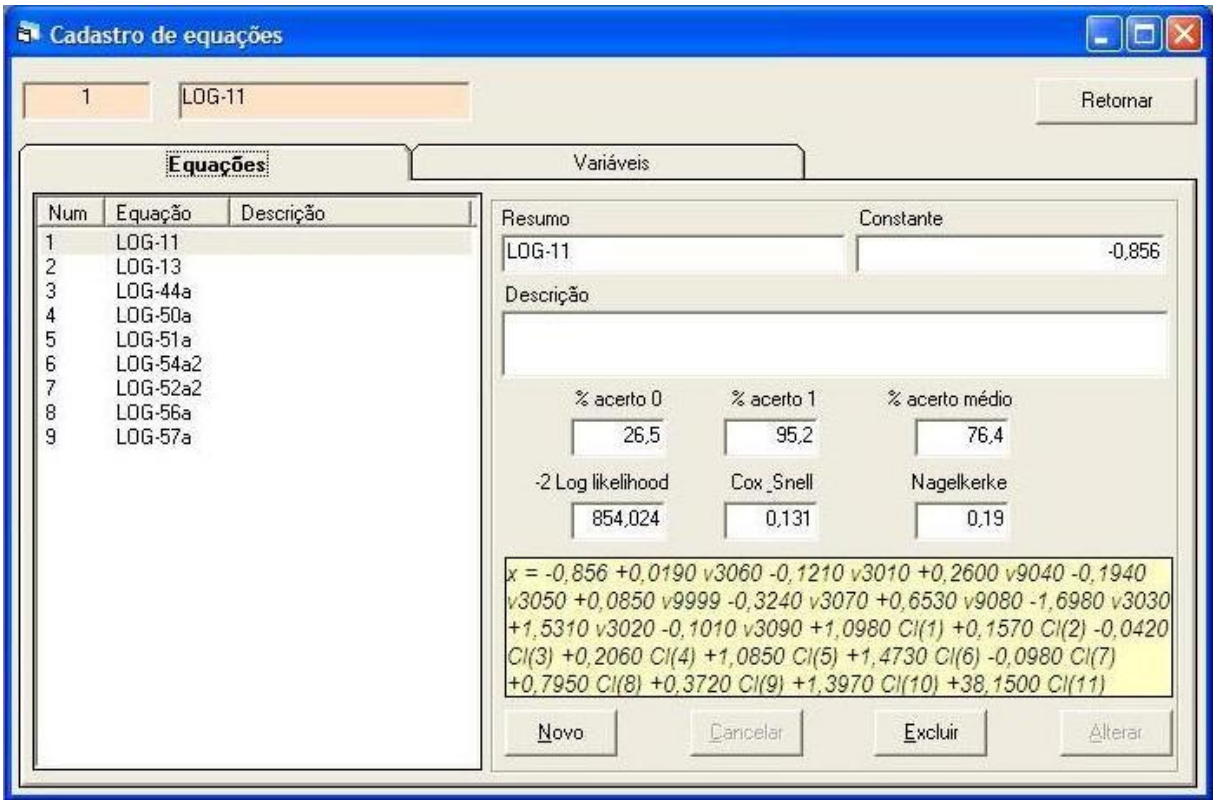

Figura 22 – Tela de cadastro de equações

Na Figura 23 observa-se a opção para cálculo de probabilidades, o operador deve selecionar qual base pretende processar: DIME ou Auditoria, em seguida escolhe o período de referência desejado. É possível processar os dados a partir de uma única equação ou usar todas as equações disponíveis. Após finalizar a execução todas as declarações da referencia escolhida terão suas probabilidades calculadas e armazenadas na base de dados. Será exibido também um sumário com os resultados obtidos, informando quantos casos foram processados em cada equação.

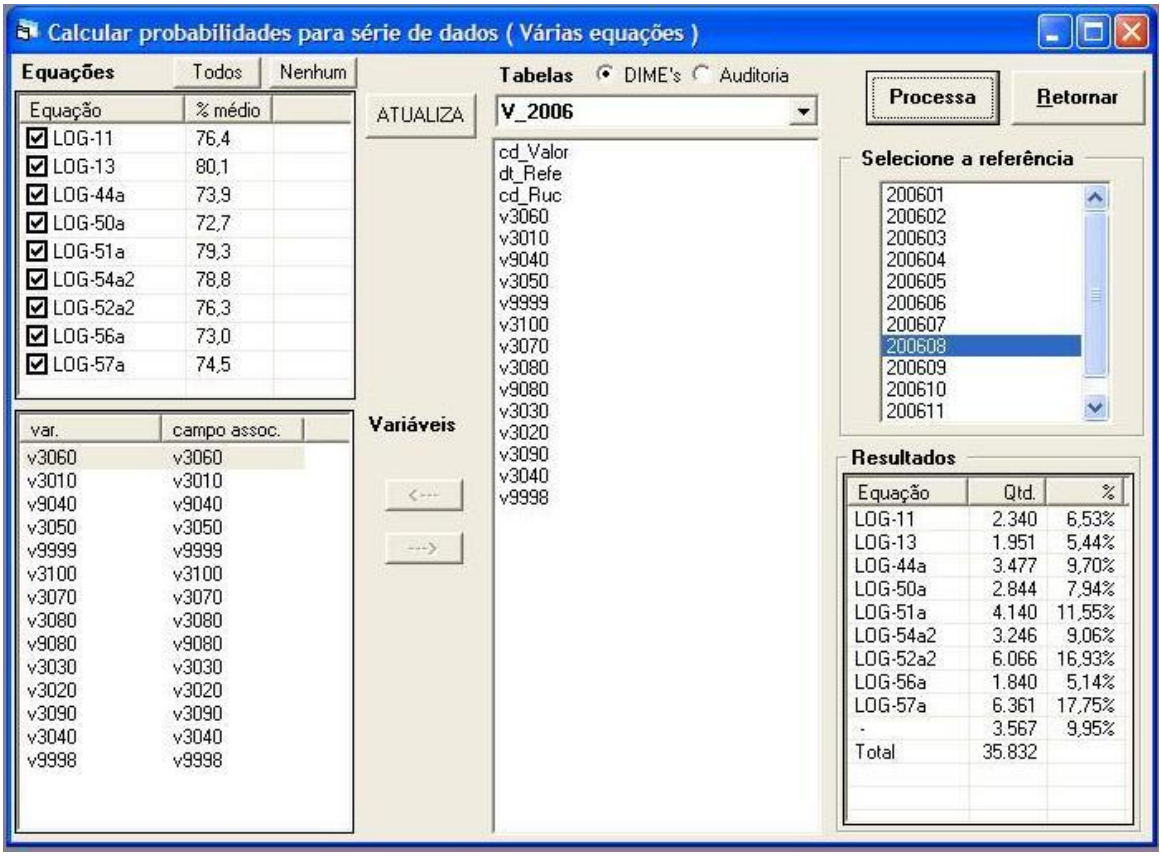

Figura 23 – Tela para cálculo de probabilidades

A opção disponível na Figura 24 apresenta uma das telas de consulta de probabilidades onde é possível filtrar as informações das DIME's por região geográfica (Usefi ou Município) e também por atividade econômica (grupo setorial, setor ou seção). O resultado da consulta é exibido na grade a direita da tela ordenado pelos maiores índices de irregularidades.

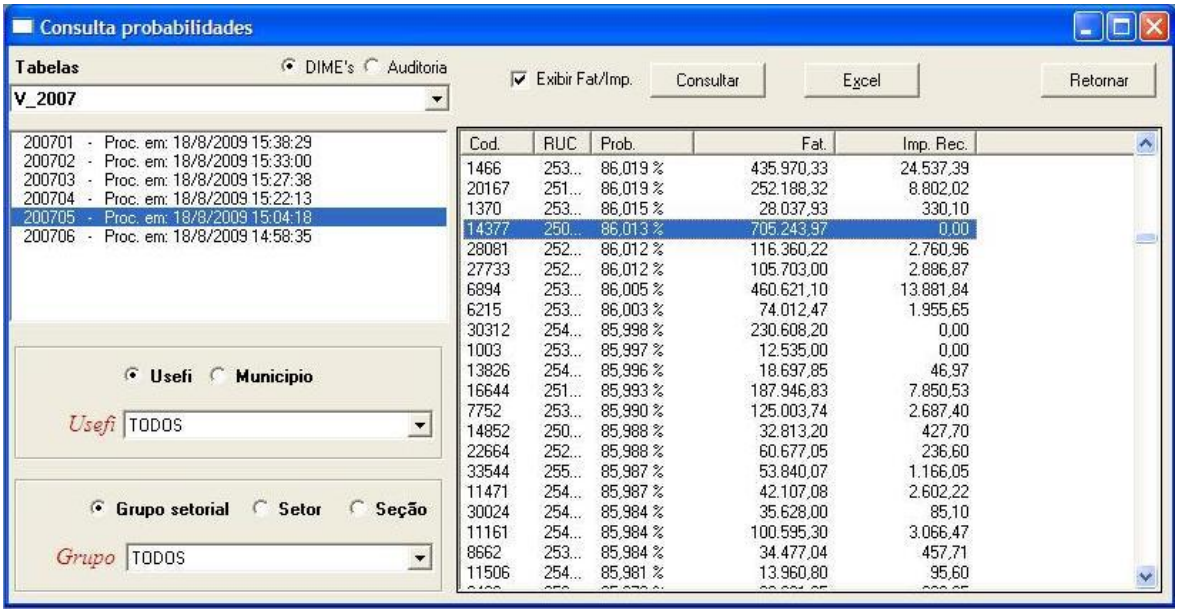

Figura 24 – Tela para consulta de probabilidades