UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

Um Estudo sobre Detecção de Desvios: Aplicação em bancos de dados da Secretaria da Saúde do Rio Grande do Sul

por

VERÔNICA LOUROZA ESTIVALET

Dissertação submetida à avaliação,
como requisito parcial para a obtenção do grau de
Mestre em Ciência da Computação

Prof. Dr. Luis Otávio Álvares
Orientador

Porto Alegre, dezembro de 2003

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Estivalet, Verônica Louroza
Estudo sobre Detecção de Desvios nos bancos de dados da saúde / por Verônica Louroza Estivalet – Porto Alegre: PPGC da UFRGS, 2003.
127 p.: il.

Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul. Instituto de Informática. Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2003. Orientador: Álvares, Luis Otávio.

1. Datamining. 2. Detecção de Desvios. I. Álvares, Luis Otávio. II. Título.

Agradecimentos

A Deus, a quem devo tudo o que conquistei na vida.

Ao meu orientador, Prof. Dr. Luis Otávio Álvares, por todo o apoio que me deu para a realização deste trabalho.

Ao meu marido, Adriano, pelo amor, carinho, dedicação, incentivo e pela compreensão nos meus períodos de estudos e viagens.

Aos meus pais, Plínio e Beatriz, pelo seu imenso amor e pela ajuda em todos os momentos.

Aos amigos da UCS e da UFRGS que conviveram comigo durante o período deste trabalho e pelos quais tenho um grande carinho.

“ O degrau de uma escada não serve simplesmente para que alguém permaneça em cima dele, destina-se a sustentar o pé de um homem pelo tempo suficiente para que ele coloque o outro um pouco mais alto. ”

Thomas Huxley

Sumário

Lista de Abreviaturas

AIH	Autorização de Internação Hospitalar
BD	Banco de Dados
DM	Data Mining
CID	Cadastro Internacional de Doenças
CRISP-DM	CRoss Industry Standard Process Model for Data Mining
CRS	Coordenadoria Regional de Saúde
DCBD	Descoberta de Conhecimento em Bases de Dados
IBM	International Business Machines Corporation
KDD	Knowledge Discovery in Databases
MD	Mineração de Dados
MS	Ministério da Saúde
SES	Secretaria Estadual de Saúde do Rio Grande do Sul
SIH	Sistema de Informações Hospitalares
SMS	Secretaria Municipal de Saúde do Rio Grande do Sul
SUS	Sistema Único de Saúde
UFRGS	Universidade Federal do Rio Grande do Sul

Lista de Figuras

Lista de Tabelas

Resumo

A mineração de dados é o núcleo do processo de descoberta de conhecimento em base de dados. Durante a mineração podem ser aplicadas diversas técnicas para a extração de conhecimento. Cada técnica disponível visa à realização de um objetivo e é executada de uma forma em particular. O foco desta dissertação é uma destas técnicas conhecida como detecção de desvios.

A detecção de desvios é baseada no reconhecimento do padrão existente nos dados avaliados e a capacidade de identificar valores que não suportem o padrão identificado. Este trabalho propõe uma sistemática de avaliação dos dados, com o objetivo de identificar os registros que destoam do padrão encontrado. Para este estudo são aplicadas algumas técnicas de avaliação estatística.

Inicialmente é apresentada uma revisão bibliográfica sobre descoberta de conhecimento em base de dados (DCBD) e mineração de dados (MD). Na seqüência, são apresentados os principais conceitos que auxiliam na definição do que é um desvio, quais as técnicas utilizadas para a detecção e a forma de avaliação do mesmo.

Dando continuidade ao trabalho, a sistemática CRISP_DM é descrita por ser aplicada aos estudos de casos realizados. A seguir, são descritos os estudos de casos realizados que utilizaram as bases da Secretaria da Saúde do Rio Grande do Sul (SES). Finalmente, são apresentados as conclusões do estudo e possíveis trabalhos futuros.

Palavras-chave: descoberta de conhecimento, mineração de dados, detecção de desvios.

TITLE: "A STUDY ON DETECTION OF DEVIATIONS: APLICATION IN DATABASE OF SECRETARIA DA SAUDE DO ESTADO DO RIO GRANDE DO SUL"

Abstract

The data mining is the core of the Process of Knowledge Discovery in Databases (KDD). Several techniques can be applied for the knowledge extraction during the mining. Each available technique seeks to the accomplishment of an objective and it is executed in a specific way in matter. The focus of this dissertation is one of these techniques known as deviations detection.

The detection of deviations is based on the recognition of the existent pattern in the appraised data and the capacity of identifying values not to support the identified pattern. This work proposes a systematic of evaluation of the data, with the objective of identifying the registrations that sound out of tune of the found pattern. For this study some techniques of statistical evaluation. are applied

Initially a bibliographical revision is presented on knowledge discovery in database (KDD) and data mining (DM). In the sequence, the main concepts are presented; the ones that aid in the definition of what is a deviation, which the techniques used for the detection and the form of evaluation of the same.

Giving continuity to the work, the CRISP_DM process model is described . Next, the realized experiments that used the bases of the Secretary of the Health of Rio Grande do Sul (SES). are described . Finally, the conclusion and futures work are dependent.

Keywords: Knowledge discovery, data mining, detection of deviation.

1 Introdução

O processo de descoberta de conhecimento em bases de dados (DCBD) descreve a busca de conhecimento implícito em grandes bases de dados e a capacidade de tornar este conhecimento acessível ao usuário.

A preocupação em validar técnicas, ferramentas e metodologias capazes de disponibilizar o conhecimento armazenado nas bases de dados e a representação deste conhecimento de forma compreensível ao usuário descrevem o principal objetivo da DCBD. Para realizar este objetivo, são utilizadas técnicas de aprendizado de máquina, inteligência artificial e de conceitos estatísticos que permitem lidar com a incerteza relacionada às descobertas.

De acordo com Parsaye [PAR 89] a informação hoje é superabundante e a capacidade de armazená-la excede a capacidade de efetivamente recuperá-la. Claramente existe uma grande preocupação na recuperação da informação armazenada, para torná-la útil ao trabalho do usuário.

Os softwares gerenciadores de base de dados (SGBD) apresentam limitação na recuperação da informação, pois disponibilizam principalmente consultas aos dados fisicamente armazenados. Existem diversos tipos de bases de dados sendo utilizadas. Esta diversificação de padrões torna difícil o trabalho de pesquisa nas mesmas. Sendo assim, a necessidade de se obter informação útil e de qualidade, abriu espaço para o estudo e a criação de diversas ferramentas que possibilitam o trabalho com as informações disponíveis.

O processo de DCBD envolve várias etapas. A mineração de dados MD (*data mining*, DM), considerada núcleo do processo, descreve o trabalho de encontrar uma informação relevante em um grande conjunto de informações diversas. Os algoritmos aplicados nesta fase utilizam técnicas de extração de padrões e detecção de desvios, e são aplicados sobre os dados pré-processados. Este processo permite a identificação de padrões e a descoberta de relacionamentos não definidos na base de dados. São utilizadas técnicas estatísticas adaptadas a algoritmos de aprendizado e híbridas simbólico/conexionistas para possibilitar a descoberta de informações significativas. Pode-se reunir as diversas propostas e agrupar os métodos em oito categorias [FRI 97]: associação, *clustering,* descrição de conceitos, detecção de desvios, seqüência, agrupamento por séries temporais, classificação e regressão. Neste trabalho será apresentado um estudo mais detalhado sobre detecção de desvios.

O processo de detecção de desvios é um conceito amplo dentro da mineração de dados. Podendo ser considerado um desvio de qualquer instância que não se enquadre nos valores esperados. As instâncias encontradas podem determinar padrões interessantes durante o processo de pesquisa.

Os algoritmos utilizados neste processo podem ser utilizados sobre modelos de dados que sejam derivados de outros algoritmos como: descrição de dependência, seqüência e ou descrições de conceitos, que podem ser obtidos automaticamente ou com a intervenção do usuário. O objetivo é encontrar informações fora dos parâmetros normais, ou seja, casos anômalos [FEL 96].

Há métodos de descoberta de objetos que não seguem padrões (classes) de valores, sendo necessário à definição de padrões de forma prévia, comumente chamados de normas [MAT 93]. Talvez este processo de classificação não diga respeito ao método de mineração, mas à aplicação dos resultados da mesma. A partir das definições de regras, classes ou comportamentos, os quais são considerados padrões corretos, é possível afirmar que um desvio é uma quebra deste padrão.

Existem definições diferentes para os desvios detectados, tais definições dependem do ambiente e das situações onde estes desvios ocorrem. Com o crescente avanço das redes de computadores, tornou-se prioridade a segurança da rede e dos processos realizados nela [LEE 97]. A detecção de intrusão trabalha com o enfoque de evitar que algum indivíduo tenha acesso as informações da rede. Para este trabalho é importante o método utilizado para executar esta detecção.

O outro tipo de desvio considerado, a fraude, pode ser exemplificado no processo de utilização de cartões de crédito. Com o aumento dos pontos de acesso ao sistema financeiro, cresce a preocupação sobre a segurança do sistema e dos clientes que utilizam este serviço, tornando necessário o estudo sobre o método de detecção de fraudes [STO 97].

A presença de desvios em bases de dados é um problema real e que cresce juntamente com a tecnologia. O tipo do desvio é diretamente determinado pelo tipo de base de dados e pela utilização dada a estes dados. Logo, a eficácia do trabalho de detecção de desvios é definida de acordo com a configuração da base de dados e com o tipo de informação que se pretende trabalhar [ARN 96].

1.1 Objetivos

O objetivo principal deste trabalho é estudar algumas técnicas de detecção de desvios e aplicá-las em dados reais da Secretaria da Saúde do Estado do Rio Grande do Sul (SES). Com a utilização de técnicas estatísticas e a aplicação de uma metodologia de trabalho, pretende-se gerar resultados que confirmem os estudos realizados.

1.2 Descrição das atividades

Este trabalho permitiu a coleta de informações qualitativas sobre experiências e testes, realizados na base de dados da SES, utilizando-se de conceitos de seleção e preparação de dados e técnicas de mineração (*data mining*), apresentadas em diversas propostas na bibliografia estudada.

Realizou-se um estudo sobre técnicas estatísticas para verificar qual se enquadraria melhor aos dados e a proposta de detectar os desvios existentes na base.

Através deste estudo foram observadas características comuns, que envolvem os mesmos processos de preparação, aplicação e análise de dados. Para aprimorar tal avaliação também foram feitos testes.

1.3 Organização do texto

Quanto à organização dos capítulos da dissertação tem-se a seguinte estrutura:

- Capítulo 2 - apresenta algumas definições e principais conceitos sobre a área de descoberta de conhecimento em base de dados e mineração de dados. Apresenta uma divisão de etapas para o processo de DCBD. Ainda no capítulo 2 é apresentado o estudo sobre a mineração de dados e as principais técnicas utilizadas neste processo.

- Capítulo 3 - apresenta as definições e os tipos de desvios encontrados na literatura, juntamente com as descrições das técnicas estatísticas utilizadas.

- Capítulo 4 - mostra a metodologia utilizada neste trabalho e encerra apresentando uma descrição simplificada do protótipo desenvolvido.

- Capítulo 5 - apresenta a descrição dos experimentos realizados sobre a base de dados da SES. Mostra todo o processo de estudo: estudo do domínio, seleção e preparação dos dados, estudo do software, aplicação das técnicas e avaliação dos resultados.

- Capítulo 6 – apresenta as considerações finais sobre este trabalho.

2 Descoberta de conhecimento em bases de dados e mineração de dados

Neste capítulo são apresentados os principais conceitos e objetivos que descrevem a descoberta de conhecimento em bases de dados (DCBD) e a mineração de dados (MD). Tais conceitos servem de embasamento teórico deste trabalho, que tem como foco principal à análise da detecção de desvios no processo de mineração de dados.

2.1 DCBD

O processo de reconhecer informações implícitas em grandes bases de dados e a capacidade de tornar este conhecimento acessível ao usuário é definido como descoberta de conhecimento em bases de dados (DCBD), também conhecido como KDD *(Knowledge Discovery in Databases)*.

Atualmente os bancos de dados apresentam uma tecnologia capaz de proporcionar o armazenamento de uma grande quantidade de informações operacionais nas empresas e locais de pesquisa. O conhecimento implícito nos dados armazenados geralmente não está disponível para o usuário.

A necessidade de informações úteis e de qualidade desencadeou o surgimento e o desenvolvimento de técnicas específicas que, quando aplicadas sobre as bases de dados, permitem a extração de conhecimento oculto e até desconhecido. O conhecimento resultante desta operação pode ser utilizado como apoio à decisão dentro de uma empresa ou instituição de pesquisa.

O objetivo principal da DCBD é validar técnicas, ferramentas e metodologias capazes de disponibilizar o conhecimento armazenado nas bases de dados e a representação deste conhecimento de forma compreensível ao usuário [FAY 96]. O processo de criação destas ferramentas não é uma tarefa simples, existem problemas relacionados a banco de dados que devem ser considerados como: a qualidade dos dados e a forma como estão armazenados no banco.

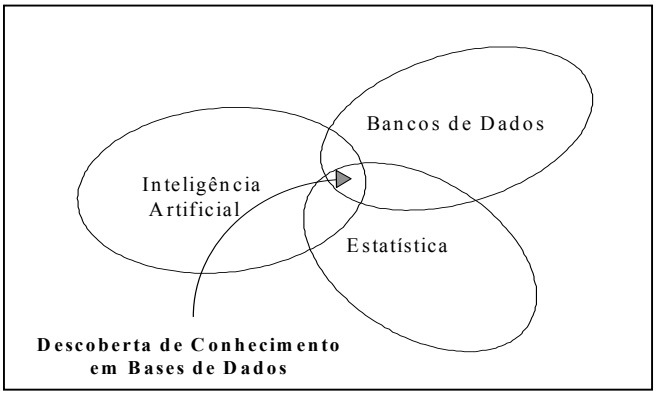


FIGURA 2.1 - Contexto da DCBD [FEL 96]

Para sanar estas dificuldades o processo de descoberta de conhecimento tem fundamentação também em outras áreas de estudos que possuem técnicas apropriadas a

este objetivo. Entre as técnicas utilizadas para o processo de DCBD podem ser utilizadas: técnicas de aprendizado de máquina, inteligência artificial e conceitos estatísticos que permitem lidar com a incerteza relacionada às descobertas [FEL 96], esta visão é representada na figura 2.1.

2.2 Etapas do DCBD

Na literatura acadêmica e nos trabalhos desenvolvidos pelos centros de pesquisas é possível encontrar diferentes definições que determinam as etapas do processo de DCBD. O processo descrito na figura 2.2 é amplamente conhecido e destaca cada etapa individualmente.

O processo de DCBD é interativo[1] e iterativo[2], o envolvimento do usuário é constante e essencial. O início do processo parte de um conjunto de dados originais, sem tratamento, seguido pela seleção e o pré-processamento dos dados. Durante esta etapa é efetuada a limpeza dos dados, com a finalidade de enquadrar os dados para a aplicação dos algoritmos de mineração. A adequação destes dados aos algoritmos torna essencial o estudo e entendimento do problema proposto e do domínio da aplicação. A etapa seguinte consiste na utilização de um algoritmo de mineração, que tem por objetivo a extração de conhecimento implícito na base de dados. Os resultados obtidos devem ser interpretados e avaliados, sendo o passo final a assimilação do conhecimento.

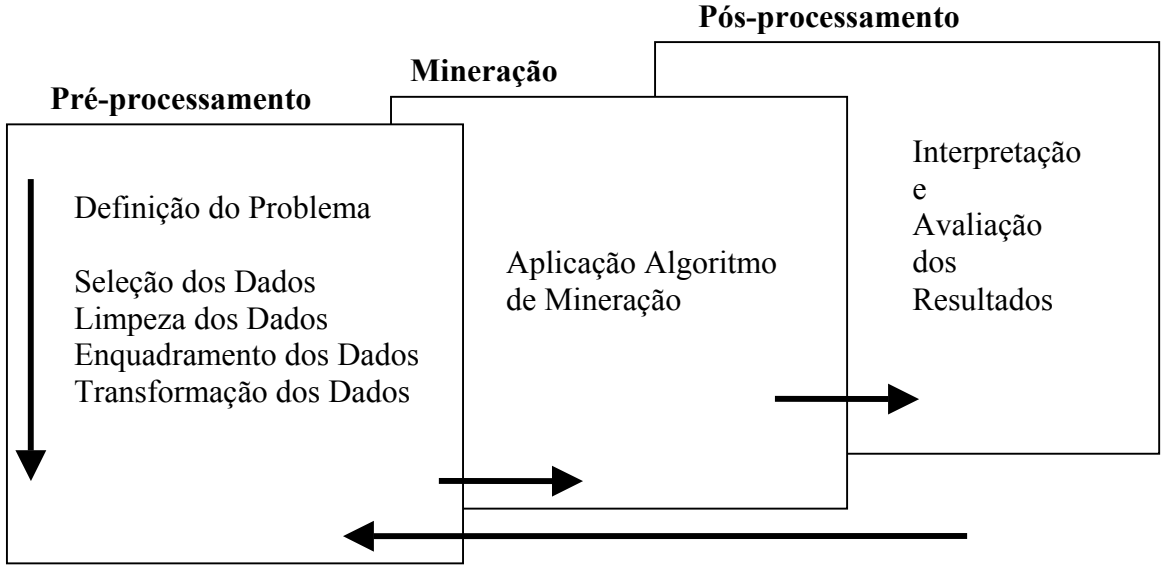


FIGURA 2.2 -Processo de mineração de dados

Os passos apresentados na figura 2.2 podem ser distribuídos em três etapas básicas, apresentadas a seguir.

1 Que possibilita uma comunicação com dois sentidos [DIC 98].

2 Feito ou repetido várias vezes [DIC 98].

2.2.1 Pré-processamento

A primeira condição necessária para a aplicação do processo de descoberta é o entendimento sobre o problema proposto. A definição do problema envolve a compreensão do domínio e a definição clara dos objetivos da pesquisa, proporcionando um trabalho mais direcionado dentro da base de dados. A partir deste entendimento é possível definir os parâmetros usados para a seleção dos dados que serão utilizados no processo.

No processo de DCBD é essencial que as pessoas envolvidas tenham uma boa compreensão do assunto trabalhado, conhecendo o tipo de informação existente nas bases de dados, bem como sua forma de armazenamento, o que determina um bom embasamento para a definição de um objetivo na descoberta de conhecimento e posteriormente uma boa avaliação dos resultados obtidos.

Este processo evita que informações importantes à pesquisa sejam perdidas. O resultado obtido sem a definição e estudo do problema pode ser considerado pouco confiável, mesmo que o desenvolvimento do processo envolva a aplicação de técnicas sofisticadas para a extração de conhecimento [TWO 98].

Antes de iniciar a preparação dos dados é preciso escolher um algoritmo de mineração, pois cada algoritmo possui limitações e parâmetros especiais de configuração. Os dados devem ser preparados de acordo com as necessidades do algoritmo selecionado.

Com os parâmetros iniciais definidos, começa a seleção, limpeza e transformação dos dados. Estas atividades envolvem o maior tempo do processo, pois devem considerar a representação da informação e a qualidade dos dados encontrados na base de dados.

Algumas poucas bases de dados estão completas e tornam-se fáceis de trabalhar. A grande maioria das bases apresenta vários problemas como: a informação redundante dentro da base, dados com informações irrelevantes, dados esparsos e campos incompletos. Ruídos podem fazer parte da base de dados e consistem em erros nos valores de atributos ou em informações de classes, que podem levar a resultados equivocados [FRI 97].

A seleção dos dados determina a criação de um conjunto de informações que será utilizado na descoberta de conhecimento. A partir do conjunto de dados é possível de definir o conjunto de atributos, amostras de dados para serem mineradas e determinar exclusões de dados irrelevantes aos objetivos da descoberta. Durante a etapa da seleção é comum ocorrer uma diminuição na quantidade de dados trabalhados, com o objetivo de otimizar o processo de descoberta.

O objetivo principal da etapa de limpeza é promover a melhoria na qualidade dos dados. Nesta etapa são aplicados algoritmos que processam a limpeza nos dados, eliminando ruídos (pequenas alterações) e aplicando estratégias para o tratamento de campos que não apresentam valores. Com os dados compreendidos, selecionados e limpos é possível prepará-los para a utilização dos algoritmos de mineração de dados, trabalho este que consiste na etapa seguinte do processo.

O enquadramento determina a adaptação dos dados a algum método de mineração, se necessário a alterar a forma estrutural do armazenamento. Este trabalho deve considerar a otimização do processo de descoberta e não desconsiderar a qualidade e a integridade dos dados utilizados.

Riddle [RID 94] define as etapas do pré-processamento como: escolha das instâncias, escolha de atributos relevantes, ajuste da representação, representação de tempo, ajuste de parâmetros indutivos.

A definição das instâncias a serem trabalhadas estabelece a amostra dos dados que serão pesquisados. Como descrito no texto acima, a escolha dos dados está diretamente ligada com o objetivo que se deseja alcançar na pesquisa e a aplicação do resultado obtido.

Para uma boa preparação dos dados é necessário considerar a escolha de atributos relevantes à pesquisa, pois nem todos os dados contidos numa determinada base de dados são interessantes ao objetivo definido anteriormente. Esta etapa pode ocasionar uma diminuição do espaço de busca, reduzindo o tempo e o custo do processo.

O passo seguinte é a preparação da representação dos dados, a qual consiste no ajuste dos dados tanto ao nível dos dados quanto ao nível de acesso aos arquivos, tornando os dados mais acessíveis ao algoritmo utilizado na descoberta. Algumas bases apresentam atributos categóricos, que podem ser classificados em faixas pré-definidas, sempre considerando o conhecimento do usuário quanto à base de dados.

Outro aspecto que pode ser considerado para a descoberta é o tempo. Assim como os dados categóricos em intervalos, o tempo pode ser tratado da mesma forma. Desta maneira, os dados de tempo descritos na base de dados podem ser informações importantes no processo de descoberta.

No processo de DCBD são utilizados algoritmos para a extração de conhecimento. Em sua maioria esses algoritmos dependem, para seu funcionamento, de informações que norteiam sua execução. Tais parâmetros são informados pelo usuário. Esta etapa é definida como ajuste de parâmetros indutivos.

Baseado nas informações apresentadas acima, conclui-se que a preparação dos dados é de máxima importância para o processo de descoberta. Os dados devem ser selecionados, a partir da base de dados utilizada para a pesquisa, sendo que esta seleção é possível com um bom conhecimento do domínio. A etapa seguinte consiste na consolidação da base de dados, promovendo a limpeza de dados que apresentem ruídos, dados incompletos e a retirada de dados irrelevantes a pesquisa.

A escolha do algoritmo que fará a tarefa de mineração é importante na preparação dos dados. Os algoritmos, aplicados na fase de mineração, podem conter particularidades para a sua configuração e utilização. Alguns algoritmos não trabalham com dados incompletos, não tratam valores categóricos entre outras limitações. O usuário deve decidir por um determinado método de pesquisa, para depois determinar qual o algoritmo será utilizado. Após a escolha do algoritmo é necessário preparar os dados para serem empregados na execução.

2.2.2 Mineração

A mineração é a aplicação do algoritmo sobre os dados já preparados, gerando resultados que podem representar padrões interessantes ou informações completamente irrelevantes à pesquisa. A execução pode gerar representações como regras de classificação, regressão e agrupamentos, entre outros.

A extração de padrões consiste no núcleo de um sistema de mineração, composto pelos algoritmos de extração de padrões, detecção de desvios, etc (Matheus *apud* [FEL 97]). A extração de padrões conta com o *background* de áreas de pesquisa como o aprendizado de máquina. São utilizadas algumas técnicas estatísticas que possibilitam algumas descobertas úteis. Estas técnicas são adaptadas a algoritmos de aprendizado e metodologias híbridas simbólico/conexionistas.

De acordo com Fayyad [FAY 96], esta fase é apresentada como mineração de dados. O mesmo autor apresenta conceitos distintos para diferenciar Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery on Databases*) e Mineração de Dados (*Data mining*). Como citado anteriormente, a descoberta de conhecimento é "o processo não trivial de identificar padrões em dados que sejam válidos, novos, potencialmente úteis e fundamentalmente compreensíveis". Mineração de Dados é, "um passo dentro do DCBD, que determina a aplicação dos algoritmos de descoberta de dados que, sob certas limitações de eficiência computacional aceitáveis, produzem uma enumeração particular de padrões sobre estes dados".

A etapa de mineração de dados pode sofrer prejuízos caso ocorra algum erro nas etapas anteriores. Sendo a etapa que mais consome recursos computacionais, existe a preocupação em agilizá-la controlando o tempo e os recursos gastos. A seção 2.3 trata exclusivamente sobre mineração de dados.

Nesta fase pode ser gerado um modelo de aplicação. A construção do modelo é um processo interativo, pois é necessário testar vários modelos para encontrar qual é o modelo ideal para a pesquisa. Este trabalho de adequar o modelo ao objetivo da pesquisa é cíclico: testar, treinar, avaliar e modificar se necessário. O trabalho de construção do modelo é efetuado sobre uma parte dos dados, só depois de testado e avaliado, a segunda parte dos dados é utilizada para validar o modelo. Esta divisão da base de dados determina uma precisão maior ao trabalho desenvolvido. Para validar o modelo existem formas de calcular a taxa de precisão do mesmo. Tais técnicas envolvem um controle de testes e treinamento do modelo utilizando uma parte da base de dados.

2.2.3 Pós-processamento

A etapa de pós-processamento trabalha com os resultados do processo de DCBD, na qual os dados obtidos são interpretados e avaliados. Esta etapa do processo consiste em consolidar o conhecimento obtido, para associá-lo a um sistema e possibilitar a sua utilização em execuções futuras, ou simplesmente efetuar o seu registro e posteriormente liberá-lo para os usuários que tenham interesse.

A comparação entre os resultados previstos e os resultados encontrados pode apresentar diferenças, e esta linha de pesquisa pode ser não ser considerada válida e útil.

Neste caso é possível revisar o processo efetuado e se necessário retornar a algum passo anterior com o objetivo de refinar os resultados. De acordo com Zytkow [ZYT 93], uma grande quantidade de dados pode ser extraída da base de dados e grande parte desta informação não ser interessante para o usuário.

O processo de avaliação destes resultados e a definição dos padrões são vitais para que um sistema de DCBD possa ser utilizado em aplicações práticas. Os termos para a avaliação dos dados são referentes: à precisão, cobertura, novidade e simplicidade [PIA 93]. É importante ressaltar a forma de apresentação dos padrões encontrados para o usuário, que pode ser por gráficos, esquemas, etc.

Os dados resultantes de um processo bem feito podem ser aplicados diretamente na solução do problema. O processo de conhecimento pode ser utilizado por diversas áreas de atividades.

Um exemplo de aplicação real pode ser a necessidade de mudança da forma de atuação de uma determinada instituição de saúde. Baseado nos dados existentes em sua base de dados à instituição pode dirigir seu trabalho na melhora do atendimento buscando a satisfação dos pacientes ou aprimorar os conhecimentos em relação às doenças apresentadas durante os períodos do ano. Isto significa que dependendo da meta específica, é possível "aumentar a taxa de resposta" ou "aumentar o valor de uma resposta" [TWO 98].

2.3 Mineração de dados

Mineração de dados como visto na seção anterior e uma das etapas do processo de descoberta de conhecimento. A seguir são apresentadas sucintamente as principais técnicas utilizadas.

2.3.1 Técnicas utilizadas na mineração

Existem dois modelos principais para o processo de mineração de dados [TWO 98], O primeiro é o modelo preditivo (*Predictive model*), definido pela utilização de dados com resultados conhecidos. Este modelo pode ser utilizado para predizer valores através de dados diferentes. Um exemplo conhecido é a avaliação dos clientes para operações de crédito. São avaliados comportamentos e características conhecidas, e a partir destes padrões já conhecidos é possível avaliar os novos clientes.

O segundo é o modelo descritivo (*descriptive model*), este é capaz de descrever padrões em dados existentes. A diferença principal entre os dois modelos é que o primeiro é capaz de inferir valores para os resultados futuros, enquanto o segundo auxilia na descrição dos dados. A aplicação destes modelos varia de acordo com o objetivo da mineração.

A mineração de dados pode utilizar mais de uma técnica para atingir seus objetivos. Estas técnicas podem ser utilizadas individualmente ou em conjunto [FAY 96]. são elas: regressão, associação, *clustering*, classificação, sumarização, detecção de desvios, seqüência, agrupamento por séries temporais. Para cada uma destas técnicas existem algoritmos específicos.

2.3.1.1 Regressão

Os algoritmos de regressão geralmente procuram descobrir fórmulas, trabalham de forma semelhante aos que efetuam o aprendizado supervisionado. Regressão é o aprendizado de uma função que mapeia cada entidade para um valor numérico [FAY 96]. Estes métodos são utilizados para ajustar curvas, úteis para modelos preditivos. A partir de um conjunto de pontos este método é capaz de calcular pontos anteriores, intermediários e posteriores. Um exemplo de aplicação é calcular a probabilidade de vida de um paciente, baseado nos exames de diagnóstico.

2.3.1.2 Associação

Utiliza mecanismos para a criação de modelos que descrevem dependências entre os dados. A técnica de associação é capaz de formular regras, a partir de ocorrências de itens em um grupo de transações, as quais têm um fator de confiança que expressa o percentual de acerto desta regra em novas transações. Associação pode também ser chamada de análise de dependências ou de modelagem de dependências.

2.3.1.3 Agrupamento

A técnica de agrupamento é uma tarefa descritiva onde é identificado um conjunto finito de categorias, ou *clusters,* para descrever os dados. Neste processo não são conhecidas as classes, o algoritmo explora as diferentes classes para detectar padrões, após os objetos são agrupados pelas suas propriedades. Este mecanismo de agrupamento é considerado como aprendizado não supervisionado.

2.3.1.4 Classificação

Consiste na execução de uma função que mapeia (classifica) um item de dado em uma entre as diversas classes pré-definidas [FAY 96]. O resultado é um conjunto de regras que possibilitem tal classificação.

Na descoberta de riscos de crédito e fraudes o método de classificação é bem adaptado, pois a partir de um conjunto de informações já classificadas é possível analisar as novas informações. Esta aplicação é freqüentemente utilizada em algoritmos de classificação baseado em redes neurais artificiais ou árvores de decisão [NOT 97].

Nesta técnica, há um conjunto de dados pré-determinados que serão utilizados no processo de classificação, estes dados são denominados conjunto de treinamento de transações.

2.3.1.5 Sumarização

Esta técnica cria descrições para uma determinada classe de dados. Os algoritmos de sumarização montam descrições para cada classe apresentada, são identificadas características comuns entre os componentes da classe as quais são utilizadas para descrever as classes. Esta técnica é classificada como aprendizado supervisionado.

2.3.1.6 Detecção de Desvios

Detecção de desvios é uma técnica que tem como objetivo a descoberta de valores ou atributos que contenham informações fora dos padrões esperados. Para a caracterização dos desvios é necessária uma definição antecipada de padrões partir desta pré-definição, os dados que não se enquadrarem são considerados desvios.

2.3.1.7 Seqüência

Define uma associação temporal dos fatos, isto é, existe uma dependência temporal dos itens relacionados existindo uma relação de causa e efeito. A seqüência é semelhante ao que acontece na técnica de associação. A diferença é que na associação a dependência existe dentro da mesma transação enquanto que na seqüência os itens que se relacionam existem dentro de transações diferentes.

As transações na seqüência acontecem em uma ordem temporal, mas são independentes. Um exemplo deste tipo de regra é o procedimento de parto, de forma que este procedimento só pode ocorrer após a detecção da gravidez da paciente. As transações seguem um ritmo cronológico.

Seguindo a distribuição das transações temporais, a seqüência pode representar a distribuição de atributos numa tabela. Dependendo da avaliação da ordem dada aos atributos na tabela é possível considerar resultados diferentes durante o processo de mineração, segundo Hermam *apud* [NOT 97].

3 Desvios

Como visto no capítulo 2, detecção de desvios[3] (*outliers*) é uma técnica de mineração de dados. Durante a mineração de dados é possível encontrar métodos que consideram e apresentam os desvios encontrados nas bases de dados. Dependendo do método utilizado e principalmente do objetivo da descoberta, tais desvios, podem ser desconsiderados dos dados avaliados.

É crescente o interesse por este assunto, pois atualmente o conhecimento gerado pelos desvios encontrados representa uma importante ajuda nas avaliações e definições de estratégias em diversos campos de atuação no mercado profissional. Em cada área de atuação existem necessidades específicas que visam à melhoria e garantia da qualidade do serviço prestado.

3.1 O que é desvio

Nos elementos que compõem uma base de dados é possível encontrar: informações consistentes, que formam um padrão de informações; dados que não se enquadram nos valores esperados; e mesmo que eventualmente, problemas com a qualidade dos dados.

Os dados que não se enquadram nos valores esperados, segundo Matheus, são considerados desvios e podem determinar padrões interessantes durante o processo de pesquisa (Matheus *apud* [FEL 97]. O mesmo autor cita ainda desvio como sendo:

- Instâncias que não se enquadram nas classes definidas;
- Superposições entre as classes;
- Classes que se diferenciam muito de suas classes pais;
- Mudanças no valor em um período de tempo;
- Discrepâncias entre valores observados e valores esperados previstos pelo modelo.

Desvios são definidos a partir de avaliações realizadas na base de dados. O conhecimento do domínio e a aplicação de técnicas de mineração possibilitam, ao analista, a identificação do padrão existente nos dados e posteriormente a identificação dos valores que destoam deste padrão.

Reafirmando a definição de desvios, Feldens cita: desvio é sempre um valor contrastante entre uma observação e um valor referencial. [FEL 96]. Da mesma forma Barnett e Lewis definem desvios como sendo uma observação que aparentemente não está consistente em relação a um conjunto de dados (Barnett e Lewis *apud* [KNO 2002] p. 3).

---

[3] Sinônimos para o termo desvio incluem: deslize, erro, falta, afastamento, distanciamento, apartamento, extravio, perda, sumiço. [Dic 98]

Para Samuels, um desvio é definido como, uma observação que é suficientemente diferente do resto dos dados [SAM 89].

Para Hawkins, uma definição intuitiva de desvios é uma observação que desvia de outras observações e considera igualmente importante a forma como estes dados foram gerados (Hawkins *apud* [KNO 2002]).

O pesquisador deve estudar e conhecer as informações descritas na base de dados e se aprofundar sobre o ambiente no qual estes dados foram gerados. Só com um conhecimento sobre o ambiente e um bom estudo do domínio é possível classificar os desvios como dados significativos para a descoberta de conhecimento.

Na área estatística são encontradas, diversas formas de avaliação dos dados, formas de avaliação dos dados e conseqüentemente as posteriores detecções dos desvios. Com a aplicação de diferentes técnicas de cálculo é possível definir os padrões existentes e conseqüentemente identificar os desvios, pois não seguem o modelo estatístico do resto dos dados. De acordo com os estudos realizados é possível determinar desvios que são definidos a partir de aplicações e resultados distintos..

Este trabalho baseou-se na aplicação de técnicas estatísticas para a detecção de desvios. Na literatura estatística, o termo desvio é identificado como *outlier*, logo, consideraram-se as definições de *outliers* como sendo definições de desvios.

Um desvio (*outlier*) é qualquer dado que estiver deslocado em relação ao resto dos dados avaliados. Para Knorr, dependendo da distribuição dos dados, um desvio pode se caracterizar por (Knorr *apud* [DOM 2003]):

- Um valor extremo ou relativamente extremo;
- Um “contaminante”, que é uma observação de alguma outra distribuição (possivelmente desconhecida);
- Um valor de dado legítimo, mas surpreendente ou inesperado.
- Um valor de dado que foi medido ou gravado incorretamente.

Considerando-se o deslocamento dos desvios em relação ao agrupamento dos dados avaliados, o desvio pode ser determinado de duas formas: desvio bruto (*gross outliers*) e desvio estrutural (*structural outliers*).

Analisando-se o resultado de uma avaliação realizada sobre uma base de dados, é possível observar um padrão de comportamento dos dados. Na seqüência do trabalho de avaliação pode-se observar a existência de uma faixa de valores que são assumidos pelas variáveis analisadas. Os desvios considerados brutos são os registros que apresentam valores afastados desta faixa padrão, i.e. não existe outro registro que apresente um valor mais extremo do que o desvio. A figura 3.1 apresenta um exemplo de desvio bruto. Como pode ser observado, o valor considerado desvio está mais abaixo e mais à esquerda do que o conjunto dos outros valores da base.

O desvio estrutural também apresenta um valor diferente dos demais valores analisados, mas este valor não é um extremo. A figura 3.2 mostra um desvio

estrutural. Enquanto na figura 3.1 o desvio bruto é um valor distante dos demais valores do agrupamento, na figura 3.2 o desvio possui um valor que não é um extremo dos demais valores registrados. O desvio estrutural está distante do agrupamento principal dos dados, por este modo é um desvio. Para a avaliação do desvio estrutural é necessário analisar toda a base de dados.

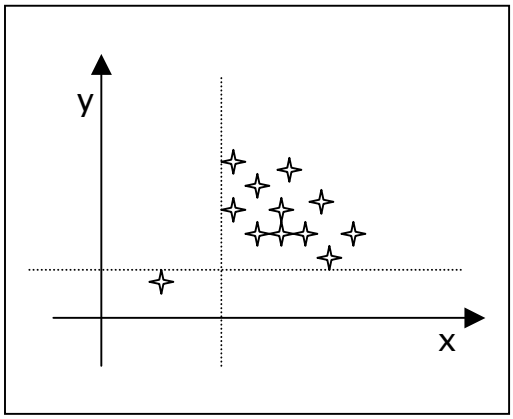


FIGURA 3.1 - Desvio Bruto (*Gross Outliers*) [KNO 2002]

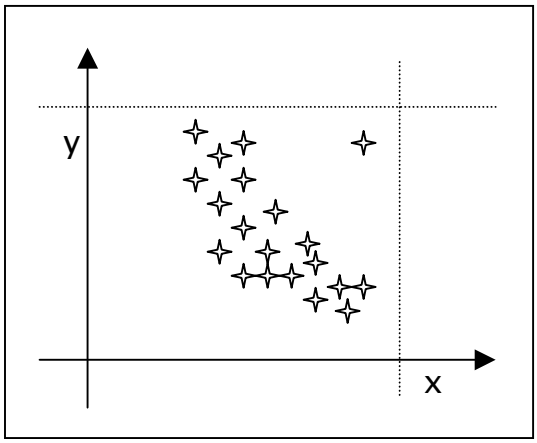


FIGURA 3.2 - Desvio Estrutural (*Structural Outliers*) [KNO 2002]

Para Parsaye, as anomalias podem ocorrer por uma série de razões, sendo que as principais são os erros nas bases de dados de grande porte (Parsaye *apud* [FEL 97]).

As principais causas de anomalias são:

a) Erros de aplicativos;

b) Falhas humanas (digitação);

c) Fraudes;

d) Casos de ocorrências raras (informações importantes para o especialista do domínio);

As duas primeiras causas apresentadas envolvem problemas com a qualidade dos dados. Nem sempre os desvios são informações que determinam uma atitude intencional, em muitos casos podem ser simplesmente falhas na qualidade dos dados. Este tipo de desvio nem sempre é importante para o especialista do domínio. As fraudes, neste caso apresentadas como erro de qualidade de dados, demonstram um erro na consistência do sistema, que gerou estas informações. No decorrer deste trabalho a fraude será tratada como um desvio de grande importância para a descoberta de conhecimento.

3.2 Detecção de desvios

Detecção de desvios é um amplo assunto para estudos. A crescente importância da detecção e posterior avaliação dos desvios encontrados, gera uma busca por uma forma mais adequada de detecção de desvios. Os trabalhos desenvolvidos são apresentados com pouca literatura sobre o assunto.

Usando diversos tipos de algoritmos de mineração de dados, os desvios podem ser identificados e considerados, mesmo que este não seja o objetivo principal do algoritmo. Estes algoritmos buscam o padrão, a partir disto é possível encontrar desvios que destoem deste padrão.

Como visto no capítulo anterior existem diversas técnicas para *data minig*, conseqüentemente existem muitas formas de detectar os desvios da base de dados. Algumas técnicas são baseadas em modelos preditivos, i.e. utilizam modelos de predição que podem determinar um padrão de comportamento dos dados. Os desvios encontrados a partir destes modelos são dados que não se enquadram no modelo pré-definido.

Outra técnica que pode ser utilizada na detecção de desvios é o agrupamento ou *clusterização*. Baseia-se na formação de *clusters* de registros semelhantes, desta forma todos os registros que não possuem as características do agrupamento formado ficam distantes do mesmo. Durante a geração dos grupos pode ocorrer a geração de agrupamentos menores e distantes do agrupamento principal. Os agrupamentos menores apresentam características diferentes do agrupamento principal, merecendo atenção, pois de acordo com a avaliação feita nos dados, estes agrupamentos podem ser considerados desvios.

Considerando-se um conjunto de n pontos de dados ou objetos, e k, o número esperado de *outliers, p*ode-se descrever os desvios em agrupamentos, como objetos k que apresentam valores máximos considerados dissimilares, excepcionais ou inconsistentes em relação aos demais dados. Nesta abordagem de trabalho, Han descreve dois problemas que ocorrem [HAN 2001]:

a) Definição de que tipo de dado pode ser considerado inconsistente em um determinado conjunto de dados;

b) Definir um método eficiente para minerar os desvios assim considerados.

A técnica de associação também pode ser utilizada. Trata-se de uma técnica simples e que pode facilmente apresentar os desvios que estão na base de dados. Como esta técnica trabalha com mecanismos para a criação de modelos que descrevem dependências entre os dados, é capaz de formular regras, a partir de ocorrências de itens em um grupo de transações, as quais tem um fator de confiança que expressa o percentual de acerto desta regra em novas transações. Avaliando-se o grau de confiança da regra é possível encontrar os desvios existentes na base de dados.

3.2.1 Análise estatística:

A estatística possui uma gama ampla de recursos para a detecção de desvios. Atualmente, a análise de detecção de desvios é feita usando técnicas estatísticas e técnicas de visualização como um processo de mineração de dados [CAB 97].

Entre as técnicas usadas para a análise tradicional de dados, a estatística é a que mais se aproxima do processo de mineração de dados, pois é possível detectar um padrão de comportamento e a partir dele reconhecer os desvios encontrados.

A aplicação das diversas técnicas de cálculo e a posterior avaliação dos resultados torna este recurso importante para este trabalho. Existem diferentes formas de avaliação dos dados, podendo ser avaliados de forma univariada ou multivariada. Para cada forma de avaliação existem variantes que podem ser consideradas.

O tipo de dado avaliado pode determinar qual o tipo de análise estatística que deve ser utilizada. Nas bases de dados trabalhadas encontram-se informações descritas em dados qualitativos e quantitativos.

O dado qualitativo determina uma representação simbólica atribuída a manifestações de evento qualitativo [PER 99]. O dado qualitativo apresenta a classificação de um fenômeno que aparentemente não poderia ser representado. Desta forma instrumentalizando o reconhecimento do evento, tanto em relação a outros eventos quando em relação as suas características. O dado qualitativo é uma quantificação do evento qualitativo, conferindo um caráter objetivo à sua observação [PER 99].

O dado quantitativo é a representação objetiva do valor a ser expresso. As informações descritas pelos dados quantitativos podem ser do tipo contínua ou discreta. De acordo com Stevenson, os dados contínuos podem assumir qualquer valor num intervalo contínuo e os dados discretos podem assumir valores inteiros sendo o resultado da contagem do número de itens considerados [STE 81].

Avaliação Univariada.

Caracteriza-se pela verificação e análise dos valores assumidos por um único atributo, o qual será avaliado isoladamente. A análise consiste basicamente na verificação dos valores apresentados, por este atributo, dentro da base de dados. Tais valores são calculados obedecendo às regras de execução da avaliação univariada utilizada, criando assim um padrão estatístico. Após calcular o padrão estatístico, a base

de dados é novamente verificada, e os desvios que destoam deste padrão são facilmente identificados.

O cálculo estatístico univariado utilizado neste trabalho é a verificação do desvio-padrão[STE 81]. Numa distribuição normal N (μ,σ^2) com dados dispostos em uma base de dados determina que a distribuição possui média μ e a variância σ^2. Os desvios (*outliers*) podem ser considerados os pontos que estão dispostos três ou mais desvios-padrão da média ($x/\ x \leq \mu + 3\sigma\ v\ x \geq \mu - 3\sigma$).

O fato de utilizar 3σ tem origem na distribuição normal, que concentra 99,73% dos valores entre $\pm 3\sigma$. Os valores além dos 3σ (maiores ou menores) apresentam baixa probabilidade de ocorrência, logo dever se investigados sempre que forem observados.

De acordo com Ross, uma base de dados que tende a infinito tende ao padrão normal de distribuição [ROS 97]. Esta afirmação é comprovada pelo teorema do limite central que determina que se $x_1, x_2,..., x_n$ é uma seqüência de valores independentes, que pertencem a uma variável aleatória, com media μ e variância σ^2. Então a distribuição de

$$\frac{X_1 + X_2 + ... + X_n - n\mu}{\sigma\sqrt{n}}$$

tende ao padrão normal se $n \rightarrow \infty$.

Por exemplo, suponha uma distribuição normal N ($10,2^2$), ou seja, $\mu=10$ e $\sigma=2$, pode-se dizer que os pontos acima de 16 e abaixo de 4 são valores atípicos (por apresentarem baixa probabilidade de ocorrência) e portanto são considerados desvios.

Análise Multivariada

A análise multivariada consiste em analisar simultaneamente mais de uma variável aleatória. Isso permite que sejam analisados os comportamentos conjuntos das variáveis. A seguir são descritos 4 métodos para a análise multivariada.

- Regressão Linear

Regressão linear consiste em avaliar o comportamento de uma variável em relação à outra variável observada. Considerando este desempenho é possível observar o padrão de comportamento das variáveis analisadas podendo com isso predizer o comportamento possível e detectar algum desvio neste comportamento.

O método de avaliação por regressão linear é baseado na equação $y = \alpha + \beta_x$. Onde α é o coeficiente linear (ponto onde a reta gerada intercepta o eixo do y) e β é o coeficiente angular da reta, isto é, a variação do y em relação à variação do x (Δ_y/Δ_x) [STE 81].

O trabalho de detectar desvios por esta técnica pode ser avaliado ao considerar-se T um conjunto de valores formados por (x_i,y_i) onde $i = 1, 2, 3,...n$ são valores do modelo. Quando, na avaliação deste modelo é possível observar valores que não se enquadram nos valores esperados, isto é, são as diferenças entre os valores observados e os valores fixados para i, tais valores são conhecidos como desvios ou resíduos e podem ser definidos por *ei*, alterando a equação, como segue $y = \alpha + \beta_x + ei$. Tais valores estão fora do padrão estabelecido e são considerados erros aleatórios, desvios ou resíduos. Os resíduos apresentam distribuição normal com média 0 e desvio padrão σ. Quando o valor de algum resíduo ultrapassar 3 σ então esse valor pode ser considerado um desvio.

Esta afirmação é confirmada por Draper e Smith que escreveram que, sendo os valores resíduos independentes e a com distribuição normal, pode-se afirmar com bases heurísticas, que os desvios estão dispostos 3 ou mais desvios-padrão da média dos resíduos.(Draper e Smith *apud* [KNO 2002]).

- Qui-quadrado

Outra técnica que pode ser utilizada para a verificação dos valores atípicos de variáveis categóricas da base de dados é o cálculo do qui-quadrado (χ^2). Consiste na análise estatística da diferença entre os valores selecionados (observados, *o*) e os valores esperados (*e*), baseados no relacionamento das variáveis [STE 81]. A equaçãoabaixo mostra a forma de cálculo do χ^2.

$$\chi^2 = \sum \left[\frac{(o-e)^2}{e} \right]$$

O χ^2 possibilita a verificação de r por k, onde k é o número de colunas e r são as categorias, assim as populações avaliadas são tratadas como multinomiais[4].

O objetivo deste cálculo é verificar as proporções das amostras e avaliar e distinguir entre amostras de populações com proporções iguais e amostras de populações com proporções diferentes. Inicialmente são definidas duas hipóteses de trabalho.

H_0 – As proporções populacionais são todas iguais;

H_1 – As proporções populacionais não são todas iguais;

O passo seguinte é a verificação dos graus de liberdade da tabela gerada. Os graus de liberdade determinam a forma de distribuição qui-quadrado, refletindo o tamanho da tabela utilizada. O valor resultante da análise dos valores gerados pelo teste

[4] Distribuição multinomial apresenta diversas categorias possíveis de resposta.

χ^2 e o valor crítico com (r-1) (k-1) graus de liberdade determina qual hipótese está correta.

O valor crítico pode ser obtido calculando-se a integral da função da distribuição qui-quadrado, com os graus de liberdade adequados, entre os pontos x e + ∞ ou verificando-se na tabela qui-quadrado.

Função:

Define-se a estatística

$$\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$$

de tal forma que se forem retiradas amostras de tamanho n de uma população normal, com desvio padrão σ, e se, para cada amostra for calculado o valor de χ^2, pode-se obter uma distribuição amostral desses valores. Essa distribuição é chamada de distribuição qui-quadrado e é dada por

$$Y = Y_0\left(\chi^2\right)^{\frac{1}{2}(\upsilon-2)} e^{\frac{-1}{2}\chi^2} = Y_0 \chi^{\upsilon-2} e^{\frac{-1}{2}\chi^2}$$

onde υ=n-1 grau de liberdade e Y_0 é uma constante dependente de υ, de modo que a área total sob a curva será 1.

A avaliação do resultado mostra se as amostras utilizadas apresentam um valor fora do normal para a situação. De acordo com a hipótese confirmada e a situação analisada é possível definir se os dados apresentam um desvio de comportamento das amostras.

Este cálculo é válido até um máximo de 30 graus de liberdade. Além deste limite, deve-se utilizar outra técnica.

- Análise de correspondência

Análise de Correspondência é uma técnica estatística multivariada para dados categóricos, de caráter gráfico, onde as posições de pontos correspondentes a variáveis ou categorias das mesmas podem ser interpretadas como associações [PER 99].. Com grande aplicação em Ciências Sociais, Marketing e Psicologia, esta técnica atualmente tem sido bastante aplicada em problemas complexos em Ciências da Saúde para análises que envolvem grande número de variáveis categorizadas e/ou qualitativas.

Pode ser utilizada quando os valores calculados para o χ^2 forem válidos, é utilizada para mostrar graficamente a relação entre as variáveis avaliadas.

- Análise de resíduos em tabela de contingência

A tabela de contingência registra as freqüências de ocorrências segundo uma dupla classificação, representada pelas duas variáveis consideradas [PER 99]. As variáveis consideradas na avaliação são selecionadas de acordo com o objetivo da análise que será realizada. A análise da tabela de contingência é feita sob a suposição de independência das variáveis usando para isso a estatística qui-quadrado. Os resíduos calculados são avaliados, considerando o valor apresentado.

A mesma lógica utilizada, para a detectar desvios numa análise univariada, aplica-se na avaliação dos valores encontrados para os resíduos calculados. Na tabela de contingência considera-se desvio todo o resíduo que apresentar um valor igual ou superior a 3 ou igual ou inferior a –3.

3.3 Avaliação dos desvios

Para identificar a importância de um desvio é necessário avaliá-lo cuidadosamente e determinar, através de um estudo do domínio, se a observação é ou não válida e o porque que esta observação é diferente das demais.

Um desvio, dependendo da sua origem e da veracidade do valor que expressa, pode ser excluído da base de dados. Mas como avaliar a situação e decidir sobre um procedimento de exclusão dos dados?

Em muitos casos a simples exclusão de um desvio pode acarretar um resultado errôneo para o processo de mineração. Existem algoritmos de *data mining* que simplesmente sugerem a retirada dos desvios encontrados. Na área estatística este processo de exclusão é considerado uma atividade comum, pois em algumas situações é possível excluir estes dados sem alterar o objetivo da pesquisa.

Em alguns casos a localização de um desvio pode ocorrer no início do processo de mineração. Neste caso o desvio deve ser avaliado e sua exclusão ou não, interfere no objetivo da mineração.

Na aplicação das técnicas estatísticas, citadas anteriormente, é possível determinar se um valor deve ser considerado desvio ou não. Esta afirmação é a base deste trabalho de mineração. O processo de verificação e avaliação dos desvios encontrados é basicamente tarefa do especialista do domínio, pois exige um conhecimento da base de dados trabalhada.

Os resultados obtidos que apresentam valores não reais podem influenciar os cálculos, dificultando a identificação do desvio, transmitindo uma informação equivocada ao usuário.

Na análise univariada, onde um valor é verificado por vez, pode-se observar a validade e a influência dos desvios encontrados. Durante este trabalho, foram feitos pequenos testes, os quais podem exemplificar esta afirmação.

Durante o processo de entendimento dos dados, foram efetuados cálculos utilizando a base estudada. Foram realizados cálculos do desvio-padrão para alguns

campos de forma aleatória. Em um dos cálculos foi verificada a quantidade de filhos das pacientes submetidas aos atendimentos equivalentes a partos e cesáreas.

Na avaliação dos dados foi encontrada uma paciente que estava registrada com 30 filhos. Analisando as informações contidas neste registro observou-se que esta paciente tinha 23 anos de idade. Este dado certamente é um erro. Uma mulher de 23 anos, não tem condições de ter 30 filhos, considerando-se que todos sejam filhos legítimos. Este dado foi excluído da base e foram refeitos os cálculos, os quais apresentaram valores diferentes.

Analisando o exemplo apresentado, conclui-se que a qualidade dos dados é muito importante para o trabalho de detecção de desvios, pois nas bases é possível encontrar informações que não tem consistência e/ou foram digitadas incorretamente.

Um exemplo que demonstra esta afirmação é um dado classificado como desvio que foi gerado a partir de um problema de qualidade dos dados, i.e. uma informação que está errada dentro da base de dados, não representando um dado real. Portanto este desvio, se considerado válido, pode alterar em muito a avaliação dos dados, transmitindo um conhecimento errôneo sobre a base avaliada.

Quando os desvios representam informações irrelevantes ou irreais, estes podem ser excluídos da base, pois não agregam conhecimento ao trabalho desenvolvido. Pois nem sempre um dado que destoa da grande maioria dos dados avaliados pode ser considerado erro de qualidade ou fraude. Um exemplo real é o nascimento de quíntuplos. Mesmo sendo um acontecimento raro é possível de ocorrer: em 2002 uma mulher de classe baixa deu a luz a 5 meninas na cidade de Farroupilha, RS.

Avaliar um desvio detectado estatisticamente está diretamente ligado ao conhecimento do analista de domínio e deve considerar:

a) Se a informação é ou não realmente uma observação válida;

b) Por que esta observação é diferente das demais.

c) Quanto um desvio é importante para os dados do sistema.

d) Problema de qualidade dos dados.

3.4 Exemplos de aplicações de detecção de desvios

Durante a execução deste trabalho, foi observado um problema que pode promover confusão em relação aos termos utilizados para a definição de desvios. Dependendo do autor e da situação avaliada, os mesmos termos são utilizados para definir desvios que ocorrem em situações diferentes. Este tópico pretende esclarecer alguns termos estudados, sempre considerando a situação na qual estão inseridas as definições apresentadas.

De acordo com o domínio dos dados avaliados, os desvios podem ser classificados por definições diferentes. São considerados desde a qualidade dos dados até a intrusão em uma rede de computadores.

Intrusões

Com o crescente avanço na utilização e expansão das redes de computadores é primordial o trabalho com a segurança da rede como um todo. A preocupação com a segurança focaliza o trabalho no controle permanente nos processos desenvolvidos e ao acesso à rede pelos usuários autorizados ou não. Um tipo de desvio detectado em um ambiente de rede é a intrusão.

Uma intrusão é definida pelo uso incorreto das "habilidades" definidas para os usuários de uma rede de computadores. Quando o usuário é cadastrado na rede, a intrusão pode ocorrer pela má utilização dos recursos disponíveis para este perfil de usuário. A outra forma de intrusão é definida pela invasão da rede por um intruso indesejado.

A grande utilização das redes de computadores e as novidades tecnológicas determinaram uma crescente preocupação com a segurança da rede e com os processos realizados nela. Esta realidade determinou o surgimento de uma área de pesquisa voltada a detecção de intrusão.

Este processo de detecção trabalha com o objetivo de evitar que um indivíduo qualquer tenha acesso as informações de uma rede privada, bem como um usuário não habilitado desenvolva tarefas que não são do seu perfil.

Padrões estatísticos, obtidos através de perfis dos usuários cadastrados, são utilizados para detectar comportamentos diferentes em um determinado ambiente. Para esta avaliação podem ser utilizados os arquivos de *log*.

O trabalho de detecção de intrusão deve estar sempre procurando formas novas de avaliação e acompanhamento do comportamento dos usuários da rede. Existem dois modos básicos para a detecção da intrusão [LEE 97]:

- Detecção de intrusão de mau uso *(Misuse Intrusion Detection)*: Padrões antigos de intrusões são utilizados para avaliar possíveis processos de intrusão.
- Detecção de intrusão por anomalias *(Anomaly Intrusion Detection)*: Reconhecer comportamentos que divergem do comportamento normal.

A estatística é utilizada neste tipo de detecção, pois pode produzir medidas estatísticas ou regras que representem o padrão normal de comportamento. A partir destas regras é possível fazer uma auditoria no comportamento do usuário e descobrir atividades que divergem dos padrões considerados normais. Este processo deve considerar o fato que o comportamento do usuário pode mudar dinamicamente e freqüentemente em determinados ambientes.

Fraudes

Fraudes[5] são caracterizadas pelos desvios de comportamento em ambientes que efetuam transações, que envolvam uma grande soma em dinheiro e em casos que envolvam a aplicação do dinheiro público, ou ainda nos processos de utilização de cartões de crédito. Considerando a definição deste desvio, identifica-se a sua importância para o mercado financeiro. O tipo de detecção aplicada às fraudes é uma das que mais detêm a atenção e estudos dos especialistas.

Com o aumento dos pontos de acesso ao sistema financeiro, cresce a preocupação com a segurança do sistema e dos clientes que utilizam este serviço. A fraude é um desvio característico do avanço da tecnologia disponível no mundo atual.

Para o controle de fraudes, as instituições financeiras utilizam máquinas que são capazes de aprender novos padrões de comportamento, juntamente com algoritmos de análise estatística para a avaliação do comportamento das transações efetuadas por cartões de crédito. Os algoritmos utilizados requerem uma grande quantidade de informações, as quais são utilizadas para a construção de modelos que definam transações irregulares.

Atualmente os bancos perceberam a necessidade do compartilhamento das informações, sobre os ataques ao sistema financeiro, e manutenção periódica das mesmas. Somente com um trabalho em conjunto é possível construir uma estrutura global de descoberta de fraudes [STO 97].

Esta estrutura global de descoberta encontra algumas dificuldades para a sua definição, são elas:

- As companhias financeiras não compartilham informações por diversas razões;
- Os banco de dados das companhias mantém registros de comportamentos de transações, que crescem e atualizam-se rapidamente.
- O processo executado em tempo-real demanda a atualização imediata dos modelos quando novos desvios são detectados.
- Fácil distribuição de modelos no ambiente de rede é essencial para a manutenção da capacidade de detecção.

[5] s. f. 1. Ato ou efeito de fraudar; logro, fraudação. 2. Abuso de confiança. 3. Contrabando.

4 Metodologia

Este capítulo descreve a metodologia utilizada nos estudos de caso realizado com dados da Secretaria da Saúde do Estado do Rio Grande do Sul (SES).

O objetivo é reconhecer e avaliar o comportamento padrão dos dados, com o auxílio das verificações estatísticas e a partir deste padrão, detectar e avaliar os desvios encontrados na base.

Foram utilizadas duas ferramentas para o desenvolvimento dos experimentos. A primeira é um software desenvolvido especialmente para este trabalho, cuja descrição completa está na seção 4.2. A segunda ferramenta é o software SPSS, o qual permite aplicação de diversas técnicas estatísticas. No trabalho desenvolvido a utilização do SPSS tem como objetivo principal à complementação da pesquisa.

Os experimentos seguiram a sistemática de trabalho CRISP_DM (*Cross Industry Process Model for Data Mining*)[CHA 99], por ser a mais detalhada encontrada na literatura. Ela apresenta um roteiro de execução adequado para as necessidades deste trabalho e não se restringe a ferramentas ou técnicas específicas.

CRISP_DM é descrito em termos de um modelo de processo hierárquico, constituído de um conjunto de tarefas distribuídas em 4 níveis de abstração [CHA 99], conforme apresentado na figura 4.1.

No primeiro nível o processo de descoberta de conhecimento é organizado em várias fases, sendo que cada fase é composta de tarefas genéricas.

O segundo nível agrega as tarefas genéricas, apresentas anteriormente. Neste nível são relacionadas todas as tarefas que pretendem abranger as necessidades encontradas no processo de mineração.

O terceiro nível é composto pelas tarefas especializadas, as quais descrevem as ações que foram utilizadas para a realização das tarefas definidas no segundo nível. Tais tarefas são a execução de passos específicos organizados logicamente, sendo possível, em determinadas situações, refazer algum passo anterior.

O quarto nível, chamado de instâncias do processo, consiste do registro das ações, decisões e resultados do processo de mineração de dados para uma aplicação em particular. Neste nível são descritos todos os passos que, de fato, foram realizados no processo de mineração.

Com a metodologia CRISP-DM é possível definir o mapeamento de modelos genéricos para modelos especializados utilizando os Contextos de Mineração de Dados, que possuem quatro dimensões diferentes (Chapman *apud* [DOM 2003] p.54).

1) Domínio da aplicação define a área específica da aplicação;

2) *O tipo de problema de MD*, parte dos objetivos do projeto de MD descrevendo suas classes específicas.

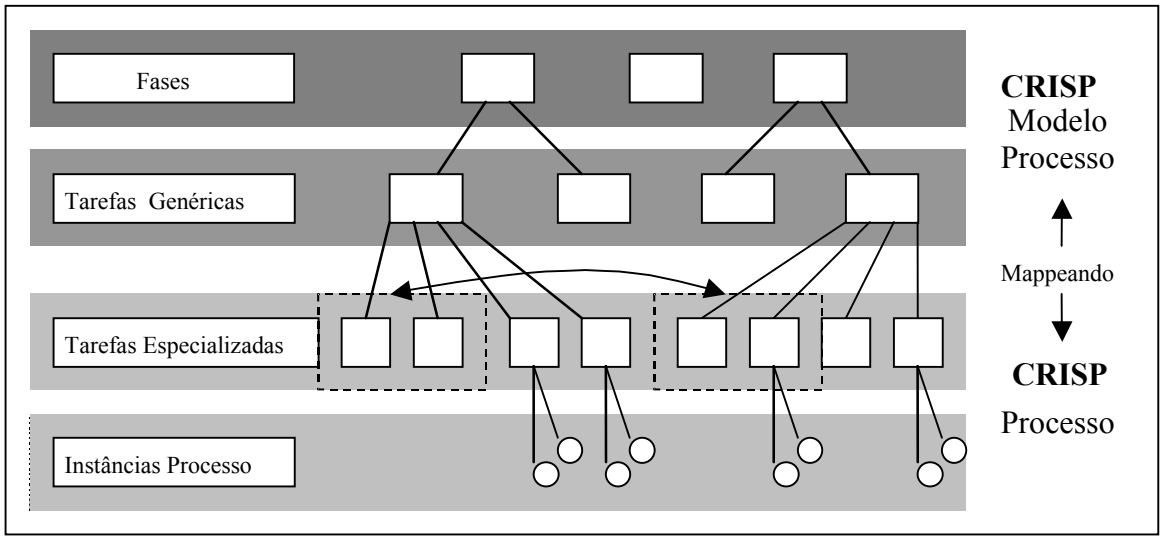


FIGURA 4.1 Metodologia CRISP_DM - 4 níveis

3) *O aspecto técnico* trata de apresentar as questões específicas em MD, descrevendo os diferentes problemas técnicos que podem ocorrer durante o processo.

4) As *ferramentas e técnicas* definem quais as ferramentas usadas, bem como as técnicas aplicadas durante o processo de MD.

Durante o desenvolvimento deste estudo foram utilizadas duas bases de dados diferentes. As bases utilizadas são:

a) Autorização de Internação Hospitalar (AIH);

b) Registro das declarações de óbitos - Mortalidade.

Maiores informações e o detalhamento sobre as bases utilizadas estão descritos no capítulo seguinte.

Em ambos os casos o objetivo definido era o mesmo: - avaliar o comportamento padrão dos dados e detectar desvios deste comportamento. O contexto de mineração de dados, para este trabalho é parametrizado, como segue:

1) Domínio da aplicação: Reconhecimento do padrão de comportamento existente nos dados avaliados e a identificação de desvios a partir deste padrão.

2) O tipo de problema de MD. Seleção dos dados e análise estatística, envolvendo a avaliação dos dados quantitativos e qualitativos.

3) O aspecto técnico: Cálculos estatísticos, aplicando as técnicas de avaliação do desvio-padrão, qui-quadrado, análise de resíduos e análise de correspondência.

4) Para a seleção, preparação e pré-avaliação dos dados é utilizado o protótipo desenvolvido para este estudo. Para uma avaliação mais detalhada utiliza-se o software de cálculo estatístico SPSS.

O detalhamento da metodologia utilizada para a avaliação dos dados obedece às fases CRISP-DM e descreve o ciclo de vida de um projeto de MD em uma seqüência de seis fases. Como citado anteriormente, esta seqüência não é rígida, permitindo que algum passo seja refeito ou até mesmo não executado, se for esta a necessidade do processo.

4.1 Fases CRISP_DM

4.1.1 Entendimento do negócio

O entendimento do negócio descreve o estudo do problema e identificação dos objetivos do usuário sobre a DCBD e a preparação de um plano inicial para o projeto. Nesta fase é necessário, se possível, o envolvimento dos analistas que trabalham com os dados que serão avaliados. Todo o conhecimento apresentado por estes profissionais é importante para a definição do plano de trabalho.

Nesta fase também é feita a análise dos recursos necessários para o desenvolvimento do projeto, as limitações devem ser avaliadas e a busca por soluções deve ser efetivada.

A técnica que será utilizada na avaliação deve ser definida, e as limitações de sua aplicação devem ser consideradas. Através da literatura existente, é possível definir qual técnica se enquadra melhor nos objetivos traçados para o trabalho.

Para este estudo foi definida a utilização de algumas técnicas estatísticas, por permitirem diferentes formas de avaliação para os dados envolvidos. A seqüência da aplicação das técnicas é definida de acordo com os resultados obtidos nas primeiras análises feitas dos dados.

A escolha da técnica estatística utilizada é definida pelo tipo de dado que será avaliado (quantitativo ou qualitativo) e qual o tipo de avaliação que deve ser feita (univariada ou multivariada).

4.1.2 Entendimento dos dados

O trabalho desenvolvido nesta fase é determinado pelo estudo dos dados envolvidos no projeto. Este estudo se inicia com a coleta, descrição, exploração dos dados e finaliza com a verificação da qualidade dos mesmos. Cada etapa do trabalho visa a compreensão de como e qual estado se encontra a base de dados trabalhada.

De acordo com o estado da base e o tipo de dado armazenado, nem sempre é possível desenvolver o trabalho desejado, em alguns casos é necessário reestruturar os dados, para que o trabalho possa ser continuado. No desenvolvimento deste trabalho é possível citar um exemplo prático: a técnica estatística análise de correspondência calculada no SPSS só trata de dados quantitativos e a base trabalhada apresenta 80% dos

valores sendo dados qualitativos. Este problema foi possível de solução, mas em certas circunstâncias os dados podem não ser possíveis de modificação.

O bom entendimento dos dados permite avaliação dos passos e do tempo necessário para a etapa seguinte. Logo, quanto maior a compreensão sobre a base trabalhada, mais fácil à definição do trabalho necessário para a fase seguinte.

O roteiro de alteração dos dados deve obedecer às limitações apresentadas pelos softwares selecionados, pois existem diversos algoritmos para a mineração de dados, sendo que cada um deles contém particularidades de utilização, necessitando de uma preparação especial dos dados. Somando a esta realidade os dados selecionados para a análise podem apresentar diferentes estruturas e formas de armazenamento, tornando praticamente impossível encontrá-los em condições ideais para a mineração, sendo necessário uma boa definição das atividades que serão desenvolvidas para a preparação destes dados.

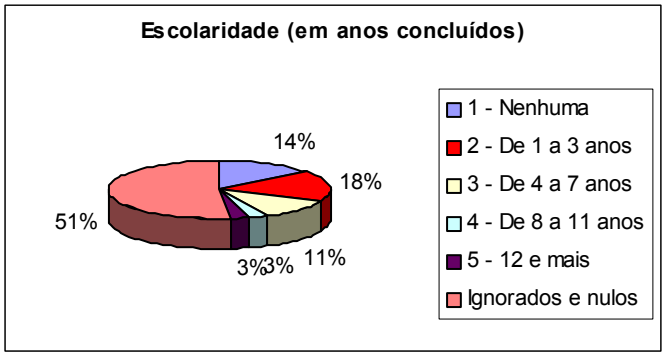


FIGURA 4.2 - Atributo escolaridade - Arquivo Mortalidade

O entendimento dos dados pode ser desenvolvido através da utilização de ferramentas simples que apresentam as condições dos dados selecionados. Estas avaliações podem mostrar como estão dispostos os atributos nos registros. Apresentando a quantidade de atributos nulos ou em branco, no caso da análise estatística, tais atributos podem influenciar nas avaliações posteriores. Como exemplo, a figura 4.2 mostra um gráfico que demonstra os atributos identificados para o campo escolaridade, do arquivo de mortalidade.

Analisando o gráfico pode-se identificar que mais de 51% dos registros são nulos ou ignorados. Desta forma um estudo sobre este campo pode não ser satisfatório para o trabalho, pois não expressa a real situação dos pacientes.

Outro tópico importante é a qualidade dos dados a ser avaliada. Em muitos casos as bases, por serem muito grandes, apresentam ruídos ou falhas nos registros. Esta avaliação permite a correção de alguns problemas que possam ocorrer, como a recuperação de partes dos arquivos que possam estar danificadas.

4.1.3 Preparação de dados

Nesta fase o trabalho é aplicar todo o conhecimento adquirido com o estudo do problema e dos dados. Esta tarefa não é simples e demanda grande parte do tempo despendido no trabalho de mineração. Devem ser descritas todas as atividades necessárias para o preparo dos dados às limitações dos softwares utilizados no processo. Em muitos casos estas atividades devem ser refeitas, até que os dados estejam prontos para a análise.

A primeira etapa da preparação dos dados é seleção dos mesmos. A qual deve apoiar-se nos objetivos definidos, no conhecimento adquirido na fase de entendimento dos dados e nas limitações das ferramentas utilizadas para o projeto. Para um resultado satisfatório e completo é necessário considerar:

a) Requisitos de tempo e espaço;

b) Simplicidade do modelo gerado;

c) Relevância dos atributos;

d) Redundância entre os atributos

A etapa seguinte envolve a limpeza dos dados selecionados. As tarefas descritas nesta etapa consideram a exclusão dos dados errôneos, a padronização de dados que expressem valores diversos, a eliminação de informações duplicadas e o tratamento de valores ausentes.

Analisando o estudo realizado, esta etapa exigiu bastante cuidado, pois como o objetivo deste trabalho é detectar desvios nas bases de dados, a atividade de limpeza não podia simplesmente desconsiderar os dados que estavam divergentes dos demais. O que poderia ocasionar uma distorção dos resultados obtidos ou a ausência de um resultado satisfatório. É muito importante a sensibilidade do analista de domínio quanto aos objetivos da mineração.

A última etapa trata da construção dos dados para a fase seguinte. Os procedimentos efetuados devem ser registrados para que possam ser repetidos ou alterados quando necessário. Durante a construção dos dados, podem ser realizados processos de normalização, transformação de dados qualitativos em dados quantitativos e a discretização dos atributos.

Para as bases utilizadas neste trabalho foram necessárias adequações quanto à qualidade dos dados, a ausência de atributos nos registros, a padronização de faixas para os atributos encontrados e a montagem dos arquivos que seriam minerados. A seqüência de tarefas desenvolvidas está descrita, em detalhes, no capítulo seguinte.

4.1.4 Modelagem

Esta fase corresponde à execução dos algoritmos de descoberta e é denominada de mineração por muitos autores.

Neste trabalho são utilizadas duas ferramentas para a avaliação dos dados, existindo uma ordem execução. Inicialmente é utilizado o protótipo desenvolvido para este trabalho, o qual possibilita a seleção dos dados e as primeiras avaliações estatísticas.

A utilização do SPSS depende do roteiro de avaliações que o usuário pretende seguir. De acordo com os resultados obtidos, no passo anterior, são realizadas no SPSS mais duas avaliações estatísticas. Para uma melhor compreensão do trabalho realizado, as etapas realizadas e os softwares utilizados serão apresentados separadamente, nos itens 4.2 e 4.3 deste capítulo.

4.1.5 Avaliação

Os modelos criados são satisfatórios para a proposta do trabalho. Deve-se ressaltar que nem todos os modelos vão apresentar a mesma seqüência de análise, isto porque os dados avaliados, por experimento, podem apresentar situações diferentes.

4.1.6 Aplicação

Os valores resultantes das análises devem ser apresentados de forma acessível ao usuário, facilitando o entendimento dos resultados. O protótipo apresenta as informações claras, permitindo o entendimento por parte do usuário. O SPSS, já exige um pouco mais de atenção do usuário. Requer do usuário um pequeno esforço de interpretação dos valores apresentados.

4.2 Descrição do protótipo

Este software foi desenvolvido para a avaliação dos dados obtidos pela SES devido às características dos dados utilizados na análise.

O problema reside em aplicar diretamente as técnicas de detecção de desvios sobre uma base de dados. Muitos softwares disponíveis apresentam detecção de desvios como um apêndice da mineração de dados.

Os dados fornecidos pela SES apresentavam determinadas características de formatação, as quais não facilitavam o trabalho da análise de desvios. Houve a necessidade de preparar os dados para avaliações estatísticas mais detalhadas para se obter um resultado satisfatório.

A necessidade de aplicação direta originou este protótipo, o qual tem a intenção de focalizar o objetivo da análise diretamente sobre a detecção de desvios.

O software parte da premissa de selecionar os dados para uma análise mais focada em objetivos pré-definidos e permitindo uma avaliação mais detalhada dos dados trabalhados.

A aplicação do algoritmo define dois tipos de avaliações possíveis: a avaliação dos dados quantitativos e a avaliação dos dados qualitativos. Baseando-se nestas duas abordagens, foram programadas técnicas estatísticas para estas análises.

- Desvio-padrão: Análise univariada, utilizada para a avaliação dos dados quantitativos.

- Qui-quadrado: Análise multivariada, utilizada para a avaliação dos dados qualitativos.

As duas técnicas estatísticas utilizadas apresentam os resultados de forma simples e clara para o usuário.

O protótipo foi adequado para preparar os dados para outra técnica estatística, permitindo a seleção e a preparação dos dados para a análise de correspondência, que é calculada através do software SPSS.

O protótipo desenvolvido pode ser utilizado como parte do processo de descoberta de conhecimento. Sendo útil para selecionar e minerar parte dos dados que compõem um grande banco de dados.

A sua facilidade está na simplicidade de selecionar uma determinada parte dos dados e permitir a preparação destes para serem aplicados em outros softwares de análise estatística.

Os dados utilizados devem ser preparados para a utilização do software, seguindo um roteiro descrito no capítulo 5. Para trabalhar com outra fonte de dados o protótipo deverá ser reformulado. Sobre estas bases o protótipo permite qualquer combinação de campos para posterior avaliação dos resultados obtidos.

4.2.1 Requisitos do Sistema

O protótipo pode ser aplicado a qualquer ambiente de trabalho, que tenha como sistema operacional Microsoft Windows.

A linguagem Delphi foi utilizada para desenvolvimento do protótipo por se tratar de uma linguagem simples e de fácil manuseio, permitindo uma boa interface com o usuário.

4.2.2 Uso do protótipo

As etapas que compõem a modelagem dos dados no protótipo estão descritas abaixo de forma clara e objetiva.

Passo 1: Seleção da base de dados

O trabalho se inicia com a seleção da base de dados que será avaliada, o protótipo apresenta uma interface amigável com o usuário, possibilitando a escolha direta da base, como mostra a figura 4.3.

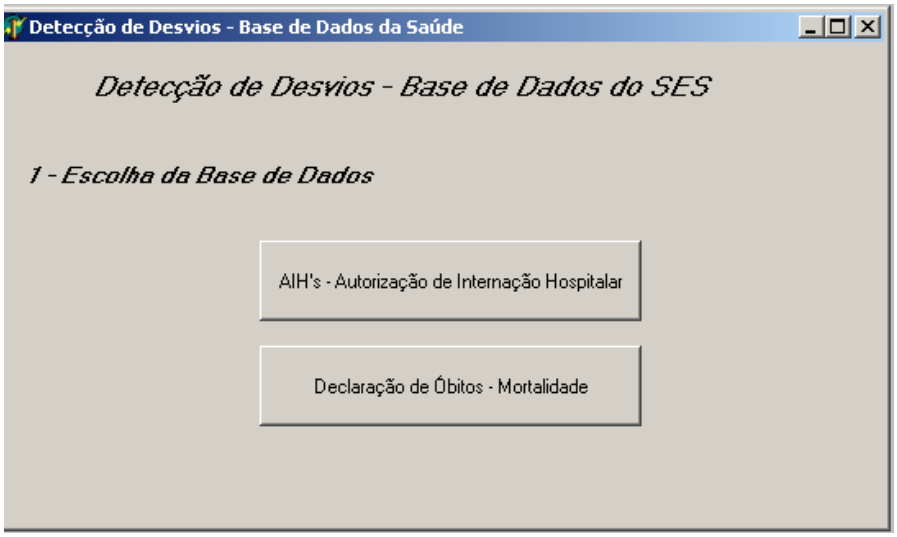


FIGURA 4.3 -Escolha da Base de Dados

Passo 2: Seleção do tipo de análise

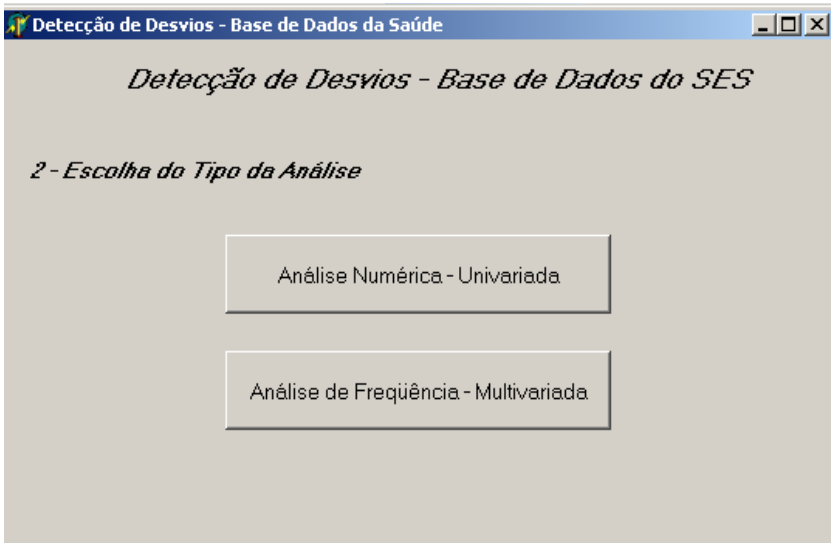


FIGURA 4.4 – Escolha da análise estatística

O passo seguinte é definir o tipo de avaliação que será desenvolvida, as avaliações estão definidas a partir do tipo de dado avaliado. Para a avaliação dos dados quantitativos é selecionada a análise numérica e a avaliação de dados qualitativos é a escolha da análise de freqüência, como pode ser visto na figura 4.4.

Passo 3: Definição dos campos

A partir deste ponto o protótipo permite a definição do campo que será utilizado para selecionar os dados que serão avaliados. Na mesma tela, obedecendo à análise estatística definida anteriormente, são selecionados os campos utilizados para a verificação estatística.

Para a seleção dos dados, o protótipo apresenta os nomes dos campos que armazenam informações qualitativas e permite ainda a seleção dos registros pela categoria da doença escolhida, i.e. informando o Código Internacional de Doenças (CID). Assim o usuário pode escolher se pretende selecionar os dados pelos atributos cadastrados no campo ou simplesmente informar o CID da doença escolhida.

Com esta facilidade a seleção dos registros em ambos as bases selecionadas é mais rápida, pois tanto a base das AIH's quanto à base da mortalidade estão diretamente relacionadas com o código das doenças.

Como os registros de CID são os mesmos para as duas bases utilizadas, adotou-se a padronização, das categorias das doenças, para a utilização das duas bases de dados. Neste caso quando for solicitada a seleção da base pelo código do CID, o protótipo selecionará as doenças de acordo com a tabela 4.1, apresentada abaixo.

TABELA 4-1 - Categorias de doenças

Faixa de valores	Descrição
A00 - B99	Algumas doenças infecciosas e parasitárias
C00 - D48	Neoplasias [tumores]
D50 - D89	Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários
E00 - E90	Doenças endócrinas, nutricionais e metabólicas
F00 - F99	Transtornos mentais e comportamentais
G00 - G99	Doenças do sistema nervoso
H00 - H59	Doenças do olho e anexos
H60 - H95	Doenças do ouvido e da apófise mastóide
I00 – I99	Doenças do aparelho circulatório
J00 - J99	Doenças do aparelho respiratório
K00 - K93	Doenças do aparelho digestivo
L00 - L99	Doenças da pele e do tecido subcutâneo
M00 - M99	Doenças do sistema osteomuscular e do tecido conjuntivo
N00 - N99	Doenças do aparelho geniturinário
O00 - O99	Gravidez, parto e puerpério
P00 - P96	Algumas afecções originadas no período perinatal
Q00 - Q99	Malformações congênitas, deformidades e anomalias cromossômicas.
R00 -R99	Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte
S00 - T98	Lesões, envenenamento e algumas outras conseqüências de causas externas
V01 - Y98	Causas externas de morbidade e de mortalidade
Z00 - Z99	Fatores que influenciam o estado de saúde e o contato com os serviços de saúde

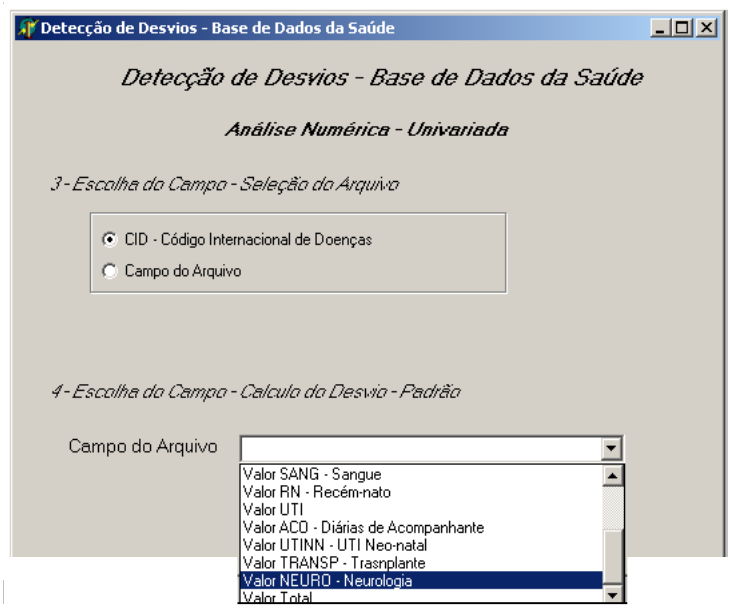


FIGURA 4.5 - Seleção do campos para análise univariada

De acordo com a análise estatística definida a segunda parte da tela pode se apresentar de dois modos distintos, como segue:

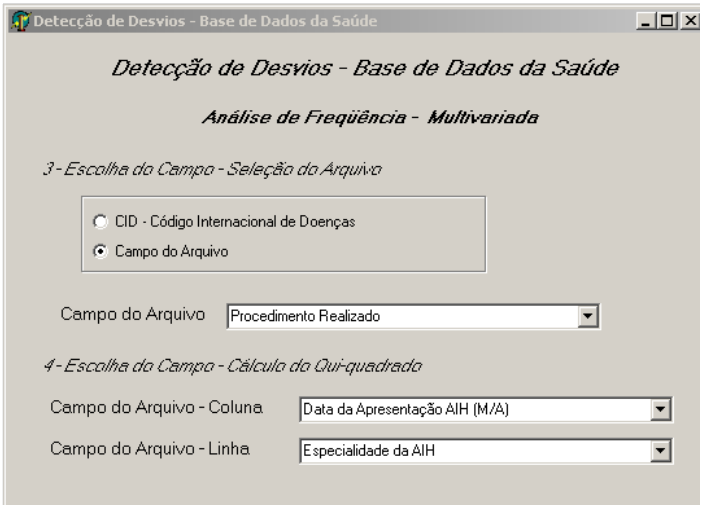


FIGURA 4.6 – Seleção de campos para análise multivariada

- Análise numérica – univariada: permite a seleção de um campo quantitativo para a avaliação estatística. Como pode ser visto na figura 4.5.

- Análise freqüência – multivariada permite a definição de dois campos qualitativos para o cálculo do qui-quadrado: um campo referente à coluna e um campo referente à linha. A tela pode ser observada na figura 4.6

Passo 4: Entrada dos atributos de seleção

Efetuado a parametrização necessária, o protótipo iniciará o recebimento dos atributos de acordo com o campo escolhido para a seleção do arquivo que será analisado. O usuário encerra a entrada de valores com a confirmação para executar os cálculos.

Passo 5: Cálculo estatístico

Realização dos cálculos estatísticos e a apresentação dos resultados. Para cada análise efetuada é apresentada uma tela diferenciada.

Para a análise numérica são apresentados os valores:

a) Média Aritmética (μ):

b) Desvio-padrão (σ):

c) Valor resultante da soma de 3 desvios-padrão e a média calculada. ($\mu + 3\sigma$):

O valor para ($\mu - 3\sigma$) não foi utilizado na avaliação, pois nos testes realizados as respostas encontradas não foram relevantes.

A análise numérica encerra com a criação de um arquivo resposta que contém os registros com valores acima de 3 desvios-padrão. Tais valores são os desvios da base avaliada.

Para a análise de freqüência, que realiza o cálculo do qui-quadrado, são apresentados os valores de:

a) *Pearson* qui-quadrado:

b) Graus de liberdade:

O programa possibilita a geração de um arquivo simplificado e codificado para ser utilizado no software SPSS.

Desta fase em diante a detecção de desvios continua com a utilização do SPSS.

4.3 Software Estatístico do SPSS

Esta seção descreve sucintamente o uso do SPSS nos experimentos realizados. Para maiores detalhes sobre o software consulte [PER 2001]

Passo 1: Entrada e definição dos dados

Ao iniciar a utilização do SPSS, é preciso abrir a base de dados preparada pelo protótipo. O software reconhece os arquivos tipo DBF (*Data Base File*), isto facilita a aplicação da ferramenta. A figura 4.7 apresenta a tela de entrada de dados e destaca a planilha *Variable View.*

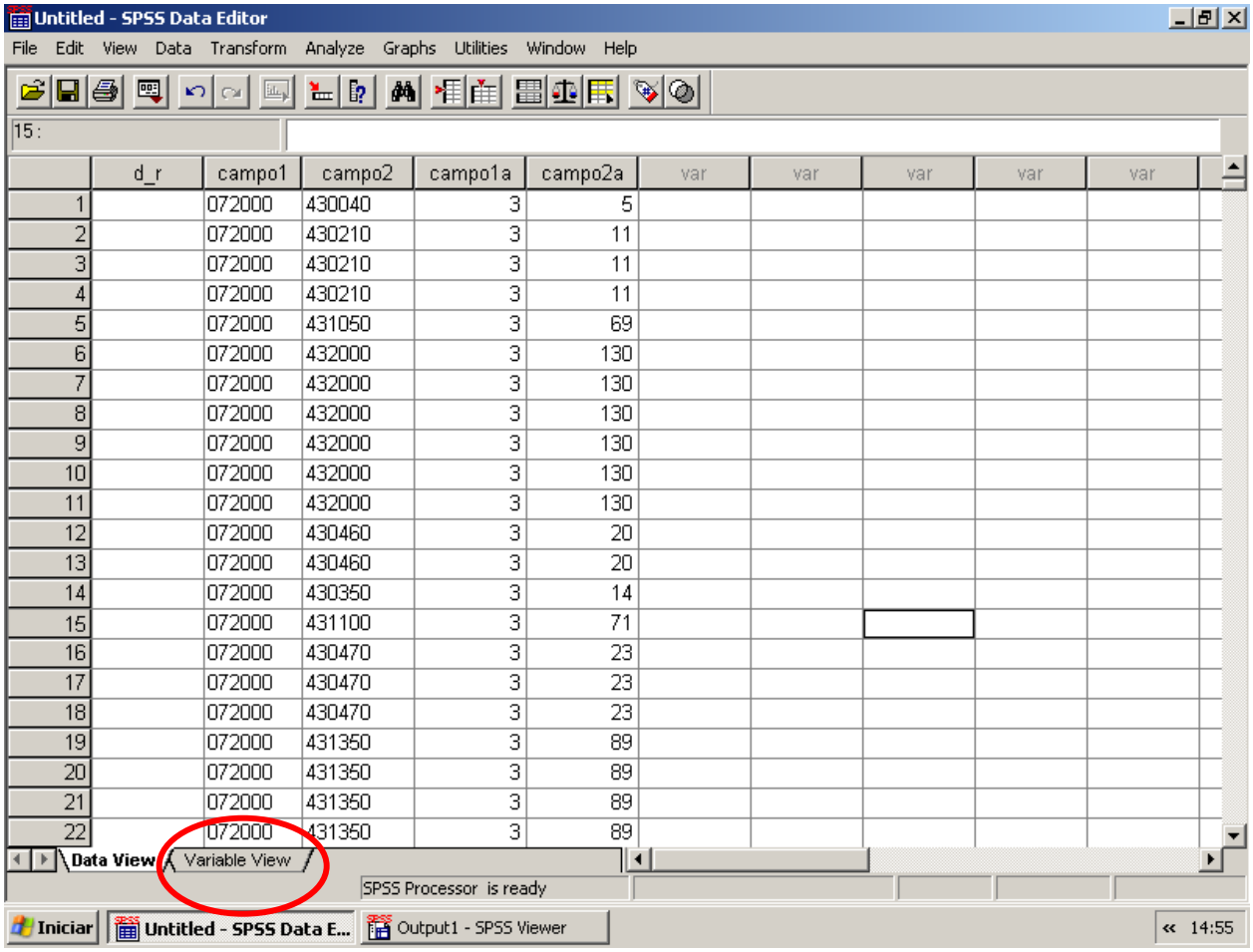


FIGURA 4.7 - Entrada de dados SPSS

Para facilitar a identificação dos campos, o SPSS permite a alteração do nome dos campos, pois o protótipo define nomes genéricos para os campos selecionados. Este recurso é possível utilizando a planilha *Variable View*. A figura 4.8 apresenta a alteração dos nomes dos campos. O campo que possui o nome geral "campo1a" é renomeado para "dta_apre".

Passo 2: Definição da técnica estatística

A seqüência de avaliações estatísticas é definida pelos resultados obtidos durante a avaliação. Este passo vai mostrar como são parametrizadas as técnicas utilizadas no SPSS.

Passo 2.1: Análise de Correspondência.

A verificação dos dados através da análise de correspondência é indicada quando os graus de liberdade verificados no cálculo do qui-quadrado não forem altos, pois a análise de correspondência, como foi citado anteriormente, é uma visualização gráfica dos resultados do qui-quadrado. A figura 4.9 apresenta a tela de configuração da análise

de correspondência. Na janela menor, observa-se a lista de valores codificados para o campo dta_apres.

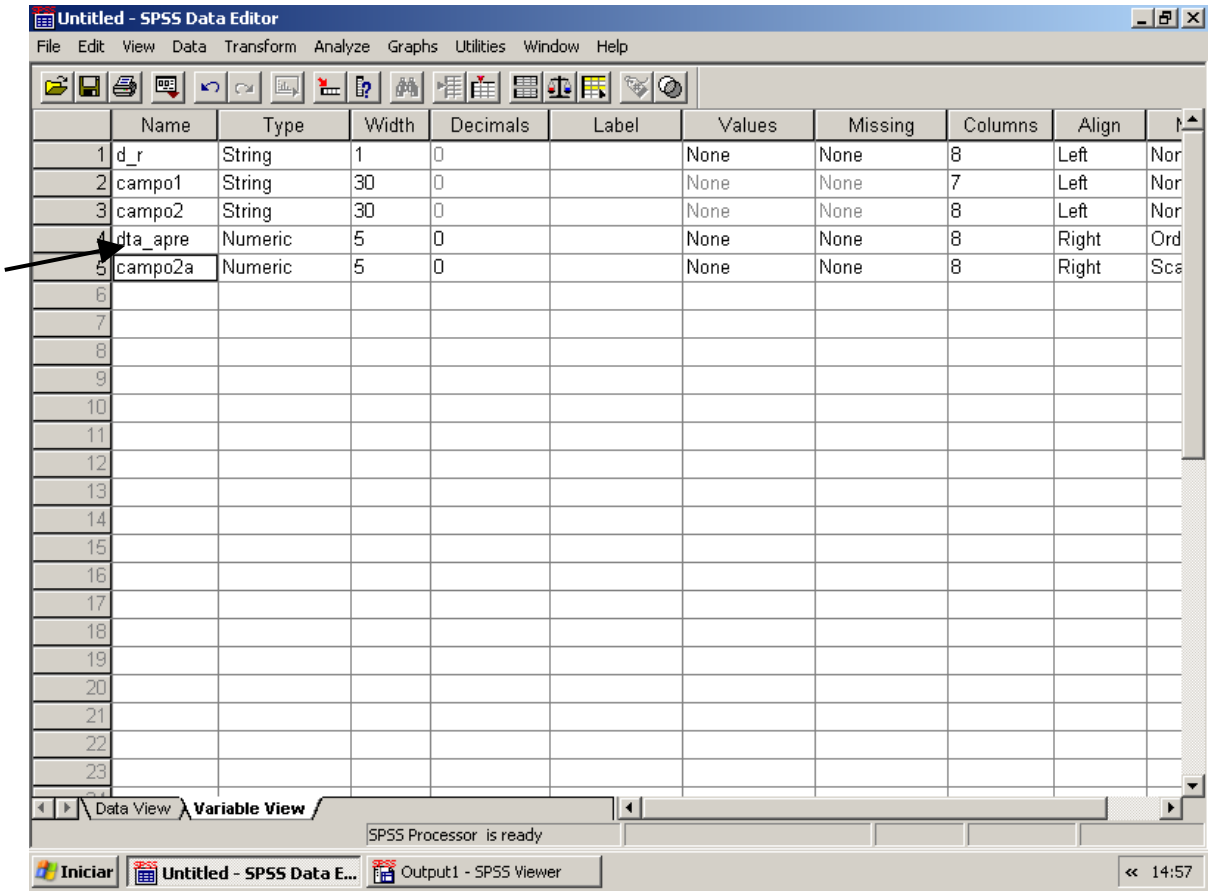


FIGURA 4.8 -Alteração do nome do campo

Passo 2.2: Análise de Resíduos

A análise de resíduos é um recurso que pode ser utilizado com mais freqüência, pois faz uma verificação entre os valores esperados e observados das células que informam a freqüência dos valores cruzados. Mesmo seguindo a mesma lógica de cálculo de resíduos do qui-quadrado, pode ser utilizado para localizar os desvios da base trabalhada.

A figura 4.10 apresenta a tela de configuração utilizada para a análise dos resíduos. Na mesma figura observa-se a solicitação de registro dos resíduos ajustados.

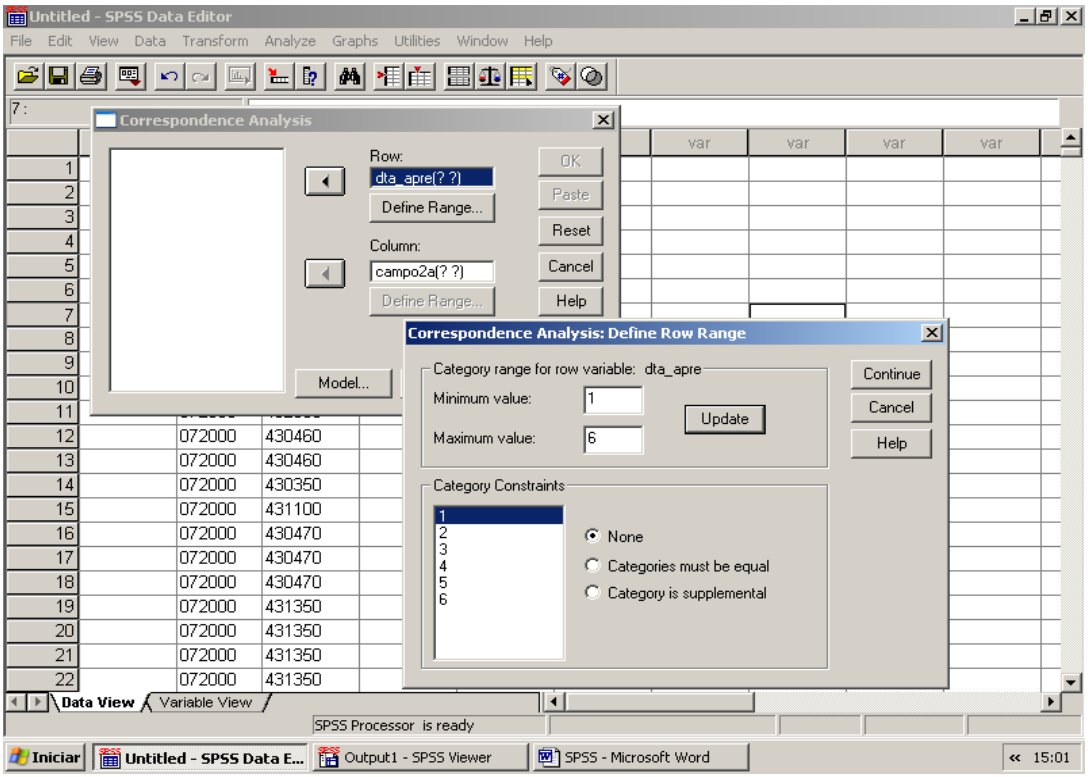


FIGURA 4.9 - Configuração - Análise de Correspondência

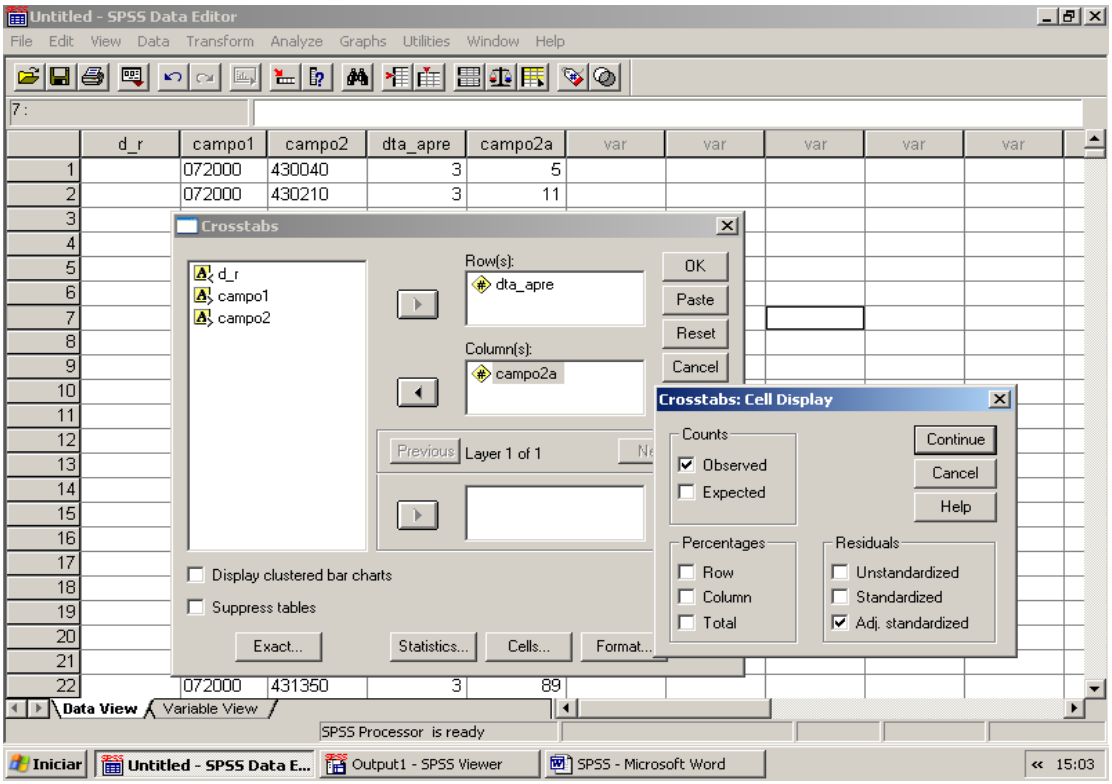


FIGURA 4.10 - Configuração - Análise de correspondência

5 Experimentos Realizados

A parte experimental deste trabalho foi realizada no contexto do projeto Desenvolvimento de Metodologia para a Extração de Conhecimento de Bases de Dados de Saúde do Estado para Avaliação e Planejamento, financiado pela FAPERGS.

Foram realizados dois estudos de caso, um com os dados das Autorizações de Internações Hospitalares (AIH) e o outro com os dados referentes aos registros de óbitos. Os dados armazenados nas bases possuem algumas semelhanças, mas o trabalho de entendimento e preparação dos dados foi executado de forma distinta.

Visando o entendimento do trabalho realizado, os estudos de caso foram descritos individualmente, seguindo a metodologia CRISP-DM, descrita no capítulo 4.

5.1 Estudo de Caso – Base das AIH's

Este estudo tem como base as AIH's do período de maio de 2000 a janeiro de 2001. As informações foram cedidas pela Secretaria da Saúde do Estado do Rio Grande do Sul (SES).

5.1.1 Entendimento do negócio

Os dados trabalhados armazenam as informações registradas nos hospitais, quando da internação dos pacientes. Tais informações são cadastradas de acordo com uma padronização estabelecida pelo Governo Federal. As normas utilizadas neste sistema estão definidas em relatórios, normas, legislações, manuais do SIH/SUS e relatórios de atividade da SES/2000, emitidos pelo governo, os quais tem o objetivo de orientar os profissionais da área de saúde.

O objetivo deste trabalho é encontrar um padrão de comportamento das informações estudadas e, a partir deste padrão, detectar informações que destoam do mesmo. Para atingir este objetivo foram aplicadas técnicas estatísticas, as quais permitiram uma boa avaliação dos dados.

As técnicas aplicadas promovem uma avaliação simples das informações contidas nas bases, mas os resultados obtidos foram satisfatórios para a avaliação do trabalho. A escolha das técnicas tem o objetivo de demonstrar passo a passo o trabalho de detecção de desvios.

Grande parte do trabalho não foi realizado no Instituto de Informática da UFRGS, o que limitou os recursos disponíveis. A base de dados foi cedida pela SES, e preparada em parte, no laboratório da UFRGS. O protótipo criado para este trabalho, foi desenvolvido em ambiente particular e o software SPSS, foi utilizado na Universidade de Caxias do Sul (UCS).

Houve algumas limitações, são elas: não se deve utilizar computadores com hardware inferior à: Pentium 3, com 256 MB de RAM e HD de 20 GB. O motivo é a grande quantidade de dados utilizada e o software SPSS.

O sistema operacional mínimo, para instalação dos softwares utilizados é o Windows XP. Neste sistema operacional o software SPSS, libera todas as possibilidades de avaliações.

5.1.2 Entendimento dos dados

Para o registro das informações coletadas, em todo o território nacional, existe uma metodologia de trabalho empregada nas definições e avaliações das AIH's. Este trabalho tende a padronizar as informações cadastradas e gera grandes bases de dados. Estas bases possuem um grande número de registros e estes registros possuem um grande número de campos.

Os arquivos iniciais foram fornecidos de acordo com a tabela 5.1, apresentada a seguir. Inicialmente foram estudados e avaliados separadamente. De acordo com a necessidade da pesquisa a ser realizada, estes dados foram preparados de formas distintas.

Durante a verificação dos dados, foram observados que os valores quantitativos, estavam quase que totalmente padronizados. Ocorrências médicas que registrassem os mesmos procedimentos apresentavam os mesmos valores gastos. Esta característica dificultaria a avaliação dos dados quantitativos, tendo em vista o objetivo deste trabalho.

TABELA 5-1 - Descrição dos arquivos

Arquivo	Descrição	N° de objetos	N° de atributos
DSMS010	Movimento das AIHs.	375.408	75
DSMS020	Procedimentos especiais autorizados de AIH nos municípios do Estado no período.	86.905	9
DSMS030	Atos profissionais autorizados de AIH nos municípios do Estado no período.	2.250.372	14
DSMS040	Movimento dos hospitais com informações de lançamentos (pagamentos e descontos).	392	9
DSMS160	Valores da AIH (faturamentos cobrados).	375.408	24
DAIH050	Tabela de Atos.	4.899	15
DAIH150	Diagnósticos de acordo com a tabela CID.	14.196	6
CONTROLE	Controle anual dos registros bloqueados.	14.282	12
BUREAU	Hospitais distribuídos por Bureau	256	2
LEITOS	Hospitais distribuídos pelo número de leitos.	379	5

5.1.3 Preparação dos dados

Na preparação dos dados, não foi necessário alterar o tipo dos arquivos fornecidos, pois as ferramentas utilizadas identificam os arquivos DBF.

As avaliações efetuadas nos arquivos necessitavam que as bases utilizadas apresentassem uma grande quantidade de campos. Desta forma a preparação de uma base para a pesquisa proposta foi definida a partir da união dos dados de dois arquivos: Movimento das AIH's (DSMS160) e Valores das AIH's (DSMS010).

O arquivo criado foi reavaliado, pois alguns campos não continham informações relevantes à pesquisa. Foram excluídos campos não preenchidos, campos de identificação do paciente e outros campos considerados desnecessários pelos usuários. Foi adicionado um campo, para registrar o total gasto pela AIH em cada registro armazenado.

Dos campos relativos aos pacientes, foram preservados os que continham informações sobre a cidade, sexo, idade e data de nascimento. Ao final do processo o arquivo apresentava 38 campos, cujos nomes e descrições podem ser observados na tabela 5.2.

TABELA 5-2 - Campos do registro - Arquivo AIH

Nome do Campo	Significado	Tipo	Faixa de valores
DCIH	Documento para cobrança de Internação Hospitalar	Texto(8)	
APRES	Mês e ano da apresentação da AIH para a Secretaria	Texto(6)	
ESPEC	Especialidade da AIH	Texto(2)	1: cirurgia geral, 2:obstetrícia, 3:clínica médica, 4:crônico e FPT,5:psiquiatria, 6: tisiologia, 7:pediatria, 8:reabilitação, 9: psiquiatria – hospital/dia
CGC	CGC do Hospital	Texto(14)	
IDENT	Identificação da IAH	Texto(1)	1:normal, 3:de continuação, 5: longa permanência
DT_NASC	Data de nascimento do paciente	Data	
IDADE	Idade do paciente	Texto(4)	Formato (30idade)
SEXO	Sexo do Paciente	Texto(1)	0 – ignorado, 1 – masculino e 3 – feminino.
N_AIH	Número da AIH	Texto(10)	
MED_SOL	CPF do médico solicitante	Texto(11)	
PROC_SOL	Procedimento Solicitado	Texto(8)	
CAR_INT	Caracter da internação	Texto(2)	1 – eletiva, 3 – Urgência (AIH emitida antes da internação), 5 – Urgência (AIH emitida depois da internação),
DT_EMIS	Data da emissão	Data	
MED_RESP	CPF do médico responsável	Texto(11)	
TOT_UTI	Total de dias de UTI	Numero(20,5)	
PROC_REA	Procedimento Realizado	Texto(8)	
DIAG_PRI	Diagnóstico principal	Texto(4)	Código CID
DIAG_SEC	Diagnóstico secundário	Texto(4)	Código CID
MOT_COB	Motivo da cobrança	Texto(2)	Alta: (11 a 19); Permanência mais de 30 dias (21 a 25), Transferência (31 a 39), Óbito com autopsia (41 a 43), Óbito sem autopsia (51 a 53), Alta reoperação (61 a 68).
NACIONAL	Nacionalidade	Texto(2)	
MUN_PAC	Município do paciente	Texto(6)	
FILHO	Numero de filhos	Numero(20,5)	
INSTRU	Grau de instrução	Texto(1)	
CIDNOTIF	Código CID	Texto(4)	

NAT	Natureza da relação Hospital e o SUS	Texto(2)	10: próprio, 20: contratado, 30: federal, 31: federal com verba própria, 40: estadual, 50: municipal, 60: filantrópico, 61: filantrópico isento (IN 01/97 SRF), 70: Universitário, 80: sindicato, 90: Universitário com pesquisa, 91: universitário de pesquisa sem fins lucrativos.
VALSH	Valor serviços hospitalares	Número(20,5)	
VALSP	Valor serviços profissionais	Número(20,5)	
VALSADT	Valor serviços auxiliares de diagnose / terapia	Número(20,5)	
VALOPM	Valor permanência a maior	Número(20,5)	
VALSANG	Valor sangue	Número(20,5)	
VALRN	Valor recén-nato	Número(20,5)	
VALUTI	Valor da UTI	Número(20,5)	
VALACO	Valor diária de acompanhante	Número(20,5)	
VALUTINN	Valor UTI neo-natal	Número(20,5)	
VALTRANSP	Valor de transplante	Número(20,5)	
VALNEURO	Valor neurologia	Número(20,5)	
TOTAL	Soma de todos os campos que armazenam os valores gastos	Número(20,5)	

5.1.4 Modelagem

A modelagem aplicada em todos os experimentos segue o roteiro descrito abaixo.

Roteiro da Mineração

- Identificação da mineração
  1) Descrição da base utilizada
     a. Nome e tipo do arquivo
     b. Critério de seleção dos dados
     c. Total de registros

  2) Análise Univariada
     a. Campo selecionado
     b. Resultados obtidos
     c. Avaliação dos resultados

  3) Análise Multivariada
     a. Técnica estatística
        i. Campos selecionados
        ii. Nome e tipo dos arquivos auxiliares (gerados pelo protótipo)
        iii. Resultados obtidos
        iv. Avaliação dos resultados

5.1.4.1 Experimento 1 – Fatores que influenciam o estado de saúde e o contato com os serviços de saúde.

1) Descrição da base utilizada

Para a seleção do arquivo foram utilizados os códigos das doenças (CID) entre Z00 e Z99, os quais refere-se às categorias que determinam os fatores que influenciam o estado de saúde e o contato com os serviços de saúde. Inicialmente foi utilizado o

protótipo para a seleção e avaliação dos dados. O arquivo resultante da seleção aplicada possui 1176 registros.

2) Análise Univariada

A análise univariada foi realizada para todos os campos quantitativos deste experimento. Mas, somente alguns serão apresentados nesta seção, os demais foram desconsiderados, pois continham grande parte de seus registros zerados.

Os campos VALNEURO e VALOPM apresentam todos os registros zerados.

Para o campo TOTAL (valor total), os valores são:

a) Média (μ): R$ 1153,74
b) Desvio-padrão (σ): R$ 392,59
c) Média + 3 desvios-padrão (μ + 3σ): R$ 3852,03

- Resultado

O arquivo de saída contém 10 registros, os quais são os desvios da base avaliada. Os registros, que compõem a resposta, estão dispostos da seguinte forma:

- Nove registros são do procedimento "31805019" (transplante renal receptor - doador cadáver) e possuem valores entre R$ 11.051,00 e R$ 14.856,00.

- Um registro é do procedimento "38017016" (curativos cirúrgicos sob anestesia geral) e apresenta o valor de R$ 4.366,00.

Avaliando a base de dados, são encontradas somente nove (9) ocorrências do procedimento "31805019", todas fazem parte dos desvios apresentados acima. Este fato determina um padrão de comportamento, onde se conclui que: os procedimentos que envolvem transplante de rim possuem um custo alto para o Estado.

Analisando o outro dado, conclui-se que é um desvio, tanto em termos quantitativos quanto em relação à freqüência de sua ocorrência na base de dados. Numa análise mais detalhada da base de dados, encontram-se 224 ocorrências do procedimento "38017016", mas somente uma é considerada desvio, pela análise univariada. As 223 ocorrências, consideradas normais, possuem um custo médio de R$ 521,00.

Concluído, a avaliação, identifica-se que todos descrevem procedimentos efetuados em pacientes do sexo masculino. Observa-se que dois pacientes reapresentaram a mesma AIH, em dois meses seguidos ambas apresentam valores distintos. Numa primeira análise este fato poderia caracterizar um desvio, mas conversando com profissionais da saúde, que trabalham com o registro de AIH's conclui-se que este fato é normal. Estes pacientes podem ter sofrido uma rejeição do órgão transplantado e o Estado é obrigado a atendê-los novamente, se este problema ocorrer.

Para o campo: VALSH (valor serviços hospitalares), os valores calculados foram:

a) Média (μ): R$ 217,22
b) Desvio-padrão (σ): R$ 561,06
c) Média + 3 desvios-padrão (μ + 3σ): R$ 1900,40

- Resultado

O arquivo de saída armazenou 9 registros referentes aos transplantes de rim e apresentam valores entre R$ 6279,00 e R$ 6798,24 reais, para o campo VALSH. Estes registros são os mesmos selecionados para a avaliação do campo TOTAL.

Para o campo: VALSP (valor serviços profissionais), os valores calculados foram:

a) Média (μ): R$ 126,43
b) Desvio-padrão (σ): R$ 242,24
c) Média + 3 desvios-padrão (μ + 3σ): R$ 853,17

- Resultado

Nove registros foram gravados no arquivo de saída. Estes registros são os mesmos dos avaliados no campo TOTAL, todos são referentes aos transplantes de rim. Apresentam valores entre R$ 2754,00 e R$ 2897,50 reais.

Para o campo: VALSADT (valor serviços auxiliares de diagnose/terapia), os valores calculados foram:

a) Média (μ): R$ 23,20
b) Desvio-padrão (σ): R$ 175,79
c) Média + 3 desvios-padrão (μ + 3σ): R$ 550,57

- Resultado

São considerados desvios os mesmos nove registros referentes aos transplantes de rim. Apresentam valores entre R$ 1980,00 e R$ 2073,10 reais.

Os campos VALSADT, VALSP, VALSH apresentam os maiores valores para os mesmos registros, os quais indicam procedimentos de transplantes de rim. Conclui-se que os atendimentos para os transplantes de rim geram altos custos para o SUS.

Para o campo: VALSANG (valor sangue), os valores calculados foram:

a) Média (μ): R$ 2,50

b) Desvio-padrão (σ): R$ 14,37
c) Média + 3 desvios-padrão (μ + 3σ): R$ 45,64

- Resultado

O arquivo de saída contém 23 registros, sendo que o valor mais alto corresponde a procedimentos de curativos cirúrgicos sob anestesia geral, e possui um valor de R$ 191,30 reais, para o campo avaliado.

Para o campo VALRN (valor recen-nato),os valores calculados foram:

a) Média (μ): R$ 2,73
b) Desvio-padrão (σ): R$ 9,59
c) Média + 3 desvios-padrão (μ + 3σ): R$ 31,50

- Resultado

O arquivo de saída contém 19 registros, considerados desvios, todos apresentam o mesmo valor para VALRN que é R$ 55,00 reais.

Avaliados estes dados identificou-se um erro de qualidade dos dados.

O procedimento descrito em todos os registros, que apresentam desvios, é o mesmo e indica "cesariana com laqueadura tubária em pacientes com cesarianas sucessivas". Ao avaliar os dados constata-se que entre as dezenove mulheres atendidas quinze possuem entre 3 a 5 filhos e quatro mulheres não possuem registro no campo FILHOS.

Estes registros mostram um erro no cadastro dos dados, o registro do procedimento realizado não é coerente com quatro pacientes atendidas, ou o registro do número de filhos das pacientes não foi realizado de forma correta.

3) Análise Multivariada

A seleção dos campos para a análise multivariada pretende seguir uma coerência de avaliação. Em alguns casos estudados as associações entre determinados campos do registro, não apresentam coerência de avaliação. Por este motivo foram testados campos que possuem informações relevantes entre si.

Qui-quadrado

A primeira análise multivariada realizada no protótipo é a técnica do qui-quadrado, baseada na definição de hipóteses. As hipóteses formuladas pretendem determinar se existe alguma dependência entre a época do ano em que são apresentadas as AIH's e os municípios que realizam estes atendimentos. Para esta verificação foram selecionados os campos:

- Data de apresentação
- Município do paciente;

As hipóteses definidas, a partir dos campos selecionados, são:

- H_0: A data de apresentação da AIH é independente do município do paciente.
- H_1: A data de apresentação da AIH é dependente do município do paciente.

- Resultado

Os resultados obtidos apresentam valores de:

- *Pearson* qui-quadrado: 1280,97
- Graus de liberdade: 1036

O valor do *Pearson* é considerado válido para a avaliação, mas o mesmo não ocorre com o valor indicado para os graus de liberdade. Quando o número de perfis (graus de liberdade) for muito grande o cálculo do qui-quadrado não revelará o grau de dependência das variáveis [PER 2001]. As tabelas usuais para a verificação dos valores do qui-quadrado apresentam, como valor máximo dos graus de liberdade, o valor de 100 *g*l (graus de liberdade).

Dando continuidade a análise multivariada, não será utilizada a verificação pela análise de correspondência, pois esta é baseada no cálculo do qui-quadrado. Se os graus de liberdade não são satisfatórios, para o qui-quadrado, também não o serão para a análise de correspondência.

Desta forma, a análise que pode ser efetuada é a verificação dos resíduos ajustados, resultantes da tabela de contingência. Os dados que serão utilizados para a avaliação dos resíduos são preparados pelo protótipo e depois aplicados no software SPSS.

Tabela de Contingência - Análise dos Resíduos

O protótipo prepara um arquivo que contém somente os campos selecionados e suas respectivas codificações, i.e. para cada valor encontrado no arquivo o protótipo codifica seqüencialmente estes valores em outro campo adicional. Ao mesmo tempo gera arquivos auxiliares para a compreensão dos códigos gerados.

Os arquivos auxiliares estão apresentados a seguir, sendo a tabela 5.3 referente à tabulação das datas de apresentação e a tabela 5.4 referente aos municípios. É importante ressaltar que é apresentado somente um trecho da tabela 5.4, o qual permite a compreensão dos dados apresentados.

- Resultado

A tabela de contingência resultante da aplicação no SPSS apresentou a matriz relacional que contêm aproximadamente 95,1% das células com valores esperados inferiores a 5 ocorrências, reafirmando a inviabilidade da análise pelo teste do qui-quadrado.

TABELA 5-3 - Arquivo Auxiliar
Data de apresentação

COD	APRES
1	12001
2	62000
3	72000
4	82000
5	92000
6	102000
7	112000
8	122000

TABELA 5-4 - Arquivo Auxiliar
Código do Município

COD	MUNICIPIO
1	421310
2	430005
3	430010
4	430020
5	430040
6	430060
7	430100
8	430160
9	430200
10	430205
11	430210
12	430310

A figura 5.1 apresenta um trecho da tabela gerada no SPSS, onde é possível verificar as discrepâncias entre os valores esperados e os valores observados.A tabela apresenta um desvio na célula que caracteriza a relação entre a data de apresentação de dezembro de 2000 e a cidade de Alvorada “430060”. Este desvio informa que houve muito mais ocorrências no mês de dezembro do que eram esperadas para esta cidade neste mês.

Avaliando os dados referentes há este mês e a esta cidade, descobriu-se que todos os registros são referentes a laqueadura tubária e o diagnóstico principal de todas as pacientes é a esterilização. Ampliando a avaliação verifica-se que dos 105 registros cadastrados neste arquivo 104 registros são referentes à laqueadura tubária e um registro e cesariana com laqueadura tubária.

			MUN_PAC					
			1	2	3	4	5	6
DTA_APRE	1	Count	0	1	0	1	0	21
		Adjusted Residual	-,4	2,4	-,4	1,4	-,6	1,5
	2	Count	0	0	0	0	0	9
		Adjusted Residual	-,3	-,3	-,3	-,5	-,5	-,6
	3	Count	0	0	0	0	1	5
		Adjusted Residual	-,3	-,3	-,3	-,5	1,8	-2,1
	4	Count	0	0	0	1	0	14
		Adjusted Residual	-,4	-,4	-,4	,7	-,5	,5
	5	Count	1	0	0	0	0	7
		Adjusted Residual	2,9	-,4	-,4	-,5	-,5	-1,5
	6	Count	0	0	0	0	0	9
		Adjusted Residual	-,4	-,4	-,4	-,5	-,5	1,2
	7	Count	0	0	0	0	0	25
		Adjusted Residual	-,4	-,4	-,4	-,5	-,5	3,6
	8	Count	0	0	1	0	1	15
		Adjusted Residual	-,4	-,4	2,3	-,6	1,3	-,6
Total		Count	1	1	1	2	2	105

Desvio

FIGURA 5.1 - Trecho da Tabela de Contingência (SPSS)

5.1.4.2 Experimento 2 – Gravidez, parto e puerpério

1) Descrição da base utilizada

Com a utilização do protótipo é selecionada a base de dados que se refere aos dados sobre Gravidez, parto e puerpério. O arquivo gerado contém 57.399 registros. Para esta seleção foram utilizados os códigos de CID entre O00 e O99.

2) Análise Univariada

Os campos de VALOPM, VALACO e VALNEURO apresentam valores zerados para todos os registros.

Numa primeira análise foram encontrados quatro registros com todos os valores zerados, os registros caracterizam partos normais, em caráter de internação de urgência

e todos efetuados em hospitais contratados e as pacientes são residentes da cidade de Muçum. Estes 4 registros são todos os apresentados neste mês.

Foram efetuadas avaliações para todos os campos quantitativos, mas estão apresentados nesta seção somente os campos com informações mais significativas.

Para o campo: TOTAL, os valores calculados foram:

a) Média (μ): R$ 258,96
b) Desvio-padrão (σ): R$ 131,77
c) Média + 3 desvios-padrão (μ + 3σ): R$ 654,29

- Resultado

O arquivo de saída contém 219 registros considerados desvios. Apenas nove destes registros não são referentes a partos e cesarianas.

Quatro registros se caracterizam pela "retirada do útero" por motivo de ruptura do útero durante o parto. A idade das pacientes indica que 50% dos casos registrados (2) ocorreram em adolescentes de 15 anos, ambas residentes na cidade de Novo Hamburgo. Os outros dois casos ocorreram com mulheres de 27 anos, residente em Canoas, e 37 anos, residente em Lajeado.

Avaliando-se a idade das pacientes identifica-se que no arquivo de saída existem 4 registros de adolescentes de 15 anos, duas submetidas à cesariana e duas à retirada de útero.

O arquivo contém 9 registros de adolescentes menores de idade, sendo que 5 são da cidade de Novo Hamburgo.

As internações mais caras apresentam valores altos para o custo de UTI para o recém-nascido ou em serviços prestados ao recém-nato.

Desconsiderando 2 registros referentes a problemas de septicemia os demais apresentam uma incidência de nascimento concentrada entre junho de 2000 a janeiro de 2001. O mês de outubro apresenta uma maior concentração de nascimentos e é possível observar que destes casos 98% das mães tinham idade entre 24 a 45 anos.

As informações mais interessantes registram grande quantidade de curetagem após aborto, sendo maior do que o número de cesarianas registradas no mesmo período. Entre estes dados aparecem três meninas de 14 anos.

3) Análise Multivariada

A seleção dos campos para a análise multivariada segue a mesma coerência aplicada ao experimento anterior.

Qui-quadrado

Os campos selecionados para esta avaliação são:

- Especialidade da AIH

- Natureza da relação entre o Hospital e o SUS

As hipóteses definidas são:

- H_0: A especialidade da AIH independe da natureza da relação entre o hospital e o SUS.
- H_1: A especialidade da AIH depende da natureza da relação entre o hospital e o SUS.

- Resultado

Os resultados obtidos apresentam valores de:

- *Pearson* qui-quadrado: 601,509

- Graus de liberdade: 14

O protótipo apresentou valor de 601,50 para o qui-quadrado com 14 graus de liberdade. Considerando o que foi estudado, pode-se observar que o valor do qui-quadrado é muito alto, pois o valor tabelado para o qui-quadrado com 14 graus de liberdade e nível de significância 0,05 é 23,7. O valor calculado está muito além do valor estabelecido na tabela.

Neste caso as variáveis avaliadas são altamente dependentes entre si, i.e. as variáveis são associadas, o que representa dizer que a maioria dos atendimentos cadastrados nesta base de dados é altamente dependente do tipo de hospital que realiza os atendimentos.

Considerando as informações registradas na base de dados, afirma-se:

- Na especialidade de obstetrícia a grande maioria dos atendimentos são efetuados em hospitais filantrópicos, o que determina um padrão de comportamento.

- Os hospitais do SUS apresentam poucos atendimentos na área de obstetrícia e na área de clínica médica, para as doenças cadastradas nas áreas relacionadas entre gravidez, parto e puerpério.

Os dados avaliados apresentam uma grande significância em relação aos cruzamentos executados. O próximo passo é a análise de correspondência, a qual permite a verificação gráfica dos resultados encontrados com o teste qui-quadrado.

Análise de Correspondência

Para esta avaliação os dados foram reformulados no protótipo, o qual seleciona os valores encontrados e grava um arquivo resposta, completamente codificado, que será utilizado na análise de correspondência.

Aplicando os dados no software do SPSS, o resultado é a tabela de correspondência, que mostra as freqüências entre a natureza de operação e as especialidades avaliadas, conforme a figura 5.2.

Correspondence Table

	NATUREZA								
ESPEC	10	20	50	60	61	63	70	90	Active Margin
1	5	367	380	1885	250	1	24	563	3475
2	0	7444	4402	31650	3743	34	132	3838	51243
3	0	490	175	1491	190	0	5	330	2681
Active Margin	5	8301	4957	35026	4183	35	161	4731	57399

FIGURA 5.2 - Tabela de Correspondência - SPSS

Apresenta também os valores correspondentes calculados para as cargas (*scores*) das colunas (figura 5.3) e linhas (figura 5.4) do gráfico analisado.

Overview Column Points[a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
NATUREZA	Mass	1	2	Inertia	1	2	1	2	Total
10	,000	-12,304	5,621	,001	,140	,068	,918	,082	1,000
20	,145	,163	-,358	,001	,041	,461	,325	,675	1,000
50	,086	-,164	,337	,001	,025	,243	,357	,643	1,000
60	,610	,108	,061	,001	,075	,056	,879	,121	1,000
61	,073	,016	,025	,000	,000	,001	,482	,518	1,000
63	,001	,633	,903	,000	,003	,012	,534	,466	1,000
70	,003	-1,110	,784	,000	,037	,043	,824	,176	1,000
90	,082	-,881	-,238	,006	,679	,116	,970	,030	1,000
Active Total	1,000			,010	1,000	1,000			

a. Symmetrical normalization

FIGURA 5.3 - *Score* das colunas - SPSS

Overview Row Points[a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
ESPEC	Mass	1	2	Inertia	1	2	1	2	Total
1	,061	-1,158	,227	,008	,862	,077	,984	,016	1,000
2	,893	,095	,031	,001	,086	,021	,957	,043	1,000
3	,047	-,322	-,882	,002	,051	,902	,237	,763	1,000
Active Total	1,000			,010	1,000	1,000			

a. Symmetrical normalization

FIGURA 5.4 - *Score* das linhas - SPSS

Depois de descrever os valores calculados é apresentado o gráfico resultante desta avaliação. O gráfico é apresentado na figura 5.5 e apresenta um agrupamento maior indicando uma grande concentração de dados com as mesmas características.

Avaliando o gráfico pode-se identificar um ponto isolado no lado superior–esquerdo. Este ponto é um desvio, pois define um ponto extremo fora do agrupamento maior.

O ponto extremo apresentado no gráfico pode ser identificado, avaliando-se as definições das linhas e das colunas apresentadas anteriormente. Observando o gráfico identifica-se:

- O ponto extremo, indicado pela seta, é um dado relativo ao campo NAT.
- Localiza-se aproximadamente entre os valores: *dimension* 1= -12 e *dimension* 2 = 5,6.

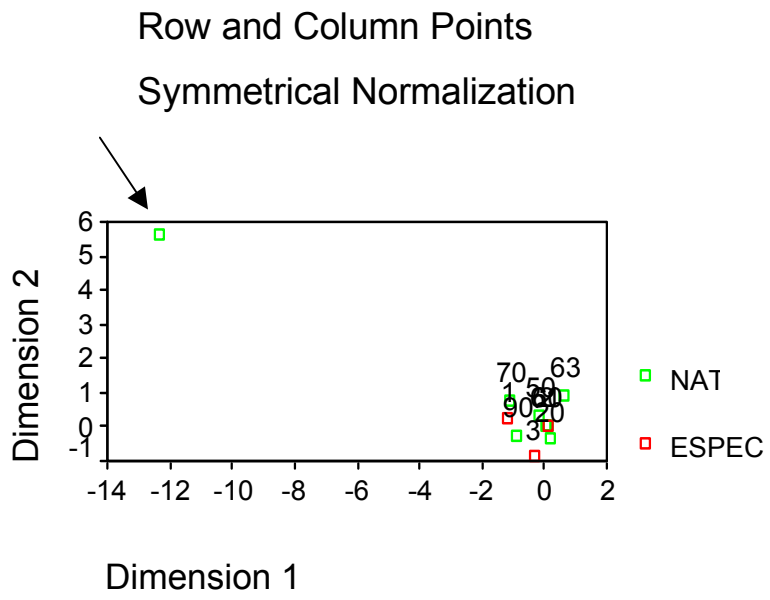


FIGURA 5-5 - Gráfico da análise de correspondência - SPSS

Observando-se a tabela formada pelos valores de NAT, identifica-se que o ponto é referente aos valores de Nat = 10, i.e. os atendimentos dos hospitais próprios do SUS.

5.1.4.3 Experimento 3 – Causas externas de morbidade e mortalidade

1) Descrição da base utilizada

As configurações escolhidas definem o trabalho com a base das AIH's e a análise numérica é efetuada sobre o campo "valor-total". O arquivo selecionado contém

os registros com o CID entre V01 e Y98, o que determina registros sobre causas externas de morbidade e mortalidade, totalizando 713 registros.

2) Análise Univariada

Para o campo: TOTAL, os valores calculados foram:

a) Média (μ): R$ 1.234,80
b) Desvio-padrão (σ): R$ 873,38
c) Média + 3 desvios-padrão (μ + 3σ): R$ 3.854,94

- Resultado

O arquivo resposta gerado contém 2 registros, que apresentam as seguintes características:

- Mesma AIH, (2294068337), logo é o mesmo paciente: um homem de 95 anos de idade da cidade de Colorado.
- Dois registros do procedimento 39003124 (artroplastia coxo-femural) e diagnostico principal, fratura do fêmur, o valor dos atendimentos está entre R$ 5.240,00 e R$ 5.404,00
- Motivo de cobrança 22: intercorrência que significa retorno ao hospital por problemas de rejeição.

Para o campo: TOTUTI (Total de dias de UTI), os valores calculados foram:

a) Média (μ): 0,12 dias
b) Desvio-padrão (σ): 1 dia
c) Média + 3 desvios-padrão (μ + 3σ): 3,13 dias

- Resultado

O arquivo resposta gerado contém 14 registros. Observa-se que destes registros somente um identifica um paciente de 42 anos de idade, os demais registros são referentes aos pacientes com idade superior a 65 anos de idade.

Considerando-se também o diagnóstico principal destas internações, verifica-se que 13 registros apresentam o diagnóstico de Artrose e procedimento de artroplastia coxo-femural. Apenas um registro tem o diagnóstico de complicações não especificadas em relação a cuidados médicos e cirúrgicos e o procedimento referente a complicações de procedimentos cirúrgicos ou médicos. Este registro é referente ao paciente de 42 anos de idade.

Avaliando-se os dados referentes ao sexo do paciente, identifica-se que 10 registros são de pacientes do sexo feminino e apenas 4 do sexo masculino.

Para o campo: VALUTI (Valor total da UTI), os valores calculados foram:

a) Média (μ): R$ 10,70
b) Desvio-padrão (σ): R$ 79,67
c) Média + 3 desvios-padrão (μ + 3σ): R$ 249,72

- Resultado

O arquivo resposta gerado contém 9 registros. Os maiores valores referentes aos gastos com UTI estão apresentados na avaliação anterior. É preciso ressaltar que nestes registros 8 são pacientes do sexo feminino e 7 delas com idades superiores a 80 anos. A outra mulher deste grupo é a paciente de 42 anos apresentada acima. Um registro interessante e do único paciente do sexo masculino, com idade de 89 anos que permaneceu 3 dias na UTI.

Para o campo: VALUTINN (Valor total de UTI neo-natal), os valores calculados foram:

a) Média (μ): R$ 15,58
b) Desvio-padrão (σ): R$ 179,94
c) Média + 3 desvios-padrão (μ + 3σ): R$ 555,40

- Resultado

De acordo com a descrição do campo, as características dos registros do arquivo de saída já são um grande desvio de qualidade de dados. O arquivo de saída possui 6 registros, sendo que nenhum paciente, cadastrado nestes registros tem idade inferior a 65 anos, sendo que três pacientes são de 95 anos e um paciente tem 97 anos. Todos cadastrados com o mesmo procedimento realizado, artroplastia coxo-femural. Dois pacientes possuem como diagnóstico principal, fratura da diáfise do fêmur e 4 com o diagnóstico de artrose NE (não especificado).

3) Análise Multivariada

Qui-quadrado

A próxima avaliação é o cálculo do qui-quadrado, os campos selecionados para esta avaliação são:

- Procedimento realizado
- Data de apresentação

As hipóteses definidas são:

- H_0: Os procedimentos realizados são independentes da data de apresentação das AIH's.
- H_1: Os procedimentos realizados são dependentes da data de apresentação das AIH's.

- Resultado

Os valores calculados são:

- *Pearson* qui-quadrado: 27,121

- Graus de liberdade: 21

Este experimento informou que existem 21 graus de liberdade e o valor do qui-quadrado é de 27,121. Avaliando o resultado obtido e a tabela de valores é possível observar que o valor apresentado é inferior ao valor apresentado na tabela, o valor da tabela é 32,7. Este resultado apresenta que as variáveis avaliadas são independentes entre si i.e., não são associadas. Logo a hipótese H_0 é validada.

Como os valores encontrados para os graus de liberdade permitem a avaliação da análise de correspondência, esta é apresentada abaixo.

Análise de Correspondência

Os campos selecionados são os mesmos, avaliados no qui-quadrado.

- Resultado

Os dados forma aplicados no SPSS, e as tabelas resultantes são as seguintes:
A tabela de correspondência gerada para os campos selecionados, esta representada na figura 5.6.

Correspondence Table

	PROCED_R				
DTA_APRE	36013099	39003124	85500798	91500141	Active Margin
012001	1	67	3	23	94
062000	0	56	3	46	105
072000	0	43	2	30	75
082000	1	75	0	38	114
092000	0	54	2	29	85
102000	0	51	0	34	85
112000	1	42	0	17	60
122000	0	64	0	31	95
Active Margin	3	452	10	248	713

FIGURA 5.6- Tabela de Correspondência - SPSS

O SPSS calcula os valores das cargas (*scores*) para as linhas e colunas da análise de correspondência. A figura 5.7 apresenta as cargas das linhas e a figura 5.8 apresenta as cargas das colunas.A figura 5.9 apresenta o gráfico da análise de correspondência, gerado para as cargas apresentadas a seguir.

Overview Row Points[a]

DTA_APRE	Mass	Score in Dimension 1	Score in Dimension 2	Inertia	Contribution: Of Point to Inertia of Dimension 1	Contribution: Of Point to Inertia of Dimension 2	Contribution: Of Dimension to Inertia of Point 1	Contribution: Of Dimension to Inertia of Point 2	Contribution: Of Dimension to Inertia of Point Total
012001	,132	-,392	,660	,010	,136	,500	,312	,684	,996
062000	,147	,605	,118	,009	,363	,018	,928	,027	,955
072000	,105	,423	,151	,003	,126	,021	,908	,089	,998
082000	,160	-,280	-,255	,003	,084	,091	,567	,364	,931
092000	,119	,139	,197	,001	,015	,040	,265	,414	,680
102000	,119	,168	-,454	,003	,023	,214	,151	,847	,998
112000	,084	-,640	-,096	,006	,232	,007	,871	,015	,886
122000	,133	-,150	-,307	,003	,020	,109	,152	,489	,640
Active Total	1,000			,038	1,000	1,000			

a. Symmetrical normalization

FIGURA 5.7 - Cargas das linhas

Overview Column Points[a]

PROCED_R	Mass	Score in Dimension 1	Score in Dimension 2	Inertia	Contribution: Of Point to Inertia of Dimension 1	Contribution: Of Point to Inertia of Dimension 2	Contribution: Of Dimension to Inertia of Point 1	Contribution: Of Dimension to Inertia of Point 2	Contribution: Of Dimension to Inertia of Point Total
36013099	,004	-2,941	,898	,008	,245	,030	,692	,050	,742
39003124	,634	-,232	,040	,006	,229	,009	,914	,021	,935
85500798	,014	1,187	2,641	,014	,133	,853	,207	,793	1,000
91500141	,348	,410	-,190	,011	,393	,109	,824	,136	,960
Active Total	1,000			,038	1,000	1,000			

a. Symmetrical normalization

FIGURA 5.8 - Cargas das colunas

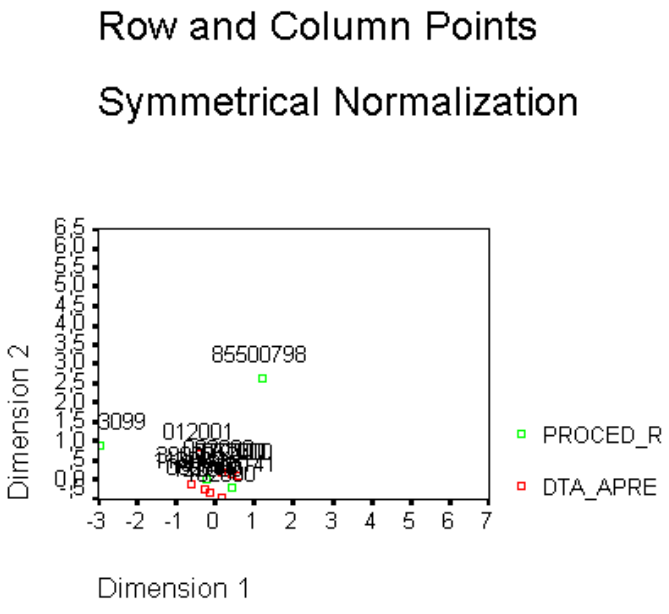


FIGURA 5.9 - Gráfico da análise de correspondência

Avaliando o gráfico, pode-se verificar que não existem valores que informem um desvio dos dados avaliados. Observa-se no gráfico dos valores distantes do agrupamento central, que se somente o gráfico fosse analisado poderia dar a impressão

de que estes dados seriam desvios em relação ao agrupamento principal. Mas com base nas outras avaliações, conclui-se que estes valores não são desvios da avaliação principal.

Para confirmar que não são encontrados valores que possam ser considerados desvios, é feita a avaliação dos resíduos.

Avaliação dos resíduos

Os dados avaliados são os mesmos que são avaliados na análise de correspondência.

- Resultado

A tabela a seguir é resultante da aplicação do software do SPSS, como mostra a figura 5.10.

APRES * PROC_REA Crosstabulation

			PROC_REA				
			36013099	39003124	85500798	91500141	Total
APRES	012001	Count	1	67	3	23	94
		Adjusted Residual	1,0	1,7	1,6	-2,3	
	062000	Count	0	56	3	46	105
		Adjusted Residual	-,7	-2,3	1,4	2,1	
	072000	Count	0	43	2	30	75
		Adjusted Residual	-,6	-1,2	1,0	1,0	
	082000	Count	1	75	0	38	114
		Adjusted Residual	,8	,6	-1,4	-,4	
	092000	Count	0	54	2	29	85
		Adjusted Residual	-,6	,0	,8	-,1	
	102000	Count	0	51	0	34	85
		Adjusted Residual	-,6	-,7	-1,2	1,1	
	112000	Count	1	42	0	17	60
		Adjusted Residual	1,6	1,1	-1,0	-1,1	
	122000	Count	0	64	0	31	95
		Adjusted Residual	-,7	,9	-1,2	-,5	
Total		Count	3	452	10	248	713

FIGURA 5-10 - Tabela de análise de resíduos

Verificando os resíduos calculados, observa-se que não existem desvios nos dados avaliados.

5.2 Estudo de Caso – Registros de óbitos

A base de dados utilizada, nestes experimentos, foi fornecida pela SES e refere-se aos registros de óbitos registrados em cartório, no período de 1999 a 2000.

5.2.1 Entendimento do negócio

Estes dados apresentam um conjunto de informações sobre o óbito ocorrido. A grande quantidade de informações procura padronizar os registros e caracterizar o falecido em relação à escolaridade, ocupação, sexo, cidade, causa da morte e outras características que serão apresentadas na seção 5.2.3.

- **Objetivo**

O objetivo destes experimentos é o mesmo apresentado no estudo de caso do item 5.1. Encontrar um padrão de comportamento para os registros de óbitos avaliados e a partir deste padrão detectar informações que se diferenciem do mesmo.

Para realizar o objetivo, também são aplicadas técnicas estatísticas, mas no caso desta base de dados, somente as técnicas multivariadas serão aplicadas. Pois não existem registros quantitativos nestas bases.

- **Recursos**

O trabalho desenvolvido utilizou os mesmos recursos descritos no estudo de caso da seção 5.1.

5.2.2 Entendimento dos dados

Esta base de dados foi utilizada em pesquisas anteriores desenvolvidas em sala de aula. E as informações aqui apresentadas são compatíveis com as informações estudadas anteriormente sobre as AIH's, pois muitas tabelas e arquivos utilizados facilitaram a compreensão destes dados.

Estes dados foram selecionados para este trabalho por já serem de conhecimento do autor desta dissertação. As informações aqui contidas foram estudadas a partir de um formulário de registro de óbito, utilizado pelo SES.

A base inicial é constituída de 138.909 registros, os quais possuem originalmente 36 campos que compõem o registro do arquivo. Os registros da tabela são definidos a partir do formulário de óbito, que está no anexo deste trabalho. Na tabela 5.5 estão apresentados os atributos utilizados nesta pesquisa.

TABELA 5-5- Descrição dos campos da tabela mortalidade

Nome do atributo	Significado	Tipo do campo	Faixa de valores
NUMERODO	Número do óbito	Texto (8)	Qualquer seqüência de caracteres numéricos de tamanho 8
DTOBITO	Data do óbito	Texto (8)	Data no formato ddmmyyyy (barras não são representadas)
NATURAL	Naturalidade do morto	Texto (10)	Código do município natal

DTNASC	Data de nascimento	Texto (8)	Data no formato ddmmyyyy (barras não são representadas)
IDADE	Idade	Texto (4)	*Este campo foi substituído pelo campo faixa etária, descrito na próxima sessão.
SEXO	Sexo	Texto (1)	F,M,I
RACACOR	Raça/cor	Texto (10)	Branca, preta, parda, indígena ,ignorada
ESTCIV	Estado Civil	Texto (15)	Solteiro, casado, viúvo, separado, união consensual, ignorado.
ESC	Escolaridade	Texto (15)	Nenhuma, de 1 a 3 anos, de 4 a 7 anos, de 8 a 11 anos, 12 ou mais, ignorado.
OCUP	Ocupação habitual e ramo de atividade	Texto (10)	Código da atividade
CEPRES	CEP residencial	Texto (10)	CEP residencial (hífen é omitido)
CODMUNRES	Código do município da residência	Texto (10)	Código do município de residência
LOCOCOR	Local de ocorrência do óbito	Texto (15)	Hospital, outros_estabelecimentos_de_saúde, via_publica, domicílio, ignorado.
ASSISTMED	Flag que indica se o morto recebeu assistência médica	Texto (5)	S,N,I
EXAME	Flag que indica se haviam sido realizados exames no morto	Texto (5)	S,N,I
CIRURGIA	Flag que indica se o morto havia sofrido cirurgia	Texto (5)	S,N,I
NECROPSIA	Flag que indica se houve necropsia	Texto (5)	S,N,I
CAUSABAS	Causa básica da morte	Texto (20)	Código Internacional de doenças
LINHAA	Indica (se houver) a doença que levou à causa básica da morte	Texto (20)	Código Internacional de doenças
LINHAB	Indica (se houver) a doença que levou à doença da linha A	Texto (20)	Código Internacional de doenças
LINHAC	Indica (se houver) a doença que levou à doença da linha B	Texto (20)	Código Internacional de doenças

5.2.3 Preparação dos dados

Analisando a base de dados, observa-se que existem muitos registros, que possuem campos incompletos ou ausentes. A figura 4.2, apresentada no capítulo 4, mostra um gráfico de freqüência que ilustra esta afirmação.

Estes registros incompletos, encontrados na base, merecem um tratamento especial, para que possam ser avaliados através das técnicas estatísticas selecionadas. Estes registros foram preenchidos com valores que não alteram a informação original. A tabela 5.6 ilustra este procedimento.

TABELA 5-6 – Preparação dos dados

Campo	Preenchimento do valor ausente	Significado
IDADE1	888	Valor nulo
ACIDTRAB	9	Valor nulo ou ignorado
CIRCOBITO	4	Outros(considera-se que o não-preenchimento deste campo implica em morte não classificada como homicídio, suicídio ou acidente, nem ignorada)
CRM	888888888888888	Valor Nulo
NECROPSIA	8	Valor Nulo
CIRURGIA	8	Valor Nulo
EXAME	8	Valor Nulo
ASSISTMED	8	Valor Nulo
CEPOCOR	88888888	Valor Nulo
LOCOCOR	8	Valor Nulo
CEPRES	88888888	Valor Nulo
OCUP	88888	Valor Nulo
ESC	9	Valor Nulo ou Ignorado
ESTCIV	9	Valor Nulo ou Ignorado
RACACOR	8	Valor Nulo
NATURAL	888	Valor Nulo
HORAOBITO	8888	Valor Nulo
CODMUNCART	8888888	Valor Nulo
DTREGCART	88888888	Valor Nulo

Como existem inúmeros códigos para os registros de Doenças e se tornava difícil e pouco proveitoso à avaliação dos dados da forma como estavam apresentados na tabela. Os dados foram reestruturados de acordo com os códigos das doenças. Os agrupamentos gerados obedecem a os tipos e as categorias das doenças, facilitando o trabalho de avaliação.

Para direcionar a pesquisa e agilizar a avaliação dos resultados, foram selecionadas algumas cidades do RS, apresentadas na tabela 5.7.

TABELA 5-7 –Relação das cidades

Código da Cidade	Descrição	Quantidade registros
4314902	Porto Alegre	14763
4314407	Pelotas	4055
4305108	Caxias do Sul	2828
4316907	Santa Maria	2361
4315602	Rio Grande	2322
4314100	Passo Fundo	1627
4322400	Uruguaiana	1515
4316808	Santa Cruz	1162
4317509	Santo Ângelo	815
	Total registros:	31448

5.2.4 Modelagem

Esta modelagem foi realizada seguindo o mesmo roteiro apresentado no estudo de caso anterior, mas com uma pequena adaptação. Os dados da mortalidade não possuem campos quantitativos, logo foi realizada apenas a análise multivariada.

5.2.4.1 Experimento 1 – Fatores que influenciam o estado de saúde e o contato com os serviços de saúde.

1) Descrição da base utilizada

Descrita no item 5.2.4.

2) Análise Univariada

Qui-quadrado

A próxima avaliação é o cálculo do qui-quadrado, os campos selecionados para esta avaliação são:

- Causa da morte
- Município de residência

As hipóteses definidas são:

- H_0: A causa da morte é independente do município de residência do falecido.
- H_1: A causa da morte é dependente do município de residência do falecido

- Resultado

Os valores calculados são:

- *Pearson* qui-quadrado: 1505
- Graus de liberdade: 128

O valor encontrado para o Pearson, é válido, mas o mesmo não ocorre para os graus de liberdade apresentados. Este valor é muito alto, dificultando qualquer avaliação através desta técnica.

Como o qui-quadrado não apresentou valores válidos para a sua avaliação, não será executada a análise de correspondência. As análises seguintes vão se concentrar na avaliação dos resíduos ajustados.

Tabela de analise de Resíduos.

A tabela gerada por esta análise é apresentada na figura 5.11. A avaliação do município em relação à causa da morte desencadeou um grande número de relações que podem ser interpretadas como desvios.

Para facilitar a interpretação da tabela e apresentar os resultados avaliados, serão descritas as relações que causam grandes distorções entre os valores observados e esperados.

Analisando a tabela encontra-se:

- Em relação às doenças "AB" – Doenças infecciosas e parasitárias:

    - Caxias do Sul, Pelotas, Santa Cruz e Santa Maria apresentam resíduos abaixo de -3σ. O que significa que estas cidades tiveram um número menor de mortes em relação ao valor observado.
    - Porto Alegre, ao contrário das demais, apresentou um número muito maior de mortes do que o esperado.

Interpretação:

Baseando-se nos resíduos calculados, conclui-se que mesmo sendo a capital do estado do RS, a incidência de mortes por doenças parasitárias é maior do que outras cidades do estado. Este tipo de verificação pode determinar um problema de saneamento em relação à população.

- Em relação às doenças "E" – Doenças endócrinas, nutricionais e metabólicas:

    - Uruguaiana e Pelotas apresentaram desvios em relação a estas doenças. Foram registrados mais casos do que era esperado nestas cidades.
    - Porto Alegre apresentou uma incidência menor do que a esperada.

Interpretação:

Observando os dados, contata-se que cidades menores possuem maior número de casos destas doenças do que um grande centro como Porto Alegre. Este estudo pode determinar que o controle destas doenças pode ser mais eficaz em maiores centros urbanos. As pessoas que residem em grandes centros se preocupam mais com a saúde nutricional, este fato esta diretamente ligado aos hábitos alimentares modernos.

- Em relação às doenças "F" – Transtornos mentais e comportamentais:

    - Santa Cruz do Sul apresentou um desvio elevado em relação a estas doenças. Foram registrados mais casos do que o esperado para a cidade.

Interpretação:

Avaliando a quantidade de casos registrados em Porto Alegre e Santa Cruz. Observa-se que Santa Cruz registra quase 20% dos números de casos do que a capital. Considerando-se a população das duas cidades, Santa Cruz apresenta um grau elevado de doenças mentais, tal fato merece um estudo mais detalhado sobre as causas deste problema.

NUMCAU * CODMUNRES Crosstabulation

			CODMUNRES									
			4305108	4314100	4314407	4314902	4315602	4316808	4316907	4317509	4322400	Total
NUMCAU	AB	Count	96	81	198	1295	123	48	99	37	112	2089
		Adjusted Residu	-7,3	-2,8	-4,8	14,3	-2,7	-3,5	-5,0	-2,4	1,2	
	CD	Count	22	6	28	129	15	8	19	6	14	247
		Adjusted Residu	,0	-2,0	-,7	1,7	-,8	-,4	,1	-,2	,6	
	E	Count	160	90	286	821	154	74	136	42	126	1889
		Adjusted Residu	-,8	-,8	3,0	-3,1	1,3	,5	-,5	-1,0	3,9	
	F	Count	13	8	18	122	14	24	20	8	12	239
		Adjusted Residu	-1,9	-1,3	-2,5	1,3	-,9	5,2	,5	,7	,1	
	G	Count	58	38	56	336	33	18	52	13	26	630
		Adjusted Residu	,2	1,0	-3,0	3,2	-2,1	-1,1	,7	-,8	-,8	
	H	Count	1	0	2	4	0	0	2	0	0	9
		Adjusted Residu	,2	-,7	,8	-,2	-,8	-,6	1,7	-,5	-,7	
	I	Count	1198	597	1890	6528	1057	500	989	338	525	13622
		Adjusted Residu	-1,1	-5,5	4,5	3,0	2,2	-,2	-1,5	-1,1	-7,0	
	J	Count	397	291	606	2354	334	211	328	165	203	4889
		Adjusted Residu	-2,3	2,7	-1,1	1,8	-1,6	2,5	-2,3	3,8	-2,4	
	K	Count	240	130	295	1023	191	60	134	60	111	2244
		Adjusted Residu	2,9	1,4	,4	-1,3	2,1	-2,7	-2,9	,3	,3	
	L	Count	2	1	12	14	9	4	8	1	2	53
		Adjusted Residu	-1,3	-1,1	2,1	-3,0	2,7	1,5	2,1	-,3	-,4	
	M	Count	25	10	21	120	6	3	11	2	13	211
		Adjusted Residu	1,5	-,3	-1,3	2,9	-2,5	-1,8	-1,3	-1,5	,9	
	N	Count	64	30	59	241	54	18	47	11	16	540
		Adjusted Residu	2,3	,4	-1,4	-1,1	2,3	-,4	1,1	-,8	-2,0	
	O	Count	4	1	4	26	0	1	3	1	2	42
		Adjusted Residu	,1	-,8	-,7	1,9	-1,8	-,5	-,1	-,1	,0	
	P	Count	196	113	260	850	173	74	134	42	194	2036
		Adjusted Residu	1,0	,8	-,2	-4,9	2,0	-,1	-1,6	-1,6	10,3	
	Q	Count	72	23	50	227	32	21	34	9	27	495
		Adjusted Residu	4,4	-,5	-1,9	-,5	-,8	,7	-,5	-1,1	,7	
	R	Count	120	146	127	238	57	47	272	54	113	1174
		Adjusted Residu	1,5	11,5	-2,2	-18,7	-3,4	,6	20,8	4,4	7,8	
	V	Count	160	62	143	435	70	51	73	26	19	1039
		Adjusted Residu	7,3	1,2	,8	-3,3	-,8	2,1	-,6	-,2	-4,6	
Total		Count	2828	1627	4055	14763	2322	1162	2361	815	1515	31448

FIGURA 5.11 – tabela do SPSS

- Em relação às doenças “G” – Doenças do sistema nervoso:

  - Pelotas registra um número muito pequeno de casos em relação à quantidade esperada, enquanto Porto Alegre registra mais casos do que o esperado.

Interpretação:

As duas cidades estão próximas ao limite estabelecido, mas mesmo assim apresentam valores que são considerados desvios em relação aos valores esperados para as suas células. De acordo com o estudo desenvolvido, pode ser interessante o estudo das causas destes dados encontrados.

- Em relação às doenças “I” – Doenças do aparelho circulatório:

    - Passo Fundo e Uruguaiana registram menos ocorrências de morte, em relação ao valor esperado para estas cidades.
    - Porto Alegre e Pelotas registram mais mortes do que a quantidade esperada.

Interpretação:

Avaliando os dados da cidade de Porto Alegre, nota-se que quase 50% das mortes registradas são decorrentes desta doença. Este mesmo raciocínio pode ser feito em relação à quantidade de mortes por problemas circulatórios em relação às cidades avaliadas, onde Porto Alegre registra quase 50% dos casos de morte em relação à doença.

Pelotas apresenta uma característica semelhante a Porto Alegre, quase 50% das mortes registradas em Pelotas é decorrente desta doença.

- Em relação às doenças “P” – Algumas afecções originadas no período perinatal:

    - Uruguaiana apresenta um desvio significativo em relação a esta doença, pois foram registrados muito mais casos de morte do que o esperado.
    - Porto Alegre registrou menos mortes do que a quantidade esperada, para esta doença.

Interpretação:

Aproximadamente 12% das mortes registradas em Uruguaiana, são decorrentes de problemas originários no período perinatal.

- Em relação às doenças “Q” – Malformação congênitas, deformidades e anomalias cromossômicas:

    - Caxias do Sul apresenta um desvio em relação a este problema. Foram registradas mortes a mais do que o esperado para a cidade.

Interpretação:

Caxias do Sul é uma cidade industrial, o que pode levantar uma suspeita, sobre a causa de haverem mais mortes de crianças com malformação congênita. Estas conclusões são baseadas no conhecimento próprio da autora, pois este vive em Caxias do Sul e conhece a realidade da cidade.

- Em relação às doenças “R” – Sintomas, sinais e achados anormais de exames clínicos e de laboratório não classificado em outra parte.

    - Passo fundo, Santa Maria, Uruguaiana e Santo Ângelo apresentam índices altos para esta categoria. Os valores observados são mais altos

do que os valores esperados. Santa Maria apresenta uma maior incidência deste problema do que Porto Alegre.
- Porto Alegre também apresenta um desvio para esta categoria, pois houve um número muito menor de ocorrências do que era o esperado.

Interpretação

Os serviços de saúde da capital classificam melhor o problema, o que não ocorre com as cidades do interior. Santa Maria apresenta o maior valor de desvio registrado nesta tabela, o que claramente indica um problema nesta área, na cidade.

- Em relação às doenças "V" – Causas externas de morbidade e mortalidade:

  - Caxias do Sul registrou mais mortes do que o esperado para a cidade em relação a esta categoria de doença
  - Porto Alegre e Uruguaiana apresentaram valores inferiores do que o esperado.

Interpretação:

Estas doenças referem-se a mortes causadas por acidentes de trânsito, agressões e problemas causados durante procedimentos cirúrgicos auxiliares, como cateterismo entre outros. No arquivo avaliado o maior número de casos é referente a agressões e mortes em acidentes de trânsito. Estes dados apresentam a cidade de Caxias do Sul como a mais violenta em relação às demais cidades avaliadas. A cidade de Porto Alegre esta abaixo do número de casos esperados para esta categoria de doença.

5.2.4.2 Experimento 2 – Causa da Morte e Raça do falecido

1) Descrição da base utilizada

Descrita no item 5.2.4.

Nesta base foram excluídos os dados sobre raça que são ignorados ou desconhecidos. Este trabalho reduziu a quantidade de registro para 29237.

2) Análise Multivariada

Qui-quadrado

Nesta análise foram selecionados os campos:

- Raça e cor
- Causa da morte

As hipóteses definidas são:

- H_0: A causa da morte é independente da raça do falecido.
- H_1: A causa da morte é dependente da raça do falecido

- Resultado

Os valores calculados são:

- *Pearson* qui-quadrado: 283,05
- Graus de liberdade: 64

Os valores são válidos, embora os graus de liberdade sejam altos. Avaliando o *Pearson* calculado na tabela encontra-se para estes graus de liberdade um valor superior a 79,1. O que define que a hipótese H_l é validada. Isto significa dizer que a causa da morte é associada com a raça do falecido.

Análise de Resíduos

Nesta análise foram selecionados os campos:

- Raça e cor
- Causa da morte

A tabela de resíduos é gerada com o auxílio do SPSS. Esta apresentada na figura 5-12. Analisando a tabela observa-se:

- Em relação às doenças "AB" – Doenças infecciosas e parasitárias:
  - A raça branca apresenta uma incidência de mortes menor do que a quantidade esperada.
  - Em contrapartida a raça negra e a parda apresentam um maior numero de incidências do que a quantidade esperada.

Interpretação:
Na sociedade gaúcha, os negros e os pardos apresentam em geral piores condições de vida, logo estão mais expostos a este tipo de doença.

- Em relação às doenças "I" – Doenças do aparelho circulatório.
  - A raça branca apresenta uma incidência de mortes maior do que a esperada.
  - A raça negra e a raça parda apresentam um menor número de incidências do que o esperado.

Interpretação:
Nesta avaliação, a interpretação é oposta a anterior. Os brancos morrem mais do que o esperado de problemas circulatórios. Enquanto com os negros e pardos a situação se inverte.

- Em relação às doenças “P” – Algumas afecções originadas no período perinatal:

    - A raça branca apresenta uma incidência de mortes menor do que a esperada.
    - Em contrapartida a raça parda apresenta um maior número de incidências do que o esperado.

Interpretação:

O problema social pode ser a causa desta disparidade entra a raça branca e a raça parda. Socialmente os brancos são melhores atendidos no período perinatal. O que pode determinar um controle maior nas doenças originadas neste período.

- Em relação às doenças “Q” – Malformação congênita, deformidades e anomalias cromossômicas.

    - A raça parda aparece como o único desvio registrado para esta doença. Os casos de morte por este problema ocorrem mais do que o esperado entre a raça parda.

Análise de Correspondência

Os campos selecionados são os mesmos, avaliados na análise de resíduos.

- Resultado

Avaliando o gráfico, pode-se verificar existe um valor que se apresenta distante do agrupamento principal, representando graficamente um desvio. Para uma avaliação mais detalhada devem ser observados os valores apresentados nos *scores* calculados.

Os valores apresentados nos *scores* representam uma característica da base de dados utilizada nas pesquisas. A coluna referente ao percentual de casos avaliados, a coluna *Mass*, mostra a baixíssima quantidade de casos relacionados à raça parda. Somente sete casos são associados a esta raça em um total de aproximadamente vinte nove mil casos avaliados. Esta realidade dá origem a valores para o *Score in Dimension* muito afastados dos demais valores apresentados.

NUMCAU * RACA Crosstabulation

			RACA					Total
			1	2	3	4	5	
NUMCAU	AB	Count	1572	291	4	130	2	1999
		Adjusted Residual	-11,4	10,4	,5	4,3	2,3	
	CD	Count	204	24	0	11	0	239
		Adjusted Residual	-,7	1,0	-,6	,0	-,2	
	E	Count	1565	186	6	71	0	1828
		Adjusted Residual	-1,7	2,9	2,0	-1,4	-,7	
	F	Count	193	27	0	12	0	232
		Adjusted Residual	-1,7	1,8	-,6	,4	-,2	
	G	Count	549	32	0	27	0	608
		Adjusted Residual	2,5	-2,8	-1,0	-,1	-,4	
	H	Count	8	1	0	0	0	9
		Adjusted Residual	,2	,3	-,1	-,7	,0	
	I	Count	11507	992	21	539	1	13060
		Adjusted Residual	5,4	-4,2	,3	-3,2	-1,6	
	J	Count	4104	357	2	201	2	4666
		Adjusted Residual	2,3	-1,9	-2,1	-,9	,9	
	K	Count	1875	187	3	84	2	2151
		Adjusted Residual	,4	,6	-,2	-1,5	2,2	
	L	Count	40	7	0	3	0	50
		Adjusted Residual	-1,5	1,4	-,3	,5	-,1	
	M	Count	177	19	0	3	0	199
		Adjusted Residual	,9	,6	-,6	-2,1	-,2	
	N	Count	451	42	0	17	0	510
		Adjusted Residual	1,0	-,1	-,9	-1,3	-,4	
	O	Count	28	7	0	3	0	38
		Adjusted Residual	-2,4	2,2	-,2	1,0	-,1	
	P	Count	931	97	5	100	0	1133
		Adjusted Residual	-4,8	,3	2,5	7,0	-,5	
	Q	Count	324	26	0	31	0	381
		Adjusted Residual	-1,1	-1,1	-,8	3,4	-,3	
	R	Count	1004	74	2	59	0	1139
		Adjusted Residual	1,3	-2,3	,2	1,0	-,5	
	V	Count	879	71	2	43	0	995
		Adjusted Residual	1,4	-1,4	,4	-,4	-,5	
Total		Count	25411	2440	45	1334	7	29237

FIGURA 5.12 - Tabela de Análise de resíduos

Overview Row Points[a]

		Score in Dimension			Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
RACA	Mass	1	2	Inertia	1	2	1	2	Total
1	,869	,110	-,008	,001	,129	,001	,996	,003	1,000
2	,083	-,773	-,391	,005	,615	,263	,866	,132	,998
3	,002	-,865	,880	,001	,014	,025	,157	,097	,253
4	,046	-,631	,854	,003	,224	,687	,474	,518	,993
5	,000	-2,506	-2,207	,000	,019	,024	,274	,127	,401
Active Total	1,000			,010	1,000	1,000			

a. Symmetrical normalization

FIGURA 5.13 - Cargas das linhas

Overview Column Points[a]

NUMCAU	Mass	Score in Dimension 1	Score in Dimension 2	Inertia	Contribution: Of Point to Inertia of Dimension 1	Contribution: Of Point to Inertia of Dimension 2	Contribution: Of Dimension to Inertia of Point 1	Contribution: Of Dimension to Inertia of Point 2	Contribution: Of Dimension to Inertia of Point Total
AB	,068	-,883	-,171	,004	,656	,041	,974	,022	,996
CD	,008	-,162	-,144	,000	,003	,004	,386	,183	,569
E	,063	-,150	-,222	,000	,017	,064	,247	,323	,570
F	,008	-,388	-,169	,000	,015	,005	,710	,080	,790
G	,021	,373	,204	,000	,036	,018	,777	,139	,916
H	,000	,142	-1,048	,000	,000	,007	,029	,933	,962
I	,447	,126	-,010	,001	,087	,001	,938	,003	,941
J	,160	,106	-,020	,000	,022	,001	,508	,010	,518
K	,074	,002	-,179	,000	,000	,048	,000	,497	,497
L	,002	-,720	-,208	,000	,011	,002	,840	,042	,881
M	,007	,174	-,656	,000	,003	,061	,101	,853	,954
N	,017	,150	-,228	,000	,005	,019	,342	,467	,808
O	,001	-1,374	-,220	,000	,030	,001	,912	,014	,926
P	,039	-,439	,806	,002	,092	,519	,326	,654	,979
Q	,013	-,134	,739	,000	,003	,147	,044	,803	,847
R	,039	,150	,271	,000	,011	,059	,339	,660	1,000
V	,034	,156	,072	,000	,010	,004	,787	,101	,888
Active Total	1,000			,010	1,000	1,000			

a. Symmetrical normalization

FIGURA 5.14 - Cargas das colunas

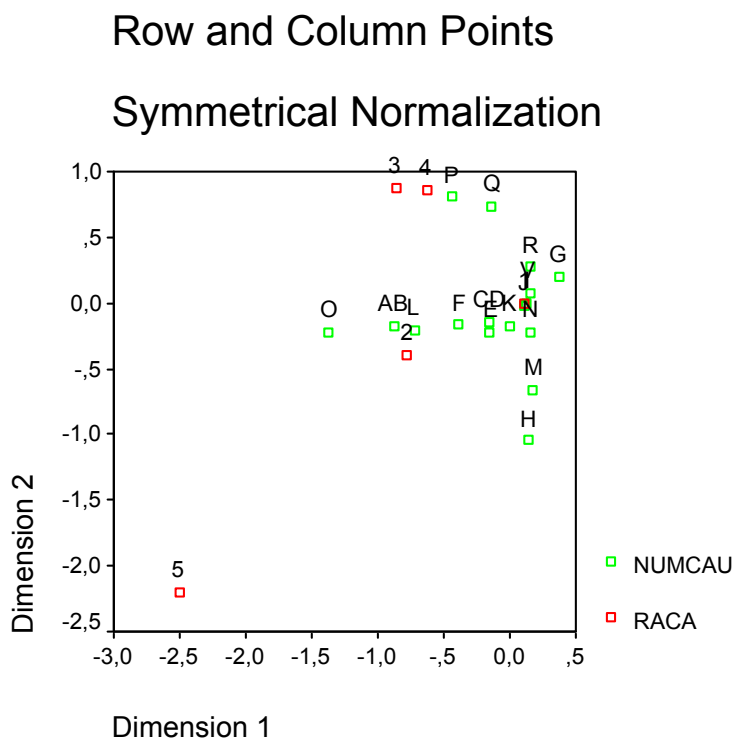


FIGURA 5-15 - Gráfico de Análise de Correspondência

6 Considerações finais

Este trabalho descreveu todas as etapas desenvolvidas para a realização do estudo sobre detecção de desvios em dados da Secretaria da Saúde do Rio Grande do Sul.

No início das atividades, foi realizado um estudo sobre a área de descoberta de conhecimento, que permitiu a verificação da existência de diversas técnicas disponíveis para o processo, como associação, agrupamento, classificação, detecção de desvios.

O foco principal deste trabalho foi voltado para a detecção de desvios em bases de dados, que é uma técnica muito importante do processo de mineração. A pesquisa realizada para identificar e validar o material literário sobre este assunto não foi uma tarefa fácil.

A grande parte da bibliografia encontrada descrevia o processo de descoberta de conhecimento como um todo, mas pouco material foi encontrado que apresentasse de forma aprofundada a técnica de detecção de desvios.

A maior parte do material encontrado, sobre detecção de desvios, está disponível em artigos isolados ou em anais de congressos das áreas afins. Logo, grande parte deste material apresenta o assunto de forma sucinta.

A partir desta situação, houve a necessidade de um estudo mais completo sobre o assunto. Baseado no fato que grande parte das técnicas de detecção de desvios utilizam a estatística como base de trabalho, esta pesquisa voltou-se ao estudo das avaliações estatísticas.

Nesta fase foram estudadas algumas técnicas estatísticas, sendo que muitas podem ser aplicadas sobre os dados estudados. A seleção das técnicas avaliadas foi baseada na relação entre o tipo de dado armazenado e de que forma está expressa a informação dentro da base de dados. Estes parâmetros podem determinar o sucesso da avaliação.

Foi necessário conhecer e identificar como as técnicas estatísticas efetuavam as avaliações. Neste ponto foi considerada a necessidade de desenvolver um software simples que aplicasse diretamente na base o estudo realizado, promovendo assim uma boa compreensão da forma de avaliação estatística dos dados.

O software permitiu a seleção e avaliação dos dados trabalhados, de forma simplificada, o que facilitou a compreensão de como são calculados os valores e avaliados os resultados obtidos. As técnicas desenvolvidas no protótipo foram o cálculo do desvio-padrão para a avaliação univariada dos registros quantitativos, e o cálculo do qui-quadrado para a avaliação multivariada dos registros qualitativos.

A utilização do software abrangia a tarefa de seleção dos dados. Permitindo que a partir de uma grande base de dados fossem formadas bases menores, as quais facilitaram o processo de avaliação.

Mesmo obtendo resultados satisfatórios, foi identificada a necessidade de ampliar as avaliações estatísticas. Pois, nem sempre os dados permitiam uma verificação válida ou ainda não resultavam valores satisfatórios para a pesquisa. Nesta fase foram definidos dois caminhos a seguir: O primeiro indicaria uma implementação a mais no software desenvolvido e o segundo implicaria na utilização de um software comercial específico.

A opção escolhida foi à utilização de um software estatístico específico o SPSS, pois permitiria uma maior agilidade nas pesquisas. A etapa seguinte foi aplicar as bases de dados no software escolhido.

Os experimentos realizados a partir da metodologia utilizada no trabalho, apresentaram resultados interessantes para a pesquisa. As avaliações de determinadas informações agregaram conhecimento ao trabalho realizado.

Depois da aplicação de todo o roteiro de trabalho citado anteriormente, conclui-se:

A metodologia aplicada foi satisfatória para a detecção de desvios, pois permitiu a identificação, avaliação e a sistematização do processo desenvolvido. Em alguns momentos da pesquisa foi necessário o retorno às etapas anteriores, pois, baseado nos resultados obtidos era necessário reorganizar os dados e avaliá-los novamente. Os resultados obtidos mostraram que a estrutura do arquivo, a definição do tipo de dado e a forma como esta expressa a informação dentro do arquivo, podem dificultar ou auxiliar no processo de detecção de desvios.

Durante um processo de detecção de desvios é primordial que sejam traçados objetivos para nortear as avaliações que serão efetuadas. E ainda para capacitar a boa avaliação dos resultados obtidos. Bons objetivos são os resultado de um bom estudo do domínio e estudo da aplicação.

As técnicas estatísticas utilizadas foram definidas de acordo com um estudo prévio das necessidades encontradas. Os resultados obtidos mostraram uma eficiência das técnicas aplicadas, mas ao mesmo tempo uma limitação das mesmas, pois a detecção de desvio engloba uma série de itens que devem ser avaliados em conjunto. E as técnicas estatísticas utilizadas permitem a avaliação simultânea de apenas duas variáveis por vez.

A aplicação do cálculo do desvio-padrão, onde somente é considerada uma variável por vez, obteve resultados satisfatórios, mesmo sendo uma avaliação muito simples. No cálculo do qui-quadrado, eram testadas duas variáveis por vez, mesmo assim os resultados obtidos nem sempre eram possíveis de avaliação. Este problema ocorria, pois o arquivo testado apresentava uma grande quantidade de valores para um determinado atributo do arquivo.

A forma como os dados estão armazenados e principalmente a informação que eles expressam são os fatores que podem determinar o tipo de avaliação que deve ser efetuada. Ou melhor, o tipo de técnica estatística que deve ser aplicada. A estatística oferece uma gama muito grande de técnicas de avaliação. Cada uma delas é definida para um tipo de aplicação, ou, definida para um tipo de resultado esperado.

A utilização do SPSS foi um recurso indispensável para a avaliação dos dados, pois permitiu um direcionamento da pesquisa, sem a preocupação de implementar o cálculo estatístico, tornando o trabalho mais fácil de ser realizado.

A sistemática de análise dos resultados obedeceu aos resultados obtidos em cada uma das técnicas aplicadas. Dependendo do resultado, de acordo com a base era definida a próxima avaliação a ser executada. Este tipo de sistemática de trabalho é altamente dependente do usuário, da mesma forma o usuário deve ter um bom conhecimento de estatística.

6.1 Trabalhos Futuros

Em relação ao protótipo as sugestões de trabalhos futuros abrangem a generalização do programa em relação a utilização de outras bases de dados, pois atualmente o protótipo só aceita duas bases de dados específicas.

O protótipo deveria permitir o retorno de alguns passos da execução. Ao encerrar a pesquisa o protótipo não permite o retorno ao passo anterior para liberar nova avaliação, desta forma o programa deve ser reiniciado. Aprimorando o protótipo seria mais simples realizar outras avaliações em um curto espaço de tempo, proporcionando assim um ganho de tempo.

Caso o protótipo fosse aprimorado, o controle da geração de arquivos de saídas deveria ser implantado. Desta forma toda a vez que uma avaliação fosse iniciada, o arquivo de resposta anterior não seria prejudicado.

Outro trabalho futuro seria identificar e estudar outras técnicas estatísticas para aplicar na detecção de desvios. Existem diversas técnicas e cada uma pode se adequar melhor aos objetivos definidos para a pesquisa realizada. Este estudo seria mais completo, se outras bases de dados com diferentes características fossem avaliadas.

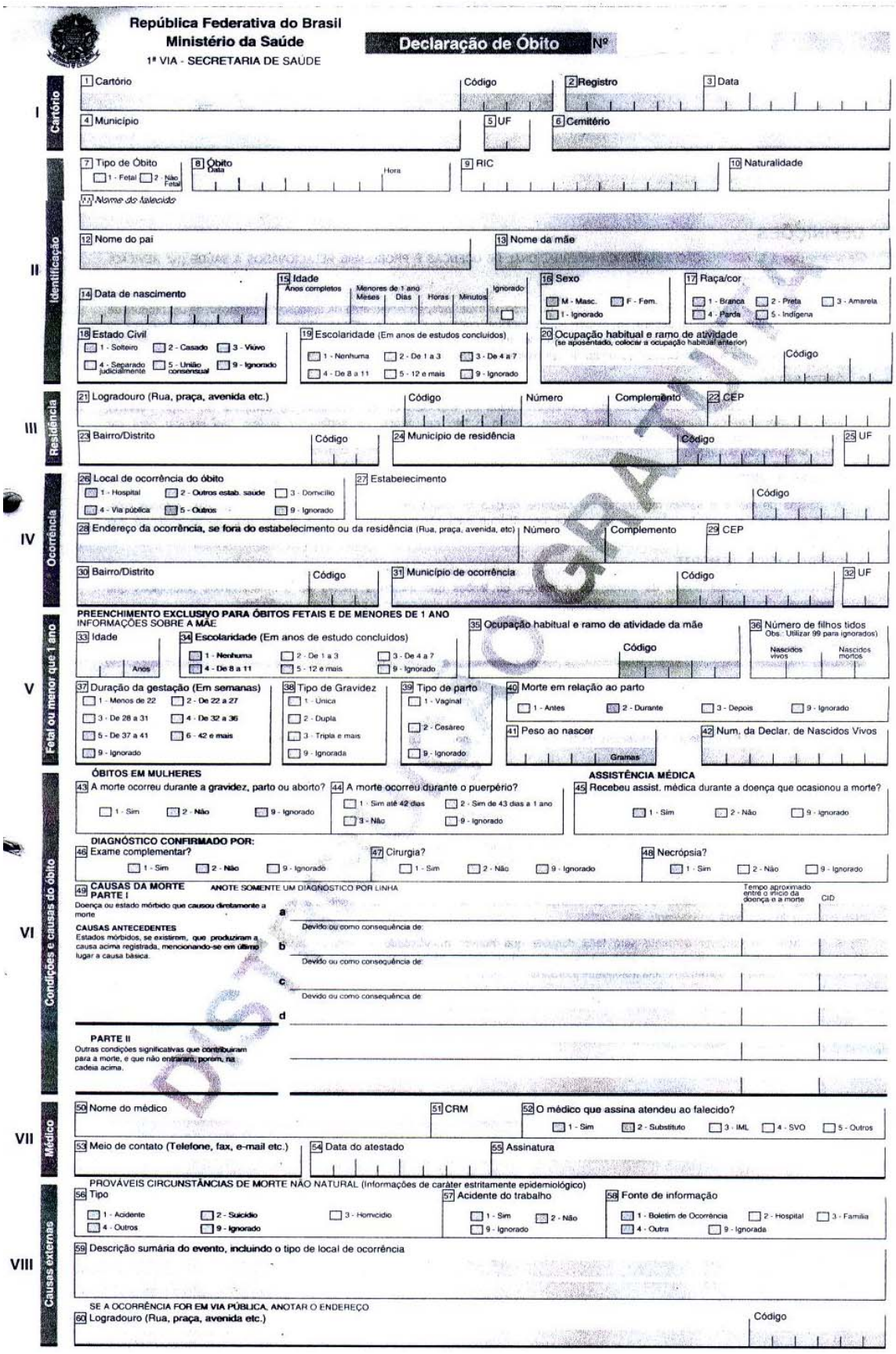
Também poderia ser feito um estudo mais aprofundado sobre a detecção de desvios como um dos objetivos a serem alcançados na aplicação de outra técnica de mineração como associação, classificação entre outras.

Anexo 1 Qui-quadrado

Graus de liberdade

	Área à direita de x											
υ	0,005	0,01	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,975	0,99	0,995
1	7,88	6,63	3,84	2,71	1,3233	0,4549	0,1015	0,0158	0,0039	0,0010	0,0002	0,0000
2	10,60	9,21	5,99	4,61	2,7726	1,3863	0,5754	0,2107	0,1026	0,0506	0,0201	0,0100
3	12,84	11,34	7,81	6,25	4,1083	2,3660	1,2125	0,5844	0,3518	0,2158	0,1148	0,0717
4	14,86	13,28	9,49	7,78	5,39	3,36	1,92	1,06	0,7107	0,4844	0,2971	0,2070
5	16,75	15,09	11,07	9,24	6,63	4,35	2,67	1,61	1,1455	0,8312	0,5543	0,4118
6	18,55	16,81	12,59	10,64	7,84	5,35	3,45	2,20	1,6354	1,2373	0,8721	0,6757
7	20,28	18,48	14,07	12,02	9,04	6,35	4,25	2,83	2,1673	1,6899	1,2390	0,9893
8	21,95	20,09	15,51	13,36	10,22	7,34	5,07	3,49	2,73	2,18	1,65	1,34
9	23,59	21,67	16,92	14,68	11,39	8,34	5,90	4,17	3,33	2,70	2,09	1,73
10	25,19	23,21	18,31	15,99	12,55	9,34	6,74	4,87	3,94	3,25	2,56	2,16
11	26,76	24,73	19,68	17,28	13,70	10,34	7,58	5,58	4,57	3,82	3,05	2,60
12	28,30	26,22	21,03	18,55	14,85	11,34	8,44	6,30	5,23	4,40	3,57	3,07
13	29,82	27,69	22,36	19,81	15,98	12,34	9,30	7,04	5,89	5,01	4,11	3,57
14	31,32	29,14	23,68	21,06	17,12	13,34	10,17	7,79	6,57	5,63	4,66	4,07
15	32,80	30,58	25,00	22,31	18,25	14,34	11,04	8,55	7,26	6,26	5,23	4,60
16	34,27	32,00	26,30	23,54	19,37	15,34	11,91	9,31	7,96	6,91	5,81	5,14
17	35,72	33,41	27,59	24,77	20,49	16,34	12,79	10,09	8,67	7,56	6,41	5,70
18	37,16	34,81	28,87	25,99	21,60	17,34	13,68	10,86	9,39	8,23	7,01	6,26
19	38,58	36,19	30,14	27,20	22,72	18,34	14,56	11,65	10,12	8,91	7,63	6,84
20	40,00	37,57	31,41	28,41	23,83	19,34	15,45	12,44	10,85	9,59	8,26	7,43
21	41,40	38,93	32,67	29,62	24,93	20,34	16,34	13,24	11,59	10,28	8,90	8,03
22	42,80	40,29	33,92	30,81	26,04	21,34	17,24	14,04	12,34	10,98	9,54	8,64
23	44,18	41,64	35,17	32,01	27,14	22,34	18,14	14,85	13,09	11,69	10,20	9,26
24	45,56	42,98	36,42	33,20	28,24	23,34	19,04	15,66	13,85	12,40	10,86	9,89
25	46,93	44,31	37,65	34,38	29,34	24,34	19,94	16,47	14,61	13,12	11,52	10,52
26	48,29	45,64	38,89	35,56	30,43	25,34	20,84	17,29	15,38	13,84	12,20	11,16
27	49,65	46,96	40,11	36,74	31,53	26,34	21,75	18,11	16,15	14,57	12,88	11,81
28	50,99	48,28	41,34	37,92	32,62	27,34	22,66	18,94	16,93	15,31	13,56	12,46
29	52,34	49,59	42,56	39,09	33,71	28,34	23,57	19,77	17,71	16,05	14,26	13,12
30	53,67	50,89	43,77	40,26	34,80	29,34	24,48	20,60	18,49	16,79	14,95	13,79
40	66,77	63,69	55,76	51,81	45,62	39,34	33,66	29,05	26,51	24,43	22,16	20,71
50	79,49	76,15	67,50	63,17	56,33	49,33	42,94	37,69	34,76	32,36	29,71	27,99
60	91,95	88,38	79,08	74,40	66,98	59,33	52,29	46,46	43,19	40,48	37,48	35,53
70	104,21	100,43	90,53	85,53	77,58	69,33	61,70	55,33	51,74	48,76	45,44	43,28
80	116,32	112,33	101,88	96,58	88,13	79,33	71,14	64,28	60,39	57,15	53,54	51,17
90	128,30	124,12	113,15	107,57	98,65	89,33	80,62	73,29	69,13	65,65	61,75	59,20
100	140,17	135,81	124,34	118,50	109,14	99,33	90,13	82,36	77,93	74,22	70,06	67,33

Anexo 2 Declaração de Óbito



República Federativa do Brasil
Ministério da Saúde
1ª VIA - SECRETARIA DE SAÚDE

Declaração de Óbito Nº

I Cartório
1 Cartório | Código | 2 Registro | 3 Data
4 Município | 5 UF | 6 Cemitério

II Identificação
7 Tipo de Óbito (1 - Fetal, 2 - Não Fetal) | 8 Óbito Data / Hora | 9 RIC | 10 Naturalidade
11 Nome do falecido
12 Nome do pai | 13 Nome da mãe
14 Data de nascimento | 15 Idade (Anos completos / Menores de 1 ano: Meses, Dias, Horas, Minutos / Ignorado) | 16 Sexo (M - Masc., F - Fem., 1 - Ignorado) | 17 Raça/cor (1 - Branca, 2 - Preta, 3 - Amarela, 4 - Parda, 5 - Indígena)
18 Estado Civil (1 - Solteiro, 2 - Casado, 3 - Viúvo, 4 - Separado judicialmente, 5 - União consensual, 9 - Ignorado)
19 Escolaridade (Em anos de estudos concluídos) (1 - Nenhuma, 2 - De 1 a 3, 3 - De 4 a 7, 4 - De 8 a 11, 5 - 12 e mais, 9 - Ignorado)
20 Ocupação habitual e ramo de atividade (se aposentado, colocar a ocupação habitual anterior) | Código

III Residência
21 Logradouro (Rua, praça, avenida etc.) | Código | Número | Complemento | 22 CEP
23 Bairro/Distrito | Código | 24 Município de residência | Código | 25 UF

IV Ocorrência
26 Local de ocorrência do óbito (1 - Hospital, 2 - Outros estab. saúde, 3 - Domicílio, 4 - Via pública, 5 - Outros, 9 - Ignorado)
27 Estabelecimento | Código
28 Endereço da ocorrência, se fora do estabelecimento ou da residência (Rua, praça, avenida, etc) | Número | Complemento | 29 CEP
30 Bairro/Distrito | Código | 31 Município de ocorrência | Código | 32 UF

V Fetal ou menor que 1 ano
PREENCHIMENTO EXCLUSIVO PARA ÓBITOS FETAIS E DE MENORES DE 1 ANO
INFORMAÇÕES SOBRE A MÃE
33 Idade (Anos)
34 Escolaridade (Em anos de estudo concluídos) (1 - Nenhuma, 2 - De 1 a 3, 3 - De 4 a 7, 4 - De 8 a 11, 5 - 12 e mais, 9 - Ignorado)
35 Ocupação habitual e ramo de atividade da mãe | Código
36 Número de filhos tidos (Obs.: Utilizar 99 para ignorados) (Nascidos vivos, Nascidos mortos)
37 Duração da gestação (Em semanas) (1 - Menos de 22, 2 - De 22 a 27, 3 - De 28 a 31, 4 - De 32 a 36, 5 - De 37 a 41, 6 - 42 e mais, 9 - Ignorado)
38 Tipo de Gravidez (1 - Única, 2 - Dupla, 3 - Tripla e mais, 9 - Ignorada)
39 Tipo de parto (1 - Vaginal, 2 - Cesáreo, 9 - Ignorado)
40 Morte em relação ao parto (1 - Antes, 2 - Durante, 3 - Depois, 9 - Ignorado)
41 Peso ao nascer (Gramas)
42 Num. da Declar. de Nascidos Vivos

VI Condições e causas do óbito
ÓBITOS EM MULHERES
43 A morte ocorreu durante a gravidez, parto ou aborto? (1 - Sim, 2 - Não, 9 - Ignorado)
44 A morte ocorreu durante o puerpério? (1 - Sim até 42 dias, 2 - Sim de 43 dias a 1 ano, 3 - Não, 9 - Ignorado)
ASSISTÊNCIA MÉDICA
45 Recebeu assist. médica durante a doença que ocasionou a morte? (1 - Sim, 2 - Não, 9 - Ignorado)
DIAGNÓSTICO CONFIRMADO POR:
46 Exame complementar? (1 - Sim, 2 - Não, 9 - Ignorado)
47 Cirurgia? (1 - Sim, 2 - Não, 9 - Ignorado)
48 Necrópsia? (1 - Sim, 2 - Não, 9 - Ignorado)
49 CAUSAS DA MORTE PARTE I — ANOTE SOMENTE UM DIAGNÓSTICO POR LINHA (Tempo aproximado entre o início da doença e a morte | CID)
Doença ou estado mórbido que causou diretamente a morte — a
CAUSAS ANTECEDENTES
Estados mórbidos, se existirem, que produziram a causa acima registrada, mencionando-se em último lugar a causa básica.
Devido ou como consequência de: b
Devido ou como consequência de: c
Devido ou como consequência de: d
PARTE II
Outras condições significativas que contribuíram para a morte, e que não entraram, porem, na cadeia acima.

VII Médico
50 Nome do médico | 51 CRM | 52 O médico que assina atendeu ao falecido? (1 - Sim, 2 - Substituto, 3 - IML, 4 - SVO, 5 - Outros)
53 Meio de contato (Telefone, fax, e-mail etc.) | 54 Data do atestado | 55 Assinatura

VIII Causas externas
PROVÁVEIS CIRCUNSTÂNCIAS DE MORTE NÃO NATURAL (Informações de caráter estritamente epidemiológico)
56 Tipo (1 - Acidente, 2 - Suicídio, 3 - Homicídio, 4 - Outros, 9 - Ignorado)
57 Acidente do trabalho (1 - Sim, 2 - Não, 9 - Ignorado)
58 Fonte de informação (1 - Boletim de Ocorrência, 2 - Hospital, 3 - Família, 4 - Outra, 9 - Ignorada)
59 Descrição sumária do evento, incluindo o tipo de local de ocorrência
SE A OCORRÊNCIA FOR EM VIA PÚBLICA, ANOTAR O ENDEREÇO
60 Logradouro (Rua, praça, avenida etc.) | Código

Referências

[ARN 96] ARNING, A.; AGRAWAL, R. ; RAGHAVAN, P. A Linear Method for Deviation Detection in Large Databases. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE IN DATABASES AND DATA MINING, 2., 1996, Portland, Oregon. **Proceedings ...** Disponível em : <http://www.almaden.ibm.com/cs/quest/publications.html#dev>. Acesso em: mar. 2000.

[CAB 97] CABENA, P. et al. **Discovering data mining:** from concept to implementation. Upper Saddle River: Prentice-Hall PTR, 1997.

[CHA 99] CHAPMAN, P. et al. **The CRISP-DM Process Model**. [S.l.]: CRISP-DM Consortium, 1999. Disponível em: <http://www.crisp-dm.org>. Acesso em: maio 2001.

[DOM 2003] DOMINGUES, Miriam Lúcia Campos Serra. **Mineração de dados utilizando aprendizado não-supervisionado**: um estudo de caso para bancos de dados da saúde. 2003. 128 p. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre

[FAY 96] FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery: An Overview. In: FAYYAD, Usama M. et al. **Advances in Knowledge Discovery and Data Mining**. Menlo Park: MIT Press, 1996. 611 p. p. 1-34.

[FEL 96] FELDENS, Miguel A. **Descoberta de Conhecimento Aplicada à Detecção de Anomalias em Bases de Dados**. Trabalho Individual (Mestrado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre,1996.

[FEL 97] FELDENS, Miguel A. **Engenharia da Descoberta do Conhecimento em Bases de Dados:** Estudo e Aplicada na Área de Saúde. 1997. Dissertação (Mestrado em ciências da Computação) - Instituto de Informática, UFRGS, Porto Alegre.

[FRI 97] FRIGERI, Sandra. **Descoberta de conhecimento em bases de dados e mineração de dados com uso de redes neurais artificiais**. 1997. 100 f. Trabalho Individual (Mestrado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

[HAN 2001] HAN, Jiawei; KAMBER, Micheline. **Data mining:** concepts and techniques. San Francisco: Morgan Kaufmann, 2001.

[KNO 2002] KNORR, Edwin. **Outliers and Data Mining:** Finding Exceptios in Data. 2002. Doctor of Philosophy thesis – Department of Computer Science, the University of British Columbia. Disponível em: <http://www.cs.ubc.ca/grads/resources/thesis/May02/Ed_Knorr.pdf>. Acesso em: fev. 2003.

[LEE 97] LEE,W.; STOLFO,S.J. **Learning Patterns from Unix Process Execution Traces for Instrusion Detection**.[S.l.]: Computer Scienc Departament Columbia University, 1997.

[MAT93] MATHEUS, C. J.; CHAN, P. K.; PIATETSKY-SHAPIRO, G. Systems for

knowledge discovery in databases. **IEEE Transactions on Knowledge and Data Engineering**, New York, v. 5, n. 6, Dec. 1993.

[NOT 97] NOTARI, Daniel. **Aplicação de redes neurais artificiais à mineração de dados**. 1997. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Departamento de Informática, Universidade de Caxias do Sul, Caxias do Sul.

[PAR 89] PARSAYE, K. et al. **Intelligent Databases**: object-oriented, deductive and hypermedia technologies. New York: John Willey, 1989.

[PER 2001] PEREZ, César. **Técnicas estatísticas con SPSS.** Madrid: Pearson Educación S. A: Prentice Hall, 2001

[PER 99] PEREIRA, Julio Cesar Rodrigues. **Análise de Dados Qualitativos. Estratégias Metodológicas para as Ciências da Saúde, Humanas e Sociais.** São Paulo: Ed. da USP, 1999.

[PIA 93] PIATETSKY-SHAPIRO, G. et al. **KDD-93**: Progress and challenges in knowledge discovery in databases. [S.l: s.n.], 1993.

[RID94] RIDDLE, P; SEGAL, R.; ETZIONI, O. Representation design and brute-force induction in a Boeing manufacturing domain. **Applied Artificial Intelligence,** [S.l.],n.8,1994.

[ROS 97] ROSS, Sheldon M. **Introduction to Probability Models**. Berkeley, California: Academic Press, 1997. p. 74-77.

[SPI 93] SPIEGEL, M. R. **Estatística** . 3.ed. São Paulo: Makron Bokks do Brasil, 1993. p.287.

[STE 81] STEVENSON, Willian J. **Estatística Aplicada a Administração.** São Paulo: Harbra, 1981.

[STO 97] STOLFO,S.J. et al. **Credit Card Fraud Detection Using Meta-Learning**.[S.l.]: Computer Scienc Departament Columbia University, 1997.

[SUS 2001] SISTEMA ÚNICO DE SAÚDE. **SIH/SUS: Sistema de Informações Hospitalares do Sistema Único de Saúde**. Brasília: Ministério da Saúde, SUS, 2001.

[TWO 98] TWO CROWS CORPORATION. **Introduction to Data Mining and Knowledge Discovery.** Potomac, USA, 1998. Disponível em: <http://www.twocrows.com> Acesso em: maio 2000.

[ZYT93] ZYTKOW, J. M.; ZEMBOWICZ, R. Database exploration in search of regularities. **Journal of Intelligent Information Systems,** Dordrecht, Holland, v. 2, n. 1, 1993.