UNIVERSIDADE FEDERAL DE SANTA CATARINA

USO DE DATA MINING E SISTEMAS DE INFORMAÇÕES GEOGRÁFICAS NO APOIO A TOMADA DE DECISÕES

Rodrigo Benincá Machado

FLORIANÓPOLIS
2005 / 1

UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

USO DE DATA MINING E SISTEMAS DE INFORMAÇÕES GEOGRÁFICAS NO APOIO A TOMADA DE DECISÕES

Rodrigo Benincá Machado

Trabalho de conclusão de curso apresentado como parte dos requisitos para obtenção do grau de Bacharel em Sistemas de Informação.

Orientador:	*Paulo José Ogliari*
Banca Avaliadora:	*Dalton F. Andrade* *José Messias Bastos*

FLORIANÓPOLIS
2005 / 1

“Todas as coisas são parecidas,

mas coisas mais próximas se parecem mais que coisas mais distantes”

Waldo Tobler
Primeira Lei da Geografia

Oferecimentos

À minha mãe Marlene Benincá Gomes.

Agradecimentos

Agradeço ao meu orientador Prof. Paulo José Ogliari por ter aceito minha proposta de trabalho de conclusão de curso e pela confiança e orientação nesse último ano.

Da mesma forma, agradeço aos professores da banca avaliadora, Prof. Dalton F. Andrade e Prof. José Messias Bastos, pelas suas sugestões, contribuições e apoio.

Ao Juliano Anderson Pacheco registro meus agradecimentos por sua ajuda, especialmente pela sua sugestão de utilizar um sistema de informações geográficas em meu trabalho.

Agradeço aos colegas de apartamento, William Sant'Ana, principalmente, por ter me apresentado ao professor Messias e ao colega Rodrigo Campiolo pela ajuda na revisão textual.

À minha amiga Larissa Costa da Mata pela revisão do texto e pelas conversas e momentos agradáveis no decorrer desses últimos anos.

Às "tias" da biblioteca do IBGE que me auxiliaram com o estudo e o fornecimento das bases de dados utilizadas nas análises.

E, se esqueci de alguém, peço desculpas e agradeço mesmo assim, todos foram muito importantes.

Resumo

MACHADO, Rodrigo Benincá. **Uso de Data Mining e Sistemas de Informações Geográficas no Apoio a Tomada de Decisões**. 2005. 72 f. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação). Curso de Sistemas de Informação, Universidade Federal de Santa Catarina, Florianópolis.

Esse trabalho de conclusão de curso objetivou desenvolver um documento que contivesse o resultado da distribuição espacial de setores censitários do município de Florianópolis semelhantes quanto ao nível de instrução e renda média do responsável pelo domicílio. O alcance do objetivo acima se deu por meio do estudo e da aplicação do processo de descoberta de conhecimento (KDD) e do emprego de um sistema de informações geográficas. A aplicação de todas as etapas do processo de descoberta de conhecimento se deu com o uso da técnica de *cluster analysis* que possibilitou a formação dos grupos de setores censitários semelhantes. A distribuição espacial dos grupos encontrados pelo KDD foi obtida pelo emprego de um sistema de informações geográficas.

Como resultado desse estudo, podemos concluir que a aplicação do processo de descoberta de conhecimento permitiu reunir os setores censitários de forma homogênea quanto à renda média e nível de instrução do responsável pelo domicílio. Além disso, o uso do sistema de informações geográficas permitiu obter uma visão clara da posição espacial dos *clusters* formados e identificar tanto as áreas que podem ser consideradas mais atraentes comercialmente, como as áreas que necessitam de maior atenção governamental.

Palavras-chave: análise de conglomerados, mineração de dados, processo de descoberta de conhecimento, sistemas de informações geográficas.

Abstract

MACHADO, Rodrigo Benincá. **Use of Data Mining and Geographic Information Systems in the Decision's Support**. 2005. 72 p. Course Conclusion Paper (Bachelor's degree in Information Systems). Information System Course. Federal University of Santa Catarina, Florianópolis.

This paper aimed to develop a report on the results of space distribution of census sectors of Florianópolis which where gotten into groups according to the education level and to the average income of the responsible for the residence. The objective mentioned above was reached by means of the study and the use of knowledge discovery in data bases (KDD) and by the use of a geographic information system. The application of all the stages of knowledge discovery in data bases was accomplished through the cluster analysis technique. In addition, the space distribution of groups formed by KDD was made by means of a geographic information systems.

As a result of this study, one may conclude that the application of knowledge discovery in data bases allowed to gather the census sectors together uniformly in relation to the average income and to the education level of the responsible for the residence. Furthermore, the use of geographic information system allowed to observe clearly the clusters location and to identify as much the commercially attractive areas as the areas which must be looked out by the city government.

Key-words: cluster analysis, data mining, knowledge discovery in data bases, geographic information systems.

Sumário

Lista de figuras

Lista de tabelas

1 – Introdução

A evolução tecnológica que ocorre nas bases materiais de nossa sociedade disponibiliza às instituições privadas e públicas a capacidade de produzir e armazenar grandes quantidades de dados referentes aos seus respectivos negócios. Esses dados podem ser utilizados para que as transações nessas empresas se tornem mais eficazes.

Nas duas últimas décadas houve um grande crescimento na quantidade de dados produzidos e armazenados em meio eletrônico. O valor desses dados está relacionado à capacidade de extrair informações úteis ao suporte de decisões operacionais, táticas e estratégicas. É possível que existam ainda, padrões ou tendências úteis que, se descobertos, podem ser empregados, por exemplo, para auxiliar em um processo de decisão em uma empresa (AEDB, 2005). Pode ser citada como exemplo de fonte de dados para análises a que se encontra nas bases de dados do censo demográfico realizado pelo IBGE.

A cada 10 anos o IBGE (Instituto Brasileiro de Geografia e Estatística) realiza o censo da população pesquisando de maneira completa variáveis demográficas, níveis de nupcialidade e fecundidade, condições de trabalho, educação e renda e características dos domicílios. O censo é efetivado em dois estágios, dos quais o primeiro se dá por meio da aplicação de um conjunto de questões básicas a toda a população, e o segundo, por meio da abordagem de uma amostra através de um questionário mais abrangente. Em todo o território nacional foram selecionados 5.304.711 domicílios para responder ao questionário da amostra, o que significou uma fração amostral da ordem de 11,7%.

É possível se extrair informações proveitosas a partir de uma base de dados como a do IBGE através de técnicas de *data mining*. O *data mining* é parte de um processo maior, chamado de processo de descoberta de conhecimento, que tem o objetivo de otimizar e automatizar a descrição das tendências e padrões presentes nos dados (OGLIARI, 2004). A informação e conhecimento obtidos nessa operação podem ser aproveitados tanto em aplicações de *business management*, controle de produção e análise de mercado como em aplicações em projetos de engenharia e exploração científica.

Esse trabalho de conclusão de curso faz uso da técnica de *data mining* conhecida como *cluster analysis* e de um sistema de informações geográficas para atingir os objetivos que serão apresentados a seguir.

1.1 Objetivos

1.1.1 Geral

O objetivo geral desse trabalho é desenvolver um documento que contenha o resultado da distribuição espacial de setores censitários que sejam semelhantes quanto ao nível de instrução e renda média do responsável pelo domicílio. Estes resultados servirão como artefato no processo de escolha do local mais adequado para instalação de empresas comerciais.

1.1.2 Específicos

Os objetivos específicos são:

- Aplicar o processo de descoberta de conhecimento para definir os grupos de setores censitários semelhantes quanto ao nível de instrução e renda média do responsável pelo domicílio.
- Empregar um sistema de informações geográficas para distribuir espacialmente os grupos de setores censitários formados no processo de descoberta de conhecimento.

1.2 Justificativa

É essencial que as empresas estejam cientes dos resultados de seus negócios bem como da implicação destes para a corporação. Tanto as instituições privadas quanto as públicas precisam de informações para que sejam tomadas as decisões mais adequadas às suas transações.

É através dos censos populacionais que as instituições governamentais obtêm informações a respeito da situação de vida da população de uma determinada área. Além disso, são os censos que fornecem informações necessárias à definição de políticas públicas municipais, estaduais ou federais e para a resolução de investimentos, sejam eles provenientes de qualquer de instância das iniciativas pública ou privada (IBGE, 2005).

A análise do censo demográfico também beneficia a sociedade dos seguintes meios: empresários podem examiná-la durante a escolha de serviços ou locais mais adequados à instalação de suas companhias (supermercados, escolas, creches, cinemas, restaurantes); a

população pode solicitar maior atenção de parte dos governantes nos problemas específicos de sua região; os sindicatos podem analisar o perfil da mão-de-obra brasileira, etc. (IBGE, 2005).

1.3 Metodologia

A metodologia de desenvolvimento desse trabalho pode ser dividida em três partes: a fundamentação teórica, a aplicação do processo de descoberta de conhecimento para formação dos grupos e a distribuição espacial dos grupos formados por meio de um sistema de informações geográficas.

A primeira seção, fundamentação teórica introduz os conceitos de cada etapa do processo de descoberta de conhecimento, enfatizando a técnica de *cluster analysis* e os conceitos de sistema de informações geográficas.

Após a conclusão da fundamentação teórica foi dado início às atividades do processo de descoberta de conhecimento por meio da análise da base de dados e da seleção, preparação e transformação das variáveis a serem consideradas na etapa de *data mining*. Estas variáveis, selecionadas na base de dados do censo demográfico, devem estar relacionadas ao nível de instrução e à renda dos responsáveis por domicílios no município de Florianópolis.

Para finalizar as atividades aqui propostas, fez-se necessário que os grupos encontrados no processo de descoberta de conhecimento fossem distribuídos espacialmente no mapa do município de Florianópolis com uso de um sistema de informações geográficas.

A execução das três fases descritas acima permitiram o alcance dos objetivos expostos na seção 1.1 desse documento.

1.4 Estrutura do trabalho

A redação do presente estudo foi desenvolvida em quatro capítulos: Introdução; Fundamentação teórica; Desenvolvimento do trabalho e Conclusões e trabalhos futuros.

Está presente no capítulo 1 a introdução do trabalho, seus objetivos, metodologia e estrutura do documento. O capítulo 2 expõe brevemente a evolução dos sistemas de banco de dados, apresenta os conceitos do processo de descoberta de conhecimento e introduz os sistemas de informações geográficas. Além disso, também é descrita a técnica de análise de conglomerados, que foi a técnica utilizada nesse trabalho.

No capítulo 3 estão expostas as atividades que foram necessárias para atingir os objetivos do trabalho, como a apresentação da base de dados utilizada e o tratamento aplicado às variáveis selecionadas, a execução do algoritmo de análise de agrupamentos e descrição de seus resultados, finalizando com a distribuição espacial dos grupos com a utilização de um sistema de informações geográficas.

A redação do trabalho é concluída no capítulo 4 com a conclusão e a sugestão de trabalhos futuros.

2 – Fundamentação Teórica

Esse capítulo expõe as considerações provenientes da pesquisa bibliográfica do trabalho distribuídas em quatro partes. A primeira delas trata da evolução dos sistemas de banco de dados, desde os sistemas de arquivos até o *data mining*. A segunda parte diz respeito ao processo de descoberta de conhecimento e apresentada as etapas deste e suas metodologias, tarefas e técnicas. Em seguida, na terceira seção, é descrita a técnica de análise de conglomerados, a qual é utilizada para o agrupamento dos setores censitários. Já o último tópico desse capítulo introduz os conceitos de sistemas de informações geográficas (GIS).

2.1 A evolução dos sistemas de banco de dados

Segundo KAMBER (2001), *data mining* pode ser visto como um dos resultados da evolução natural da tecnologia da informação. Pode-se observar um caminho evolutivo na indústria de banco de dados pelo desenvolvimento das seguintes funcionalidades:

- *Processamento primitivo de arquivos:* a própria aplicação era responsável por manter e gerenciar a estrutura de dados definida pelo desenvolvedor do sistema;
- *Gerenciamento de dados:* incluindo recuperação e base de dados de processamento transacional;
- *Entendimento e análise de dados:* envolvendo *data warehouse* e *data mining*.

A partir dos anos sessenta, tecnologia da informação e banco de dados tem se envolvido sistematicamente desde os primitivos sistemas de processamento de arquivos até os sofisticados e poderosos sistemas de banco de dados. Com a evolução da tecnologia de banco de dados os usuários ganharam um acesso mais flexível e conveniente aos dados através de linguagens de consulta, interfaces gráficas com o usuário, otimização no processamento de consultas e gerenciamento de transação. Métodos eficientes para processamento de transações on-line (OLTP) têm contribuído para a evolução e aceitação em larga escala da tecnologia relacional como a principal ferramenta para armazenamento eficiente, recuperação e gerenciamento de grande quantidade de dados (KAMBER, 2001).

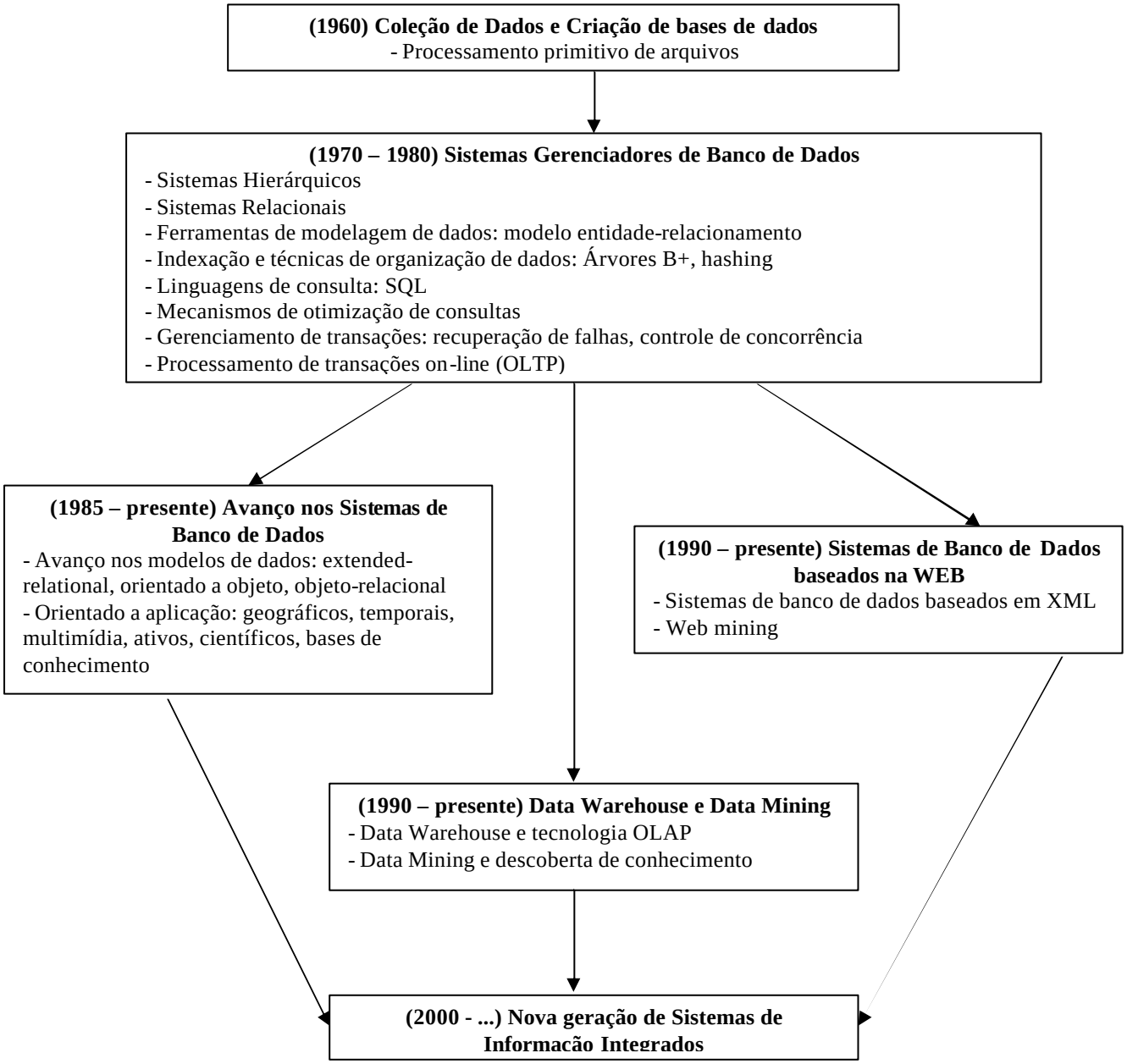

Figura 2-1 - A evolução da tecnologia de banco de dados. Adaptado de (KAMBER, 2001)

Segundo ELMASRI (2002), em 1970 Ted Codd da IBM apresentou o modelo relacional de banco de dados, que chamou a atenção devido à sua simplicidade e sua fundamentação matemática. Esse modelo utiliza o conceito de uma relação matemática com sua base teórica na teoria de conjuntos e na lógica de predicados de primeira ordem e está sendo utilizado amplamente em sistemas comerciais ao longo dos últimos 20 anos.

Os dados podem ser armazenados em diferentes tipos de bancos de dados, como por exemplo, o modelo relacional, o orientado a objetos, o temporal, o geográfico, o multimídia,

entre outros. A escolha está condicionada ao tipo de aplicação. Não está no escopo desse trabalho apresentar maiores detalhes sobre os diferentes tipos de banco de dados e suas aplicações. Essas informações podem ser encontradas em ELMASRI (2002).

Recentemente, desenvolveu-se uma nova arquitetura de banco de dados chamada de *data warehouse*. *Data warehouse* é um repositório de dados que organiza num esquema unificado e em apenas um local as informações provenientes de diversas fontes de dados com o objetivo de facilitar o gerenciamento da decisão. A tecnologia de data warehouse inclui a limpeza e integração dos dados e o processo analítico on-line (OLAP). OLAP é uma técnica de análise com funcionalidades como a sumarização, consolidação e agregação, assim como a habilidade de observar a informação de diferentes ângulos. Embora as ferramentas OLAP suportem análise multidimensional e tomada de decisão, ainda podem ser requeridas análises adicionais de dados, como por exemplo, classificação, associação ou agrupamentos. Essas análises adicionais podem ser encontradas nas ferramentas de *data* mining (KAMBER, 2001).

Segundo KAMBER (2001), as ferramentas de *data mining* realizam análises nos dados e podem descobrir importantes informações em forma de padrões, contribuindo para a estratégia do negócio, bases de conhecimento, pesquisas médicas e científicas, administração governamental e controle ambiental. Maiores detalhes sobre data mining são apresentados no tópico sobre o processo de descoberta de conhecimento (KDD).

2.2 Processo de Descoberta de Conhecimento

Ao longo das últimas três décadas, muitas organizações produziram grande quantidade de dados, sendo armazenados e processados utilizando a tecnologia de banco de dados através da linguagem SQL, que é a principal linguagem para acesso e manipulação dos dados na maioria dos bancos de dados transacionais (KAMBER, 2001). Como o acesso à informação é um fator de sucesso competitivo para grande parte das organizações, cada vez mais gerentes e executivos necessitam obter com rapidez e facilidade informações sobre seu negócio para auxiliarem em suas decisões. É impraticável a obtenção desta informação pelos profissionais do nível tático e estratégico da organização através da utilização da linguagem SQL, pois ela é altamente estruturada e pré-supõe que seu usuário esteja ciente da estrutura do banco de dados[1]. A linguagem SQL é mais utilizada por profissionais predominantemente do nível operacional da empresa (ELMASRI, 2002).

[1] *Estrutura de banco de dados* é definida como os tipos de dados, relacionamentos e restrições que devem existir entre os dados (Elmasri, 2002).

Baseado nesse contexto, foi concebida a arquitetura de dados do data warehouse (DW), viabilizada diante do crescente aumento de processamento e armazenamento dos computadores atuais e da sofisticação das técnicas e ferramentas analíticas (ELMASRI, 2002). O Data warehouse organiza num esquema unificado e em apenas um local as informações provenientes de diversas fontes de dados com o objetivo de facilitar o gerenciamento da decisão utilizando ferramentas de análise OLAP. Apesar de o data warehouse facilitar o gerenciamento da decisão, há algumas questões que necessitam de outros tipos de análise (KAMBER, 2001). Neste caso, pode-se utilizar o *data mining*.

Para KAMBER (2001), data *mining* é um campo novo e interdisciplinar desenhado a partir de áreas como os sistemas de banco de dados, data warehousing, estatística, inteligência artificial, visualização de dados, computação de alta performance, entre outros. Data mining é uma das etapas no processo de descoberta de conhecimento (KDD), como apresentado na figura 2.2.

O processo de descoberta de conhecimento é um processo composto por diversas etapas, envolvendo metodologias e técnicas de data mining. O seu objetivo é o de "otimizar e automatizar o processo de descrição das tendências e dos padrões contidos nesse processo, potencialmente úteis e interpretáveis" (OGLIARI, 2004). Para PRASS (2004), o processo de descoberta de conhecimento compreende todo o ciclo que os dados percorrem até virar informação.

Nessa seção, são apresentados os conceitos do processo de descoberta de conhecimento com suas etapas, metodologias, tarefas e técnicas.

Figura 2-2 - As etapas do processo de descoberta de conhecimento. Fonte: Adaptação de OGLIARI (2004), PRASS (2004) e KAMBER (2001).

2.2.1 Seleção dos dados

Nessa primeira fase do processo de descoberta de conhecimento é selecionado um conjunto de dados contendo as variáveis que serão utilizadas em análises nas fases

posteriores. Essas variáveis são selecionadas de acordo com o objetivo em questão e deve ser realizada com auxílio de um especialista no assunto, pois é difícil fazer uma seleção adequada sem ter um bom domínio do problema em estudo e, muitas vezes, é dessa escolha de variáveis que depende o resultado bem sucedido de todo o KDD (OGLIARI, 2004).

2.2.2 Pré-Processamento

A etapa de pré-processamento é importante para data warehouse e para data mining, pois nessa etapa são identificados e corrigidos problemas presentes nos dados selecionados. Esses problemas podem ser dados inconsistentes, dados faltantes (*missing*) ou valores discrepantes (*outliers*) (KAMBER, 2001).

Para obter bons resultados no KDD, além da correta seleção das variáveis a serem utilizadas nas análises, deve-se ter os dados limpos e corretos, pois as ferramentas de mineração de dados são altamente sensíveis a ruídos nos dados. A identificação desses problemas pode ser obtida através da análise exploratória dos dados.

A análise exploratória dos dados "emprega técnicas estatísticas descritivas e gráficas para estudar um conjunto de dados, detectando *outliers* e anomalias, e testando as suposições do modelo" (OGLIARI, 2004).

a. Valores faltantes (*missing*)

Os valores faltantes (*missing values*) ocorrem quando um determinado registro (tupla) possui uma ou mais variáveis sem valor. Para (KAMBER, 2001) os seguintes métodos podem ser utilizados para contornar esse problema:

- *Excluir a tupla*: A exclusão de todo um registro implica em perda de informação e é mais aconselhável quando a quantidade de variáveis com valores ausentes for grande ou quando essa variável ausente for importante para utilização com técnicas de classificação ou descrição.
- *Preencher manualmente os valores faltantes:* Esse método torna-se impraticável caso a quantidade de valores faltantes for alta ou caso a base de dados seja muito grande, pois envolve muito tempo e especialistas no problema.
- *Usar uma constante global para valores faltantes*: utilizando um valor como "desconhecido" para preencher os valores faltantes. Não é recomendando, pois

as técnicas de data mining podem errar ao considerar esse valor como uma relação interessante entre os casos analisados, por exemplo, por terem em comum esse valor.

- *Usar a média aritmética da variável:* Nesse método utiliza-se da média aritmética da variável em questão para preencher o valor faltante.
- *Usar o valor mais provável:* Através de modelos de predição (como análise de regressão, árvores de decisão) ou técnicas de imputação pode-se fazer uso das demais variáveis do registro para predizer o valor ausente. Com esse método a chance de obter um valor correto ou mais próximo do real é superior aos dos modelos sugeridos acima.

b. Valores Discrepantes (*outliers*)

Os valores extremos, atípicos ou distintos dos demais encontrados no conjunto de dados são chamados de valores discrepantes ou outliers.

Os valores discrepantes devem ser bem analisados, pois podem estar indicando uma tendência ou comportamento incomum ou transações fraudulentas. Caso seja necessária alguma ação pode-se optar por excluir esse registro ou substituir seus valores por valores encontrados com o uso de regressão de dados.

2.2.3 Transformação e Integração

Nesse ponto os dados já estão devidamente selecionados, limpos e pré-processados. Porém, para aplicar os algoritmos de data mining, os dados devem estar integrados e transformados, ou seja, estarem em um formato padrão para utilização.

A integração de dados consiste em reunir num único arquivo (semelhante a uma planilha eletrônica) ou tabela de banco de dados todos os dados de diferentes fontes que foram selecionados para utilização nas análises. Na integração deve-se ter cuidado para não incluir variáveis duplicadas (que podem ter diferentes nomes em cada fonte de dados selecionada) e variáveis redundantes (que podem ser verificadas através de análise de correlação). Tomando esses cuidados se contribui para aumentar a precisão e a velocidade nos processos de mineração de dados.

A transformação dos dados, segundo OGLIARI (2004), tem o objetivo de “converter um conjunto bruto de dados em uma forma padrão de uso”. Com esse objetivo, as seguintes técnicas podem ser empregadas:

- *Discretização:* envolve a conversão de variáveis contínuas em categorizadas e depois em discretas ou a conversão de variáveis categorizadas em discretas (OGLIARI, 2004).
- *Agregação:* é aplicação de operações nos dados para obter dados resumidos. Por exemplo, obter vendas mensais ou semanais a partir dos dados de vendas diárias (KAMBER, 2001).
- *Normalização:* consiste em padronizar os valores de uma variável para uma faixa padrão de valores, como por exemplo, [-1, 1] ou [0, 1]. A normalização é particularmente útil para os algoritmos de classificação envolvendo redes neurais (melhorando o desempenho no aprendizado) e *clustering* (nos cálculos das distâncias). Um tipo de normalização é a normalização Min-Max que mapeia os valores atuais de uma variável A em valores V’ na faixa informada em new_max e new_min, conforme a fórmula abaixo (KAMBER, 2001):

$$v' = \frac{v - \min_A}{\max_A - \min_A}\left(new\_\max_A - new\_\min_A\right) + new\_\min_A \quad \text{(Equação 1)}$$

  Como nesse trabalho a única transformação necessária é para a faixa [0, 1], a fórmula acima pode ser simplificada para:

$$v' = \frac{v - \min_A}{\max_A - \min_A} \quad \text{(Equação 2)}$$

- *Alisamento:* é uma técnica para remover valores discrepantes. Por exemplo, análise de regressão.
- *Criação de novas variáveis a partir de outro conjunto de variáveis:* Por exemplo, a partir de dados com informações de largura e comprimento criar uma nova variável calculada chamada área.

2.2.4 Data Mining

Data mining é uma atividade que emprega técnicas estatísticas e de inteligência artificial para identificar informações relevantes em grandes quantidades de dados, apresentando-as em formas de padrões que sejam úteis para as corporações montarem suas

estratégias com objetivo de melhorarem seus negócios quanto a vendas, marketing, suporte, entre outros (PRASS, 2004).

2.2.4.1 Metodologias

Segundo OGLIARI (2004), a mineração de dados pode seguir as seguintes metodologias conforme o conhecimento que se tenha do assunto:

- A metodologia de *teste de hipóteses* é empregada quando se tem muito conhecimento do assunto (campo de atuação da empresa ou o assunto em análise) ou alguma idéia de qual relação está se buscando. Dessa forma é montada uma ou mais hipóteses e busca-se testá-las utilizando alguma técnica de *data mining.*
- A metodologia de *descoberta supervisionada de conhecimento* é empregada quando se tem algum conhecimento do assunto em questão e geralmente busca-se justificar alguma variável (chamada de variável alvo ou resultado) em termos de outras variáveis (entradas), ou seja, encontrar um modelo que corretamente associe as variáveis de entrada com a variável alvo. Como exemplos de técnicas temos redes neurais, regressão linear múltipla e regressão logística.
- A *descoberta não supervisionada de conhecimento* é empregada quando o grau de conhecimento do assunto é muito baixo ou nulo. Não é definida a variável alvo. Como exemplo, os modelos de variáveis latentes (como a análise de componentes principais) que buscam a redução da dimensionalidade, ou seja, apenas as variáveis mais não significativas são mantidas. Outro exemplo é a análise de agrupamentos que tem o objetivo de formar grupos de registros semelhantes com base em diversas variáveis.

2.2.4.2 Tarefas desempenhadas

As tarefas de data mining estão relacionadas com o tipo de conhecimento que se deseja extrair dos dados. A seguir estão apresentadas algumas dessas tarefas e quais as técnicas (ou ferramentas) de data mining que empregam.

- *Associação:* também é referenciada como *market basket association* (associação de carrinho de supermercado). Seu objetivo é identificar quais fatos ou objetos tendem a ocorrerem juntos para um mesmo evento no conjunto

de dados, por exemplo, informa o quanto a presença de um determinado produto numa compra implica a presença de outro. A técnica de data mining responsável por essa tarefa é a análise de associação (OGLIARI, 2004).

- *Classificação:* Associa ou classifica um registro em classes predefinidas, ou seja, determina com que grupo de entidades já classificadas determinado dado apresenta mais semelhança. (AEDB, 2005). Podem ser utilizadas as seguintes técnicas de data mining: redes neurais artificiais, árvores de decisão, estatística (análise discriminante e regressão logística).

- *Estimação:* "Estimar uma grandeza é avaliá-la tendo como base casos semelhantes onde esse valor está presente" (AEDB, 2005). As técnicas de redes neurais, algoritmos genéticos e estatística (intervalos de confiança, intervalos de predição (regressão)) são utilizadas para essa tarefa.

- *Previsão e predição:* consiste em determinar o valor que uma variável irá assumir no futuro tendo como base seus valores anteriores, ou seja, os valores que essa variável assumiu em períodos de tempo passados são utilizados para predizer seu valor futuro. Utilizam-se as técnicas de redes neurais, árvores de decisão e estatística (regressão múltipla, regressão logística binária) (OGLIARI, 2004).

- *Agrupamento ou segmentação*: consiste em agrupar os registros mais semelhantes entre si de acordo com algumas variáveis. As técnicas utilizadas são redes neurais artificiais e estatística (análise de conglomerados).

2.2.4.3 Técnicas

Nessa sessão são apresentadas sucintamente as técnicas (ou ferramentas) de data mining. No tópico 2.3 é apresentado em maiores detalhes a técnica de análise de agrupamento (*cluster analysis*), que é a técnica utilizada nesse trabalho de conclusão de curso.

- *Regras de Indução ou Análise de Associação*: PRASS (2004) e CARVALHO (2003) dizem que essa técnica é altamente automatizada e, possivelmente é a melhor técnica de mineração de dados para expor todas as possibilidades de padrões existentes no banco de dados. As regras são apresentadas de uma forma bastante simples: **se** <condição> **então** <conseqüência>. Alguns exemplos:

- **se** comprou cereal **então** comprou também leite.
- **se** comprou presunto e queijo **então** comprou também pão.

Essas regras também vêm acompanhadas com sua precisão (indica com que freqüência a regra está correta) e a cobertura (com que freqüência essa regra pode ser usada).

Padrão gerado: conjunto de regras na forma "**se** <condição> **então** <conseqüência>".

- *Árvores de Decisão:* é um modelo preditivo onde o usuário deve selecionar uma das variáveis para ser a variável alvo e o algoritmo de árvore de decisão analisa o conjunto das demais variáveis disponíveis em função da variável alvo, criando e organizando regras de classificação e decisão no formato de diagramas de árvores. Nesse diagrama os ramos indicam uma questão de classificação e as folhas indicam os conjuntos de dados que satisfazem a essa classificação (PRASS (2004); CARVALHO (2003); BARBIERI (2001)).

  Um exemplo de aplicação de regras de decisão é em malas diretas, onde geralmente há um custo de envio bastante elevado se comparado comas taxas de retorno. BARBIERI (2001) informa que a taxa de retorno das malas diretas é na ordem de 20% e que as árvores de decisão são úteis por permitirem ao setor de marketing direcionar suas campanhas ao grupo de clientes que possivelmente terá mais retorno. A figura 2.3 ilustra uma árvore de decisão.

  Padrão gerado: estrutura em forma de árvore de decisão como apresentado na figura 2.3.

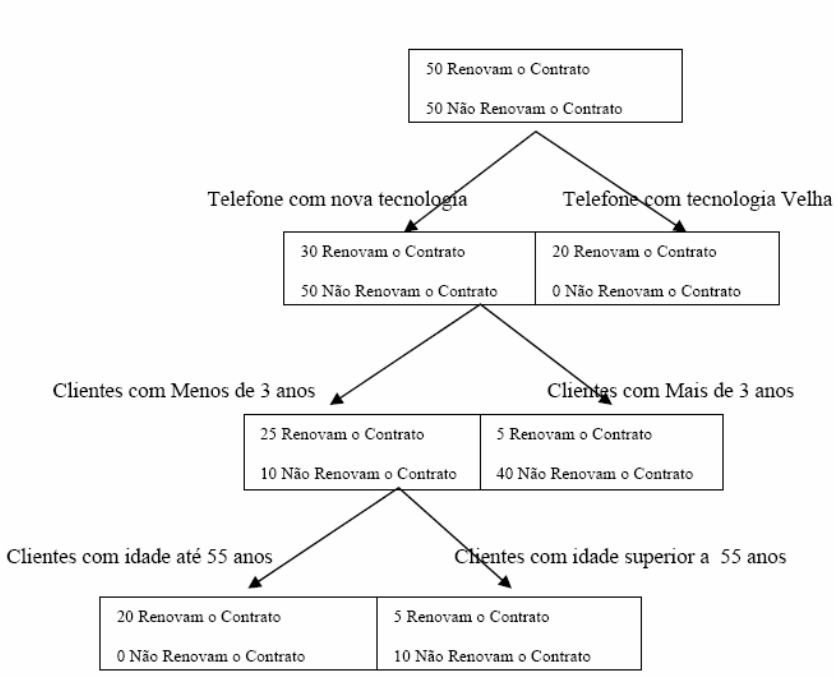

Figura 2-3 – Exemplo de resposta da técnica de árvore de decisão. Fonte: (CARVALHO, 2003).

- *Redes Neurais Artificiais:* são técnicas que procuram reproduzir de maneira simplificada as conexões do sistema biológico neural (COMRIE in PRASS (2004)). A estrutura é composta de diversos nós (chamados neurônios) interconectados e geralmente organizados em três ou mais camadas. A primeira camada é a camada de entrada e recebe o *input* das variáveis do banco de dados, que devem ser valores numéricos entre 0 e 1, pois as redes neurais trabalham melhor com esses valores. Isso implica em transformação dos dados para ficarem nessa faixa de valores. Entre as duas camadas há a camada intermediária, que pode ser composta por diversas camadas.

  Nessa técnica, a base de dados deve ser dividida em três partes, o arquivo de treinamento, o arquivo de validação e o arquivo de testes. O arquivo de treinamento é utilizado para definir os melhores valores das conexões entre os neurônios. O arquivo de validação é utilizado para encontrar o valor da taxa de erros da rede neural, e é uma boa estimativa para a margem de erro ao se utilizar a base de testes ou novos registros.

  A técnica de redes neurais é utilizada para realizar predição, classificação, agrupamentos e previsões (séries temporais), sendo mais comumente utilizada para classificações e predições.

Padrão gerado: um modelo de redes neurais que irá predizer ou classificar um determinado conjunto de dados de entrada.

- *Análise de Conglomerados (cluster analysis):* uma análise de agrupamento tem o objetivo de agrupar os dados em subconjuntos de tal forma "que o grau de associação entre os casos de um determinado grupo é forte e é fraca entre casos de diferentes grupos" (OGLIARI, 2004). Os detalhes dessa técnica são apresentados no tópico 2.3.

  Padrão gerado: retorna o conjunto de dados original acrescido de um novo atributo, que é a identificação de um grupo ao qual pertence o corrente registro. Dependendo do software utilizado pode retornar a importância que cada variável teve na formação dos grupos, etc..

2.2.5 Análise e Interpretação

Nessa fase do processo de descoberta de conhecimento, em conjunto com um especialista no assunto, devem-se interpretar os resultados e analisá-los quanto a sua relevância e qualidade. Segundo OGLIARI (2004), os padrões podem ser avaliados quanto ao serem:

- *Úteis*: o quanto ajuda a responder os objetivos traçados para a realização da mineração? Por exemplo, pode-se descartar um padrão encontrado por de representar informações já conhecidas ou que não seja pertinente aos objetivos.
- *Interpretáveis:* devem ser de fácil compreensão pelos analistas. Por exemplo, pode-se descartar um conjunto de regras por ser muito grande para ser devidamente avaliado ou por conter muita informação redundante.
- *Válidos:* deve-se avaliar o quanto um resultado é confiável e correto através das medidas estatísticas fornecidas pela maioria dos algoritmos de data mining sobre a validade do padrão apresentado (CARVALHO, *2003).*
- *Novos:* avaliar no sentido de ser algo interessante ou incomum.

Dependendo da avaliação do modelo pode ser necessário voltar alguns passos no processo de descoberta de conhecimento ou mesmo reiniciar todo o processo visando remover redundâncias ou variáveis sem significado para análise.

2.3 Análise de Conglomerados

Nessa seção são apresentados os requisitos necessários para realizar a análise de agrupamentos. São apresentadas as medidas de distância empregadas na avaliação das similaridades entre objetos (cada registro da tabela) e como devem ser aplicadas para cada tipo de variável envolvida na análise. Para finalizar o capítulo são apresentados os métodos de formação de agrupamentos.

Segundo KAMBER (2001), *clustering* (cluster analysis ou análise de conglomerados) é a técnica de reunir objetos em grupos, de tal forma que os objetos que estão no mesmo grupo são mais semelhantes entre si do que os objetos que estão em outro grupo definido. Ou seja, "é utilizado para combinar registros em grupos de forma que cada grupo formado seja mais homogêneo para um determinado conjunto de variáveis" (OGLIARI, 2004).

2.3.1 Medidas de distância

A avaliação das semelhanças entre objetos é realizada através de medidas de distâncias aplicadas às variáveis da análise. Cada tipo de variável possui uma ou mais medidas de similaridade a ser aplicada.

Vale salientar que o usuário de uma ferramenta de data mining não precisa calcular a medida de distância entre os objetos, mas precisa transformar os dados para que a aplicação possa realizar corretamente a mensuração de distância.

A seguir são apresentados os diferentes tipos de variáveis comumente encontrados em *data mining* e como processá-las para sua utilização em *cluster analysis*. Inicialmente apresentam-se técnicas a serem aplicadas em análises que envolvam apenas um tipo de variável, e por último mostra-se como proceder em análises que envolvam diferentes tipos de variáveis.

2.3.1.1 Variáveis intervalares

Segundo KAMBER (2001), variáveis intervalares são mensuradas numa escala linear. Como exemplos de variáveis intervalares podemos citar as variáveis que representam peso, altura, temperatura. As diferentes unidades de mensuração representadas por esse tipo de variável podem afetar a análise de agrupamentos, pois os objetos podem ter um valor de similaridade, em termos numéricos, maior ou menor.

Segundo KAMBER (2001), para evitar esse efeito na medida de distancia os dados precisam antes ser padronizados. A padronização tende a dar a todas as variáveis o mesmo peso e pode ser realizada através do *z-score,* dado por:

$$z = \frac{x_i - \bar{x}}{s} \qquad \text{(Equação 3)}$$

onde x_i é o i-ésimo elemento de um conjunto com *n* elementos, $\bar{x}$ é a média e *s* é o desvio absoluto dado.

Ainda segundo KAMBER (2001), o uso do desvio absoluto em vez do desvio padrão é justificado pelo fato de o desvio absoluto minimizar os efeitos dos outliers. Após a padronização (ou não, dependendo da aplicação) das variáveis intervalares pode-se mensurar a distância utilizando o cálculo da *distância euclidiana*. A distancia euclidiana é a mais comumente utilizada, sendo definida por:

$$d(i, j) = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{il})^2} \qquad \text{(Equação 4)}$$

onde *i* e *j* representam dois objetos e *p* representa o número de variáveis.

KAMBER (2001) apresenta outras medidas de distancia, como a *Manhattan*, porém não será apresentada por não ser comumente utilizada.

2.3.1.2 Variáveis binárias

As variáveis binárias assumem apenas dois valores possíveis (0 e 1) onde zero indica a ausência de determinada característica e um indica que a característica está presente. Por exemplo, considerando a variável *fumante, zero* indicaria que a pessoa não fuma e *um* que a pessoa fuma. (KAMBER, 2001). Além de indicar presença ou não de uma característica ela pode ser utilizada para mapear uma variável nominal com dois estado para zero e um, como por exemplo, sexo masculino como sendo 0 e feminino como 1.

As variáveis binárias podem ser simétricas ou assimétricas. As variáveis simétricas possuem os dois estados equivalentes (tem o mesmo peso), ou seja, não há diferença entre a característica estar presente ou não para a análise em estudo (por exemplo, masculino X feminino). Para as variáveis assimétricas há diferença de pesos entre os dois estados em uma análise. Como exemplo desse tipo de variável cita-se a presença ou não de determinada

doença. Nesse caso mapeia-se como sendo 1 o valor que possui maior importância no contexto em análise. (KAMBER, 2001).

Segundo KAMBER (2001), a medida de semelhança para esse tipo de variável pode ser calculada utilizando a tabela de contingência para variáveis binárias.

Tabela 2-1 - Tabela de contingência para variáveis binárias. Fonte: KAMBER (2001).

		Objeto j		
		1	0	Soma
Objeto i	1	q	r	q + r
	0	s	t	s + t
	Soma	q + s	r + t	p = q+r+s+t

A leitura dessa tabela se dá da seguinte forma, *r* representa os casos onde o objeto *i* possui a característica e o objeto *j* não possui. *S* representa os casos onde o objeto *j* possui a característica e o objeto *i* não a possui, e assim por diante nos demais casos.

Para as variáveis binárias simétricas a medida de distância é calculada através de

$$d(i,\ j) = \frac{r+s}{p} \qquad \text{(Equação 5)}$$

e, caso seja assimétrica utiliza-se o coeficiente de *Jaccard*, que ignora os casos onde a variável apresenta os dois estados como 0 (não presença da característica):

$$d(i,\ j) = \frac{r+s}{q+r+s} \qquad \text{(Equação 6)}$$

2.3.1.3 Variáveis nominais

Variáveis nominais são uma generalização de variáveis binárias, se diferenciando pela possibilidade de representarem mais que dois estados. Por exemplo, uma variável nominal com cinco estados representando cinco cores (vermelho, verde, azul, amarelo, violeta). Os estados dessas variáveis podem ser representados através de números inteiros, letras ou símbolos. Assim os nomes das cores poderiam ser substituídos, arbitrariamente, por 1, 2, 3, 4 e 5 (KAMBER, 2001).

A medida de distância pode ser calculada pela fórmula abaixo, onde *m* representa o número de variáveis em que *i* e *j* possuem o mesmo valor e *p* é a quantidade de variáveis.

$$d(i,\ j) = \frac{p-m}{p} \qquad \text{(Equação 7)}$$

Outra forma de codificar os valores presentes em variáveis nominais seria transformá-las em variáveis binárias. Para cada valor diferente de uma variável nominal seria criado a mesma quantidade de variáveis binárias e atribui-se o valor 1 quando a variável indicar o valor original e zero para as demais. A medida de distância seria calculada da mesma forma apresentada anteriormente para variáveis binárias.

2.3.1.4 Variáveis ordinais

As variáveis ordinais são semelhantes as nominais diferindo pelo fato que nas ordinais seus valores apresentarem uma seqüência de importância. Esses tipos de variáveis podem representar conceitos acadêmicos, níveis de instrução, etc.

Antes de mensurar a distância entre os objetos, os valores das variáveis ordinais precisam ser mapeadas para *ranks.* Dessa forma, uma variável ordinal com M estados seria mapeada para os valores 1, 2...M. Como cada variável ordinal pode ter diferentes números de estados, torna-se necessário transformá-los para a faixa de valores [0, 1], garantindo assim que os valores tenham pesos iguais antes de se avaliar as distâncias, e é realizado através da fórmula abaixo.

$$z_{if} = \frac{r_{if} - 1}{M_f - 1} \qquad \text{(Equação 8)}$$

Após essas transformações pode-se utilizar as medidas de distâncias apresentadas para a variável do tipo intervalar com os valores z_{if} encontrados acima.

2.3.1.5 Variáveis de tipos diferentes

Nos tópicos anteriores foram discutidas como calcular semelhanças entre objetos formados por apenas um tipo de variável. Porém, é bastante comum realizar análises que envolvam mais de um tipo de variável.

Para realizar as medidas de distâncias deve-se inicialmente transformar cada tipo de variável como descrito nos tópicos anteriores. Estando cada tipo de variáveis devidamente processadas deve-se transformar todos os valores para que fiquem na faixa de valores [0, 1].

A transformação para a faixa [0, 1] depende de cada tipo de variável. As variáveis binárias não necessitam de transformação, pois já estão na faixa correta de valores. Para as variáveis nominais e ordinais pode-se utilizar as possibilidades apresentadas anteriormente. Já as variáveis intervalares podem ser transformadas para a faixa [0, 1] através da equação 2 apresentada na seção 2.2.3.

Com essas etapas intermediárias concluídas a medida de distancia entre os objetos pode ser realizada através do cálculo da distancia euclidiana apresentada anteriormente.

2.3.2 Métodos de formação do agrupamento

Os algoritmos de análise de conglomerados podem ser classificados quanto ao método de formação em hierárquicos, de partição, baseados em modelo e baseados em densidade. Em PRASS (2004) pode ser encontrada uma comparação entre algoritmos que implementam cada método de formação de agrupamento citados acima.

2.3.2.1 Métodos Hierárquicos

Os métodos hierárquicos criam uma decomposição hierárquica dos dados e podem ser classificados em divisivos e aglomerativos dependendo da maneira como a decomposição é realizada (KAMBER, 2001).

- *Métodos Hierárquicos Aglomerativos*: esta estratégia é também conhecida como *bottom-up*. Inicia considerando cada objeto como sendo um *cluster* atômico e une os clusters mais semelhantes entre si formando grupos cada vez maiores. O algoritmo de agrupamento vai encerrar quando todos os objetos pertencerem a apenas um *cluster* ou até que uma condição de término seja atingida (determinação da quantidade de *clusters* a serem formados). A maioria dos métodos de agrupamento hierárquicos são métodos hierárquicos aglomerativos. Na obra de KAMBER (2001) está detalhado o algoritmo *AGNES*, que implementa esse método.
- *Métodos Hierárquicos Divisivos*: a estratégia adotada nesses métodos é o oposto dos métodos aglomerativos e é também conhecida como *top-down*. Inicia com todos os objetos fazendo parte de apenas um *cluster* que é sucessivamente dividido até que cada objeto forme um grupo ou até que atinja uma quantidade de clusters pré-determinada. Pode ser encontrado na obra de Kamber (2001) a descrição do algoritmo *DIANA*, que implementa esse método.

Os métodos hierárquicos possuem a desvantagem de que uma vez um passo sendo efetuado (junção de grupos no aglomerativo ou divisão de grupos no divisivo) eles não podem ser desfeitos, ou seja, erros de decisão não podem ser corrigidos. Mas há a vantagem de se conhecer o histórico da formação dos agrupamentos, permitindo saber de onde determinado objeto se originou.

Esses métodos possuem o inconveniente de serem impraticáveis em grandes bases de dados devido ao alto custo computacional (MICHAUD (1997) in PRASS (2004)).

2.3.2.2 Métodos de Partição

Os algoritmos de partição buscam a formação dos grupos pela divisão dos objetos da base de dados sem a necessidade de associações hierárquicas. Dado *k*, o número de partições (*clusters*) a serem construídas, o método de partição cria um particionamento inicial. Ele então usa técnicas de realocação iterativas que tenta melhorar o particionamento movendo os objetos de um grupo para outro. O critério geral para um bom particionamento é que os objetos no mesmo cluster são mais parecidos entre si e mais diferentes ou distantes dos objetos presentes em outros grupos. (KAMBER, 2001).

Os dois principais algoritmos que implementam o método de partição são o *k-means* e o *k-medoid*. Entretanto o segundo método é mais robusto que o *k-means* na presença de valores *outliers* porque trabalha com o objeto mais centralmente localizado (mais próximo a mediana do grupo) em vez do objeto mais próximo do valor médio do *cluster*. Em ambos os métodos é necessário que o usuário informe a quantidade de partições (*clusters*) a serem formados, ou seja, é necessário informar o valor *k*. Segundo KAMBER (2001) pode-se optar por utilizar os métodos hierárquicos para a determinação da quantidade de grupos, e utilizar esse valor como o valor *K* nos métodos de partição. KAMBER (2001) apresenta o algoritmo *CLARANS* que se adapta melhor para grandes bases de dados.

2.3.2.3 Métodos de densidade

"Nos métodos baseados em densidade, um agrupamento é uma região que tem uma densidade maior de objetos do que outra região vizinha" (PRASS, 2004). A idéia geral desse método é que para um dado ponto (objeto) dos dados exista, em uma dada área, uma quantidade mínima de outros objetos (KAMBER, 2001). Ou seja, o número de objetos que circulam um determinado ponto deve exceder algum limite para que haja a formação do agrupamento.

"O método inicia usa execução por um objeto arbitrário e, se sua vizinhança satisfaz o mínimo de densidade, inclui o objeto e os que estão em sua vizinhança no mesmo agrupamento. O processo é então repetido para novos pontos adicionados" (PRASS, 2004).

O algoritmo *DBSCAN* apresentado em PRASS (2004) e KAMBER (2001) é um dos algoritmos que implementa o método de densidade.

2.4 Sistemas de Informações Geográficas (GIS)

Esse capítulo tem o objetivo de introduzir sucintamente os sistemas de informações geográficas, suas definições disponíveis na literatura, descrever sua arquitetura interna, a estrutura das bases de dados utilizadas por eles e os tipos de dados presentes em análise espacial de dados.

A definições encontradas na literatura sobre GIS:

> *Um conjunto manual ou computacional de procedimentos utilizados para armazenar e manipular dados geo-referenciados (ARONOFF in CÂMARA et al. (2000).*

> *Sistemas que realizam o tratamento computacional de dados geográficos e manipulam a geometria e os atributos dos dados que estão geo-referenciados, ou seja, localizados na superfície terrestre e representados numa projeção cartográfica (CÂMARA et al., 2000).*

> *Um sistema de suporte à decisão que integra dados referenciados espacialmente num ambiente de respostas a problemas (COWEN in CÂMARA et. al (2000).*

Cada uma dessas definições reflete a diversidade de usos e visões para essa tecnologia e, a partir desses conceitos, CÂMARA et. al. (2000) apresenta as principais características de um sistema de informações geográficas, conforme abaixo:

- Devem possibilitar a integração e inserção, em uma base de dados única, de informações provenientes de censos populacionais, cadastros urbanos e rurais, imagens de satélite, redes e informações espaciais provenientes de dados cartográficos[2].
- Fornecer mecanismos para combinar as informações citadas acima por meio de algoritmos de manipulação e análise e, inclusive possibilitar consultas, visualizações e plotagem do conteúdo da base de dados geo-referenciadas.

[2] Segundo dicionário eletrônico Aurélio, cartografia pode ser definido como:
1. Arte ou ciência de compor cartas geográficas.
2. Tratado sobre mapas.

2.4.1 A estrutura de um GIS

De uma forma simplificada um sistema de informações geográficas pode ser composto pelas seguintes camadas, como apresentado na figura 2.4.

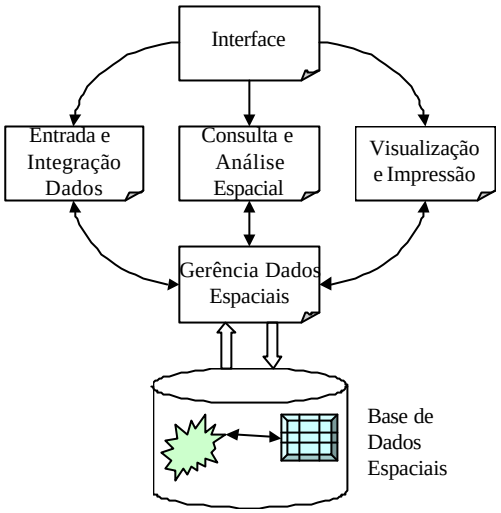

Figura 2-4 – Estrutura geral de Sistemas de informações geográficas. Fonte: CÂMARA et al. (2000).

Conforme vemos na figura 2.4, na camada mais próxima ao usuário temos a interface que define como o sistema é operado e controlado. Na camada intermediária temos três componentes responsáveis pelos mecanismos de processamento de dados espaciais, como a entrada, a edição, a análise, a visualização e a saída. No nível mais baixo de um GIS está presente o componente de gerência de banco de dados geográfico que controla o armazenamento e recuperação dos dados (CÂMARA et al., 2000).

2.4.2 Gerência de dados no GIS

O componente de gerência de dados espaciais presente na figura 2.4 interage diretamente com a base de dados espaciais. Dependendo da forma como a base de dados está estruturada o componente de gerência de dados do GIS pode seguir a arquitetura *dual* ou a arquitetura integrada (CÂMARA et. al. 2000).

2.4.2.1 Arquitetura *dual*

Um sistema de informações geográficas que implementa a arquitetura *dual* possui sua base de dados geo-referenciada composta de pelo menos dois arquivos. Um dos arquivos possui armazenado os dados (atributos ou variáveis) de cada unidade geográfica em forma de uma tabela. Esse arquivo é chamado de componente alfa-numérica. O outro arquivo armazena

a representação geográfica dos dados. Para cada identificador de unidade geográfica deve haver na componente alfa-numérica uma entrada correspondente.

Essa arquitetura apresenta a vantagem de se utilizar os SGBDs relacionais disponíveis no mercado para a manipulação do arquivo alfa-numérico, porém, o arquivo de representação geográfica não é gerenciado pelo SGBD. Por essa razão essa arquitetura apresenta as seguintes desvantagens, segundo CÂMARA et al. (2000):

- Dificuldade de manter a integridade entre a componente alfa-numérica e a representação geográfica.
- Consultas mais lentas, pois os arquivos são processados separadamente pelo GIS, sendo necessário uma consulta a componente alfa-numérica através do SGBD seguida pelo processamento do arquivo de representação geográfica (em geral em formato proprietário).
- Cada sistema de informações geográficas pode utilizar seu formato próprio de representação geográfica, dificultando assim a interoperabilidade entre os dados.

A base de dados fornecida pelo IBGE e utilizada nesse trabalho implementa a arquitetura *dual* seguindo o formato de arquivos da ESRI (2005), chamado de *shape*. Esse tipo de formato é composto de três arquivos: um com extensão *SHP*, onde está armazenada a geometria do mapeamento; outro com extensão *DBF*, que corresponde aos dados associados a cada elemento da geometria; e finalmente um de extensão *SHX*, que possibilita a conexão entre as entidades da geometria e os seus respectivos atributos. Esquematicamente na figura 2.5 está a representação desse tipo de formato. Podemos analisar essa figura sendo a área da esquerda com o ID "Aa" e a da direita "Bb", logo estas têm, respectivamente, associadas os valores 2 e 3 como atributos (PACHECO, 2005).

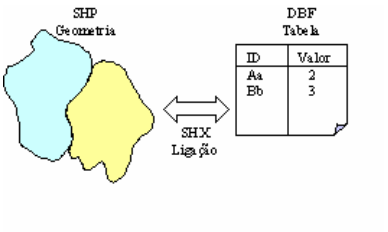

Figura 2-5 – Estrutura do formato shape da ESRI (2005). Fonte: PACHECO (2005).

Geralmente os arquivos de mapeamento disponíveis possuem poucos dados, sendo necessária a inserção de outros dados nesses arquivos para efetuar-se as análises desejadas.

Para atender aos objetivos desse trabalho foi inserida a informação do *cluster* que cada setor censitário está associado. O procedimento necessário para realizar a inserção está exposto na sessão 3.6.

2.4.2.2 Arquitetura integrada

Nesse tipo de arquitetura tanto a componente alfa-numérica quanto a espacial (representação geográfica) é armazenada em um SGBD. A principal vantagem dessa arquitetura é que sob responsabilidade do SGBD o controle e a manipulação dos dados, a gerência de transações, o controle de integridade e a concorrência.

A figura 2.6 ilustra as duas arquiteturas para a gerência de dados nos sistemas de informações geográficas.

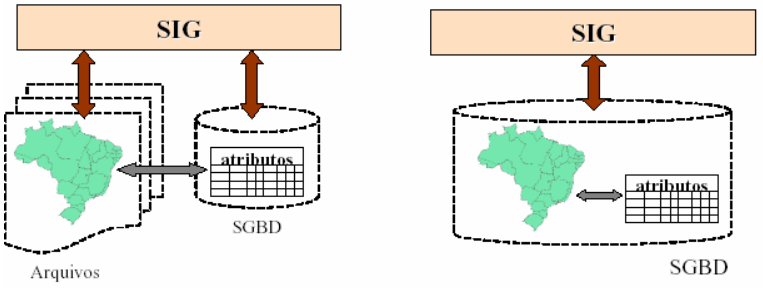

Figura 2-6 – A arquitetura *dual* (esquerda) e a arquitetura integrada (direita). Fonte: CÂMARA et al. (2000).

2.4.3 Análise espacial de dados geográficos

Segundo CÂMARA et. al. (2000), a ênfase da análise espacial de dados é "mensurar propriedades e relacionamentos, levando em conta a localização espacial do fenômeno em estudo. Sendo a idéia central incorporar o espaço à análise desejada".

Em análise espacial de dados geográficos, três tipos de dados podem ser identificados para caracterizar os problemas. São eles:

- *Análise de padrões de pontos*: analisa se os pontos em que os eventos ocorrem estão distribuídos de forma aleatória ou formando estruturas aglomeradas ou regulares, sendo também possível verificar se a distribuição dos pontos está associada à outra variável. Como exemplo verificar se a distribuições dos pontos de criminalidade estão relacionados com alguma variável sócio-econômica (CÂMARA et al., 2000).

- *Análise de superfícies*: consiste em “reconstruir a superfície analisada por interpolação com base na coleta de amostras, que podem estar associadas a pontos, linhas ou polígonos, necessitando definir o modelo de dependência espacial entre essas amostras” (PACHECO, 2005). Esse tipo de dados é geralmente proveniente de levantamentos de recursos naturais como mapas geológicos, topográficos, ecológicos, etc..
- *Análise de áreas*: os dados normalmente associados com áreas em geral referem-se a dados agregados de levantamentos populacionais (como censos e estatísticas de saúde). São disponibilizados em agregados para manter a confidencialidade dos dados (CÂMARA et al, 2000). “Supondo que as áreas são supostamente homogêneas internamente, mudanças importantes só ocorrem nos limites, permitindo a determinação, por exemplo, de índices populacionais com base nas medidas realizadas nessas áreas ou determinação de equações de regressão ajustadas aos dado” (PACHECO, 2005).

Nesse trabalho não se fez necessário o uso de análise de dados espaciais. Informações sobre como os dados devem ser gerados e armazenados, bem como maiores detalhes de como proceder com as análises podem ser encontrados nas obras de CÂMARA et al. (2000) e PACHECO (2005).

3 – Desenvolvimento do Trabalho

Nessa sessão são descritas as atividades envolvidas para a realização desse trabalho. A estrutura do desenvolvimento do trabalho segue as etapas do processo de descoberta de conhecimento apresentadas no capítulo de fundamentação teórica, finalizando com a distribuição espacial dos grupos no município de Florianópolis.

3.1 Considerações sobre a base de dados

A base de dados utilizada nas análises foi adquirida na Biblioteca do IBGE localizada no centro do município de Florianópolis. Está contida em um CD-ROM comercializado pelo IBGE sob o título de "Agregado por Setores Censitários dos Resultados do Universo – 2ª Edição" para os estados de Santa Catarina e Paraná. Os dados foram fornecidos gratuitamente, por se tratarem apenas de uma parcela do produto comercial.

Os dados desse produto estão distribuídos em vinte e uma planilhas eletrônicas no formato do MS Excel (para cada Unidade da Federação). Entretanto, os dados necessários para as análises estão armazenados em 2 planilhas eletrônicas, a planilha *responsavel1.xls* e *basico_uf.xls*. A planilha *responsável1.xls* fornece informações sobre os responsáveis por domicílios particulares permanentes por sexo, idade, alfabetização, anos de estudo e rendimento (IBGEa). A planilha *basico_uf.xls* fornece os códigos e nomes das subdivisões geográficas e a informação básica do cadastro de áreas (totais, médias e variâncias) (IBGEa).

Para o município de Florianópolis, cada planilha selecionada contém 460 registros e cada registro corresponde a um determinado setor censitário do município de Florianópolis.

Para o melhor entendimento dos termos presentes nos objetivos desse trabalho é citado abaixo a descrição de setor censitário, domicílio particular permanente, renda nominal mensal e anos de estudo (nível de instrução), conforme a documentação presente em IBGEa.

> *O setor censitário é a menor unidade territorial, com limites físicos identificáveis em campo, com dimensão adequada à operação de pesquisas e cujo conjunto esgota a totalidade do Território Nacional, o que permite assegurar a plena cobertura do País (*IBGEa*).*

> *Domicílio particular: quando o relacionamento entre seus ocupantes era ditado por laços de parentesco, de dependência doméstica ou por normas de convivência. (...), permanente: quando construído para servir exclusivamente à habitação e, na data de referência, tinha a finalidade de servir de moradia a uma ou mais pessoa (IBGEa).*

> *Considerou-se como rendimento nominal mensal da pessoa de 10 anos ou mais de idade, responsável pelo domicílio particular permanente, a soma do rendimento nominal mensal de trabalho com o proveniente de outras fontes que tinha na semana de referência, que foi a de 23 a 29 de julho de 2000 (IBGEa).*

> *A classificação de anos de estudo foi estabelecida com objetivo de compatibilizar os sistemas de ensino anteriores e atual. Essa classificação foi obtida em função da última série concluída com aprovação no nível ou grau mais elevado que a pessoa de 10 anos ou mais de idade, responsável pelo domicílio particular permanente, estava freqüentando ou havia freqüentado (IBGEa).*

Abaixo está apresentado o mapeamento da última série concluída com aprovação e a correspondência em anos de estudo, conforme documentação de IBGEa:

- Sem instrução e menos de 1 ano de estudo, para a pessoa que nunca freqüentou escola ou, embora tenha freqüentado, não concluiu pelo menos a 1ª série do ensino fundamental, 1º grau ou elementar;
- 1 ano de estudo, para a pessoa que concluiu: curso de alfabetização de adultos; ou a 1ª série do ensino fundamental, 1º grau ou elementar;
- 2 anos de estudo, para a pessoa que concluiu a 2ª série do ensino fundamental, 1º grau ou elementar;
- 3 anos de estudo, para a pessoa que concluiu a 3ª série do ensino fundamental, 1º grau ou elementar;
- 4 anos de estudo, para a pessoa que concluiu: a 4ª série do ensino fundamental ou 1º grau; ou, no mínimo, a 4ª série e, no máximo, a 6ª série do elementar;
- 5 anos de estudo, para a pessoa que concluiu: a 5ª série do ensino fundamental ou 1º grau; ou a 1ª série do médio 1º ciclo;

- 6 anos de estudo, para a pessoa que concluiu: a 6ª série do ensino fundamental ou 1º grau; ou a 2ª série do médio 1º ciclo;
- 7 anos de estudo, para a pessoa que concluiu: a 7ª série do ensino fundamental ou 1º grau; ou a 3ª série do médio 1º ciclo;
- 8 anos de estudo, para a pessoa que concluiu: a 8ª série do ensino fundamental ou 1º grau; ou, no mínimo, a 4ª série e, no máximo, a 5ª série do médio 1º ciclo;
- 9 anos de estudo, para a pessoa que concluiu a 1ª série do ensino médio, 2º grau ou médio 2º ciclo;
- 10 anos de estudo, para a pessoa que concluiu a 2ª série do ensino médio, 2º grau ou médio 2º ciclo;
- 11 anos de estudo, para a pessoa que concluiu, no mínimo, a 3ª série e, no máximo, a 4ª série do ensino médio, 2º grau ou médio 2º ciclo;
- 12 anos de estudo, para a pessoa que concluiu a 1ª série do superior;
- 13 anos de estudo, para a pessoa que concluiu a 2ª série do superior;
- 14 anos de estudo, para a pessoa que concluiu a 3ª série do superior;
- 15 anos de estudo, para a pessoa que concluiu a 4ª série do superior;
- 16 anos de estudo, para a pessoa que concluiu a 5ª série do superior;
- 17 anos de estudo ou mais, para a pessoa que concluiu a 6a série do superior ou mestrado ou doutorado.

Ainda conforme a documentação disponível em IBGEa, há a possibilidade de alguns registros (setores censitários) não apresentarem valores (estarem nulos ou *missing*) para as variáveis de renda e anos de estudo. Para não permitir a identificação dos entrevistados, esses valores e outros foram omitidos nos setores censitários que possuem menos de cinco domicílios particulares permanentes.

Nesse trabalho o termo “responsáveis por domicílio particular permanente” pode ser identificado apenas como “responsáveis” ou “responsáveis por domicílio”.

3.2 Seleção das variáveis

Para satisfazer os objetivos descritos no capítulo 1, foram identificas nas planilhas *Basico_UF.xls* e *Responsavel1.xls* as variáveis relevantes para as análises. Essas variáveis estão apresentadas nas tabelas 3.1 e 3.2.

Tabela 3-1 – Variáveis selecionadas na planilha *Basico_UF*.

Variável	**Descrição**
cod_setor	Código do setor censitário.
var03	Média do rendimento nominal mensal dos responsáveis no setor.
var10	Média do número de anos de estudo dos responsáveis no setor.

Tabela 3-2 – Variáveis selecionadas na *planilha Responsavel1_UF*.

Variável	Descrição
cod_setor	Código do setor censitário.
v0580	Responsáveis por domicílios particulares permanentes sem instrução ou com menos de 1 ano de estudo.
v0581	Responsáveis por domicílios particulares permanentes com 1 ano de estudo.
v0582	Responsáveis por domicílios particulares permanentes com 2 anos de estudo.
v0583	Responsáveis por domicílios particulares permanentes com 3 anos de estudo.
v0584	Responsáveis por domicílios particulares permanentes com 4 anos de estudo.
v0585	Responsáveis por domicílios particulares permanentes com 5 anos de estudo.
v0586	Responsáveis por domicílios particulares permanentes com 6 anos de estudo.
v0587	Responsáveis por domicílios particulares permanentes com 7 anos de estudo.
v0588	Responsáveis por domicílios particulares permanentes com 8 anos de estudo.
v0589	Responsáveis por domicílios particulares permanentes com 9 anos de estudo.
v0590	Responsáveis por domicílios particulares permanentes com 10 anos de estudo.
v0591	Responsáveis por domicílios particulares permanentes com 11 anos de estudo.
v0592	Responsáveis por domicílios particulares permanentes com 12 anos de estudo.
v0593	Responsáveis por domicílios particulares permanentes com 13 anos de estudo.
v0594	Responsáveis por domicílios particulares permanentes com 14 anos de estudo.
v0595	Responsáveis por domicílios particulares permanentes com 15 anos de estudo.
v0596	Responsáveis por domicílios particulares permanentes com 16 anos de estudo.
v0597	Responsáveis por domicílios particulares permanentes com 17 ou mais anos de estudo.
v0599	Responsáveis por domicílios particulares permanentes com anos de estudo determinado.

A variável v0599 indica a contagem de responsáveis que possuem anos de estudo determinados para cada setor, ou seja, representa a soma dos valores das variáveis v0581 à v0597. A variável v0598, totalizando número de responsáveis com anos de estudo não determinado, não foi utilizada pois seu valor é insignificante (representando duzentos e onze responsáveis num universo de mais de cem mil).

A variável v10 (mediaAnosEstudo) apesar de selecionada não é utilizada na análise de conglomerados, sendo utilizada apenas para resumir as informações presentes nas variáveis v0580 à v0597 nas análises descritivas apresentadas adiante.

3.3 Pré-processamento

Conforme descrito na seção 3.1, há a possibilidade de alguns setores censitários não apresentarem as informações de renda e escolaridade. Por essa razão foram identificados e removidos antes da análise exploratória das variáveis selecionadas.

Utilizando o software MS Excel para a identificação dos registros com valores *missing* observamos a presença de dezessete tuplas com essa característica. Esses registros foram

removidos de ambas as planilhas de dados. Os valores da variável *cod_setor*, que identifica o setor censitário dos itens excluídos, foram armazenados em um arquivo para futura utilização na distribuição espacial no mapa do município de Florianópolis.

Para facilitar a compreensão da análise exploratória dos dados selecionados, optou-se por agrupar as variáveis de anos de estudos em categorias (ver tabela 3.3), seguindo as classificações citadas na seção 3.1.

Tabela 3-3 –Categorização das variáveis de anos de estudo.

Variável	*Descrição*
cod_setor	Código do setor censitário.
semInstrucao	Responsáveis por domicílios sem instrução ou com menos de 1 ano de estudo.
primeiraQuarta	Responsáveis por domicílios que tiveram o grau máximo de estudo entre a primeira e a quarta séries do ensino fundamental.
quintaOitava	Responsáveis por domicílios que tiveram o grau máximo de estudo entre a quinta e a oitava séries do ensino fundamental
ensinoMedio	Responsáveis por domicílios que tiveram o grau máximo de estudo alguma série do ensino médio.
ensinoSuperior	Responsáveis por domicílios que tiveram o grau máximo de estudo alguma série do ensino superior.
mestradoDoutorado	Responsáveis por domicílios que tiveram o grau máximo de estudo o mestrado ou doutorado.
anosEstudoDet	Responsáveis por domicílios permanentes com anos de estudo determinado.

Também com a intenção de facilitar a análise dos dados, a variável var03, que representa a renda média dos responsáveis no setor, foi utilizada para a criação de outra variável representando a renda média em salários mínimos no setor. A nova variável, chamada rendaMediaSM, teve seus valores calculados através da razão de var03 pelo valor do salário mínimo (R$ 151,00 segundo IBGEa) vigente na época do censo 2000. A variável var10 também foi renomeada para mediaAnosEstudo, conforme a tabela 3.4.

Tabela 3-4 - Variáveis selecionadas na planilha *Basico_UF*: alteração de renda média nominal para renda média em salários mínimos.

Variável	*Descrição*
cod_setor	Código do setor censitário.
rendaMediaSM	Média do rendimento mensal em salários mínimos das pessoas responsáveis pelo domicílio.
mediaAnosEstudo	Média do número de anos de estudo das pessoas responsáveis por domicílios.

Uma nova modificação nos valores das variáveis presentes na tabela 3.3 foi necessária. Pelo fato de os setores censitários possuírem diferentes quantidades de responsáveis por

domicílio não seria adequado analisar essas variáveis utilizando a contagem de responsáveis para cada categoria de séries cursadas. Optou-se por utilizar a proporção que cada variável possui em relação ao total de responsáveis por setor com anos de estudo determinado (v0599 ou anosEstudoDet), estando apresentadas na tabela 3.5.

Tabela 3-5 - Variáveis selecionadas na planilha *Responsavel1_UF* representando a proporção de cada categoria de ensino no setor.

Variável	*Descrição*
cod_setor	Código do setor censitário.
per_semInstrucao	Proporção, no setor, de responsáveis por domicílios sem instrução ou com menos de 1 ano de estudo.
per_primeiraQuarta	Proporção, no setor, de responsáveis por domicílios que tiveram o grau máximo de estudo entre a primeira e quarta séries do ensino fundamental.
per_quintaOitava	Proporção, no setor, de responsáveis por domicílios que tiveram o grau máximo de estudo entre a quinta e oitava séries do ensino fundamental.
per_ensinoMedio	Proporção, no setor, de responsáveis por domicílios que tiveram o grau máximo de estudo alguma série do ensino médio.
per_ensinoSuperior	Proporção, no setor, de responsáveis por domicílios que tiveram o grau máximo de estudo alguma série do ensino superior.
per_mestradoDoutorado	Proporção, no setor, de responsáveis por domicílios que tiveram o grau máximo de estudo o mestrado ou doutorado.

Nas análises são consideradas as variáveis presentes nas tabelas 3.4 e 3.5. Cada variável possui 443 observações, pois foram removidos os dezessete registros que possuem valores *missing*.

3.3.1 Análise exploratória dos dados

Nessa seção apresentamos resumos estatísticos e gráficos para as variáveis selecionadas, com objetivo de estudar o conjunto de dados a serem analisados.

Tabela 3-6 – Estatística descritiva para as variáveis mediaAnosEstudo e rendaMediaSM.

	mediaAnosEstudo	**rendaMédiaSM**
Média	9,76	10,88
Mediana	9,69	8,98
Desvio padrão	2,70	8,12
Mínimo	2,61	1,28
Máximo	14,82	87,48

Pela tabela 3.6 observamos que a média de anos de estudos dos setores censitários é correspondente a primeira série do ensino médio. A variável rendaMediaSM destaca-se por apresentar valor máximo em torno de 87 salários mínimos, ou seja, um valor distante da

média dos setores e quase setenta vezes maior que seu valor mínimo. Essa assimetria na distribuição da renda pode ser observada na figura 3.2.

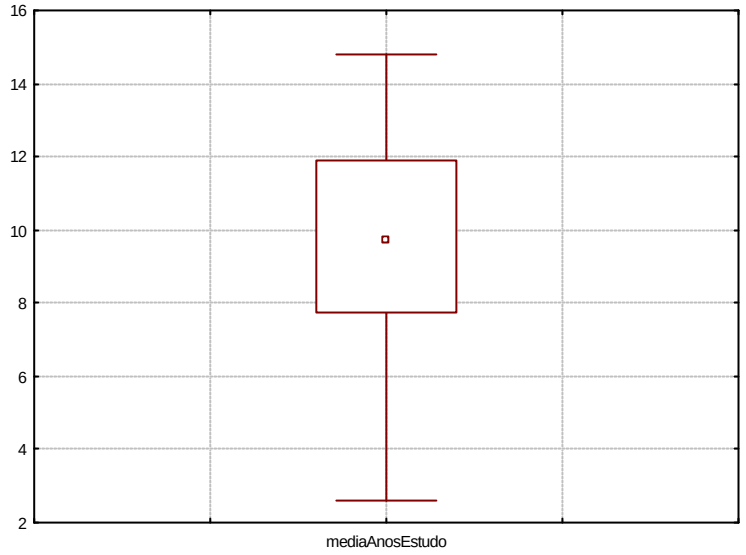

Figura 3-1 - Box plot para a varável mediaAnosEstudo

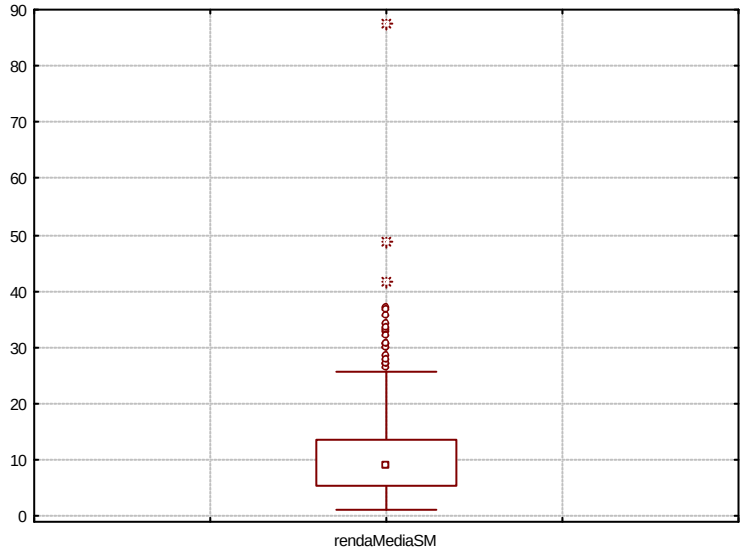

Figura 3-2 - Box plot para a variável rendaMediaSM

Analisando as tabelas 3.7 e 3.8 e a figura 3.3 podemos notar que a distribuição das variáveis per_semInstrucao e per_mestradoDoutorado são semelhantes, ou seja, da mesma maneira que existem setores censitários com, aproximadamente, 30% de seus responsáveis sem instrução, há a presença de setores com aproximadamente a mesma proporção de responsáveis com mestrado ou doutorado. Através dessas tabelas é possível verificar que a soma das proporções de responsáveis no setor com até o ensino médio concluído é de 66,32% enquanto que a soma das proporções de responsáveis com ensino superior e mestrado é de

33,69%. Essas medidas serão úteis para se comparar cada grupo formado entre si e com os valores acima calculados.

Deve-se notar também que, em média, os setores possuem em torno de 28% de seus responsáveis por domicílio com ensino superior. Essa proporção é maior que qualquer outro nível de ensino apresentado na tabela 3.7 e 3.8.

Nessas tabelas, a presença de zeros como o valor mínimo de cada faixa de instrução justifica-se pela existência de setores censitários com baixa quantidade de responsáveis por domicílios, podendo ocorrer assim setores sem nenhum responsável para determinada variável.

Tabela 3-7 - Estatística descritiva para as variáveis per_semInstrucao, per_primeiraQuarta, per_quintaOitava e per_ensinoMedio.

	per_semInstrucao	per_primeiraQuarta	per_quintaOitava	per_ensinoMedio
Média	3.25	19.38	18.62	25.07
Mediana	2.05	17.50	17.73	25.19
Desvio padrão	4.00	13.47	10.42	8.65
Mínimo	0.00	0.00	0.00	0.00
Máximo	28.99	6.67	66.67	46.12

Tabela 3-8 - Estatística descritiva para as variáveis per_ensinoSuperior e per_MestradoDoutorado.

	per_ensinoSuperior	per_mestradoDoutorado
Média	27.99	5.70
Mediana	25.66	3.36
Desvio padrão	18.81	5.90
Mínimo	0.00	0.00
Máximo	69.08	29.13

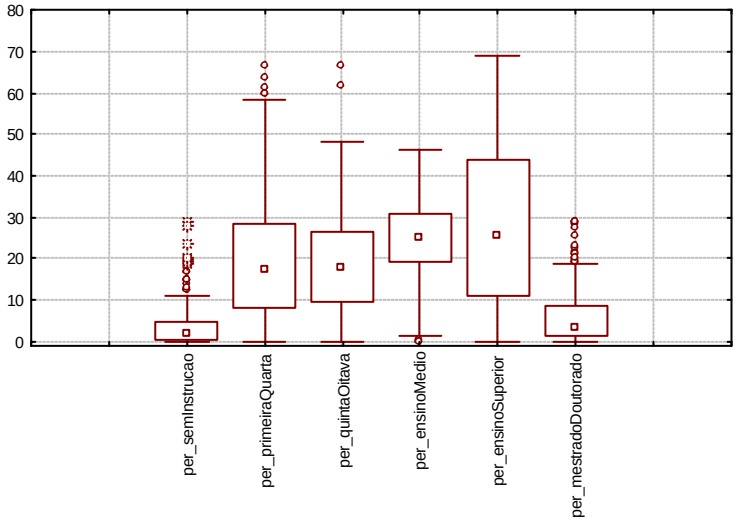

Figura 3-3 - Box plot para as variáveis que representam a proporção do nível de instrução em cada setor censitário.

Com a análise exploratória observarmos a presença de diversos setores com valores discrepantes ou extremos na variável rendaMediaSM (figura 3.2). Por opção de modelagem, para atender aos objetivos do agrupamento, optou-se por não excluí-los das análises ou de usar técnicas de regressão para substituir seus valores.

Como a variável mediaAnosEstudo não possui valores outliers ou extremos (figura 3.1), decidimos não remover os valores discrepantes nas variáveis que representam a proporção de responsáveis por domicílio de acordo com a faixa de escolaridade, pois essas variáveis representam um particionamento da variável mediaAnosEstudo.

Apenas por curiosidade identificamos os três setores censitários com valores extremos de renda (tabela 3.9) para serem posteriormente localizados no município de Florianópolis (figura 3.17).

Tabela 3-9 – Setores censitários da base de dados que possuem valores extremos na variável rendaMediaSM.

cod_setor	mediaAnosEstudo	rendaMediaSM	bairro
420540705000033	14.11	41.54	Centro
420540705000039	14.14	48.62	Centro
420540705000059	14.58	87.48	Centro

3.4 Integração e transformação dos dados

Após os dados serem selecionados, limpos e pré-processados, eles precisam ser formatados e armazenados adequadamente para que os algoritmos de *data mining* possam ser aplicados. Os próximos sub-tópicos apresentam as transformações e integrações implementadas para esse trabalho.

3.4.1 Transformação dos dados

Seguindo as orientações sobre análise de agrupamentos para variáveis envolvendo grandezas diferentes (apresentado no tópico 2.3.1.5), nossas variáveis selecionadas foram transportadas para a faixa de valores [0, 1].

Para as variáveis per_semInstrucao, per_primeiraQuarta, per_quintaOitava, per_ensinoMedio, per_ensinoSuperior e per_mestradoDoutorado a transformação para faixa [0, 1] foi facilitada pelo fato de estarem indicando uma proporção (percentual) para cada setor. Dessa forma, bastou-se dividirmos seus valores por 100 (cem) para transportá-los para a faixa adequada.

Para a variável rendaMediaSM que possui uma distribuição assimétrica de seus valores, foi aplicada a função logarítmica (equação 9) para deixar seus valores com uma distribuição normal.

$$\text{logRendaMediaSM} = \log_{10}\{\text{rendaMediaSM}\} \qquad \text{(Equação 9)}$$

Após a aplicação da função logarítmica foi utilizada a equação 2, apresentada na seção 2.2.3, para transformar seus valores para a faixa [0, 1]. Dessa forma, foi criada outra variável representando essas transformações, a variável *mmlog_rendaMediaSM*.

Embora algumas das variáveis de proporção da escolaridade por setor também apresentem assimetria de distribuição, com valores outliers e extremos, esses não foram removidos ou transformados para não deformar a proporcionalidade existente entre as seis variáveis que indicam o nível de instrução.

3.4.2 Integração dos dados

A integração dos dados foi realizada utilizando o sistema gerenciador de banco de dados MS SQL Server versão 2000 da Microsoft. As planilhas eletrônicas contendo as variáveis selecionadas e transformadas foram importadas para o MS SQL Server. A execução do script SQL abaixo resultou na criação da tabela *"selecao"* contendo a integração das duas planilhas com as variáveis selecionadas.

```
select   sb.cod_setor, sb.rendaMediaSM, sb.mmlog_rendaMediaSM, sb.mediaAnosEstudo,
sr.per_semInstrucao, sr. per_primeira_quarta, sr. per_quinta_oitava, sr. per_ensino_medio,
sr. per_superior, sr. per_mestrado_doutorado
into     selecao
from     selecaoBasico sb, selecaoResponsavel1 sr
where    sb.cod_setor = sr.cod_setor
```

Através do recurso de exportação dos dados disponível no ambiente do MS SQL Server exportou-se a tabela "selecao" para uma planilha eletrônica no formato do MS Excel. Na tabela 3.10 estão apresentadas as variáveis selecionadas, transformadas e integradas que serão utilizadas na etapa de data mining.

Tabela 3-10 - Variáveis a utilizar na análise de conglomerados após serem transformadas e integradas.

Variável	*Descrição*
cod_setor	Código do setor censitário.
mmlog_rendaMediaSM	Variável rendaMediaSM após transformação.
per_semInstrucao	Proporção, no setor, de responsáveis por domicílios sem instrução ou com menos de 1 ano de estudo.
per_primeiraQuarta	Proporção, no setor, de responsáveis por domicílios que tiveram o grau máximo de estudo entre a primeira e quarta séries do ensino fundamental.
per_quintaOitava	Proporção, no setor, de responsáveis por domicílios que tiveram o grau máximo de estudo entre a quinta e oitava séries do ensino fundamental.
per_ensinoMedio	Proporção, no setor, de responsáveis por domicílios que tiveram o grau máximo de estudo alguma série do ensino médio.
per_ensinoSuperior	Proporção, no setor, de responsáveis por domicílios que tiveram o grau máximo de estudo alguma série do ensino superior.
per_mestradoDoutorado	Proporção, no setor, de responsáveis por domicílios que tiveram o grau máximo de estudo o mestrado ou doutorado.

3.5 Data mining

A tarefa de *data mining* utilizada corresponde a tarefa de agrupamento. Essa tarefa visa agrupar os registros mais semelhantes entre si de acordo com as variáveis selecionadas para a análise (tabela 10).

A tarefa de agrupamento segue a metodologia não supervisionada de conhecimento e a técnica aplicada foi a de *cluster analysis* (ou análise de conglomerados), estando apresentada em maiores detalhes na seção 2.3.

O software utilizado para a mineração de dados foi o *The SAS System for Windows* versão 8.2, mais especificamente seu módulo *SAS Enterprise Miner*.

No módulo *SAS Enterprise Miner* utilizou-se os componentes *Input Data Source*, *Clustering* e *Insight*. No primeiro componente apenas selecionou-se a base de dados que será utilizada nas análises, embora outras opções estivessem disponíveis. A base de dados selecionada foi o arquivo exportado pelo MS SQL Server na etapa de integração e transformação dos dados.

O componente *Clustering* é o responsável pela execução do algoritmo de agrupamento. Nele são indicadas as variáveis utilizadas, o algoritmo de agrupamento (WARD) e o número máximo de clusters a serem formados (6). As demais opções disponíveis foram mantidas com o valor padrão do componente. Optamos por definir o número máximo

de cluster como seis para facilitar as análises nesse trabalho como também formar grupos mais distintos entre si.

Por fim, o componente *Insight* permite a exploração e análise dos dados em que está conectado, e nesse caso seriam os dados do componente *Clustering*. Entretanto, esse componente foi utilizado apenas para exportar os registros com seus respectivos *clusters* identificados.

Após a configuração dos componentes necessários para realizar o agrupamento foi selecionada a opção *run* (executar) no componente *Clustering*. Sem a necessidade de intervenção do usuário, o software se encarrega de realizar os agrupamentos e ao concluir são exibidos os resultados.

3.5.1 Os resultados da análise de agrupamento

Os resultados da análise de agrupamento que são apresentados pelos *SAS Enterprise Miner* são facilmente interpretáveis e possuem informações suficientes para se verificar a composição de cada *cluster*. Nesse trabalho os clusters são analisados através das seguintes opções dos resultados gerados:

- *Tabela de variáveis*: Exibe todas as variáveis que foram utilizadas no *clustering analysis* acompanhadas por uma medida de importância, que é apresentada numa faixa entre 0 e 1. A medida de importância representa o quanto a variável foi importante no processo de formação dos agrupamentos, ou seja, o quanto ela foi significante no processo de divisão dos dados em grupos. Se alguma variável apresentar o valor de importância zero, significa que ela não foi utilizada como uma variável de particionamento durante a execução do *clustering analysis*, ao contrário das variáveis que possuem valores mais próximos de 1.

- *Tabela estatística*: apresenta em uma tabela as informações de cada cluster. Dentre as informações disponíveis, as principais para esse trabalho foram a quantidade de observações em cada grupo, a média para cada variável de entrada, e o *cluster* mais próximo.

- *Resumo estatístico*: semelhante a tabela estatística, porém disponibiliza maiores informações das variáveis para cada *cluster*, como valores mínimos, máximos, média e desvio padrão.

- *Perfis dos cluster*: exibe em forma de um gráfico de barras tridimensional a distribuição de uma determinada variável em cada cluster formado. Permite visualizar a distribuição de qualquer variável presente na base de dados, inclusive as variáveis que não foram utilizadas na formação do agrupamento (como por exemplo, a variável mediaAnosEstudo e rendaMediaSM).

3.5.1.1 Grau de importância das variáveis

Como resultado da análise de conglomerados obtivemos a formação de cinco grupos de setores distintos quanto a renda média dos responsáveis e do nível de instrução. Essa quantidade de setores censitários foi definido manualmente para possibilitar encontrar grupos de setores censitários mais distintos entre si e facilitar as análises apresentadas nesse relatório.

Pela tabela 3.11 pode-se verificar que a variável com maior grau de importância na formação dos conglomerados foi a variável per_ensinoSuperior. Essa variável recebeu o valor máximo de importância, ou seja, 1 (um). Observa-se a variável que representa a renda média em salários mínimos no setor (mmlog_rendaMediaSM) como a segunda variável em ordem de significância, com 0,6421 pontos, seguida pela proporção de responsáveis com até a quarta concluída (0,5417) e pela proporção de responsáveis com até a oitava série concluída (0,2747).

As variáveis per_ensinoMedio e per_mestradoDoutorado foram pouco significativas no processo. A variável de proporção de responsáveis sem instrução no setor não teve nenhuma importância na divisão dos registros em clusters, pois apresenta o valor zero.

Tabela 3-11 – Importância de cada variável para a formação dos agrupamentos.

Variável	Importância
logRendaMediaSM	0,6421
per_semInstrucao	0
per_primeiraQuarta	0,5417
per_quintaOitava	0,2747
per_ensinoMedio	0,0910
per_ensinoSuperior	1
per_mestradoDoutorado	0,0925

3.5.1.2 Resumo estatístico dos grupos formados

Através da tabela 3.12 observamos a quantidade de setores que constituem cada cluster e verificamos a existência de uma relação direta entre a rendaMediaSM e mediaAnosEstudo. Através dela pode-se traçar um *ranking* dos grupos formados utilizando como critério a renda e os anos de estudo. Dessa forma, em ordem decrescente de renda e instrução, os grupos seriam *cluster* 1, *cluster* 3, *cluster* 4, *cluster* 2 e *cluster* 5.

Tabela 3-12 – Freqüência de setores por cluster, renda média e media de anos de estudo dos grupos formados.

Cluster	Freqüência	log_RendaMediaSM	rendaMediaSM	mediaAnosEstudo
1	51	0.709003768	27,02	13,75
2	16	0.256944719	3,92	6,54
3	149	0.542390372	13,08	11,75
4	185	0.391324909	7,03	8.37
5	42	0.193644054	3,06	5,18

Comparando-a com os dados da tabela 3.6,observamos que a renda média em salários mínimos do grupo 1 chama a atenção por ser duas vezes e meia superior a média apresentada no município e sua média de anos de estudo ser 1,4 vezes maior que a média municipal. O grupo 5, ao contrário, apresenta uma renda 3,5 vezes inferior a de Florianópolis e os anos de estudo pouco superior a metade. Os demais clusters formados apresentam valores intermediários aos apresentados nesse parágrafo, sendo que apenas o cluster 3 possui valores médios superiores aos encontrados em Florianópolis.

Tabela 3-13 – Proporção média de responsáveis por cluster para cada nível de instrução.

Cluster	per_semInstrucao	per_primeiraQuarta	per_quintaOitava	per_EnsinoMedio	Sub Total
1	0,22	3,64	5,81	17,66	**27,33**
2	7,01	27,00	42,21	18,79	**95,01**
3	1,21	9,51	11,28	27,94	**49,94**
4	3,76	25,08	24,24	28,06	**81,14**
5	10,51	45,57	26,48	13,10	**95,66**

Tabela 3-14 – Proporção média de responsáveis por cluster para cada nível de instrução.

Cluster	per_ensinoSuperior	per_mestradoDoutorado	Sub Total
1	57,96	14,71	**72,67**
2	4,43	0,55	**4,98**
3	41,69	8,37	**50,06**
4	16,30	2,55	**18,85**
5	3,53	0,81	**4,34**

As tabelas 3.13 e 3.14 detalham os níveis de instrução em cada grupo formado. É interessante perceber o cluster 1, classificado como o grupo de maior renda e escolaridade, possui quase três quartos de seus responsáveis por domicílio com ensino superior ou mestrado/doutorado. Contrastando com o cluster 5, o de menor renda e escolaridade, onde apenas 4,94% de seus responsáveis possuem ensino superior ou mestrado/doutorado e, 82,56% possuem até a oitava série do ensino fundamental concluída.

3.5.1.3 Representações gráficas dos clusters

Apesar do software *SAS Enterprise Miner* disponibilizar uma representação gráfica tridimensional para comparar-se os agrupamentos formados, optamos por utilizar o software *Statistica* para gerar gráficos *box plot* para comparações gráficas dos agrupamentos.

O gráfico apresentado na figura 3.4 está exibindo a distribuição das médias de anos de estudo que cada grupo apresenta em seus setores. São visíveis as diferenças apresentadas entre os grupos formados e pode-se destacar a diferença de anos de estudo entre os clusters 1 e 5 (como já foi comentado na seção anterior).

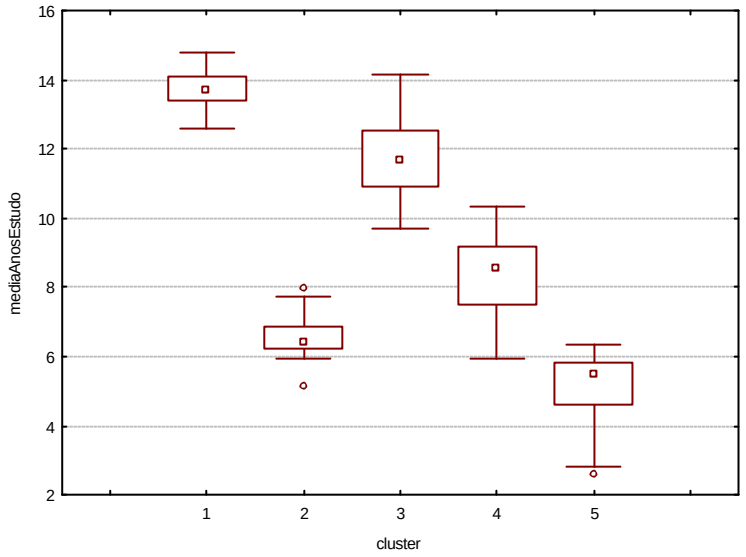

Figura 3-4 - Box plot da variável mediaAnosEstudo para cada cluster.

As figuras 3.5 e 3.6 representam a distribuição da mesma grandeza entre os clusters, a renda média em salários mínimos. A diferença é que na primeira figura são apresentados os valores transformados, ou seja, variando do valor mínimo zero ao valor máximo 1, enquanto que na figura 3.6 são apresentados os valores reais em salários mínimos para o ano 2000. É interessante perceber que o gráfico da figura 3.5 é bastante semelhante ao exibido na figura 3.4 (*box plot* para a média de anos de estudo).

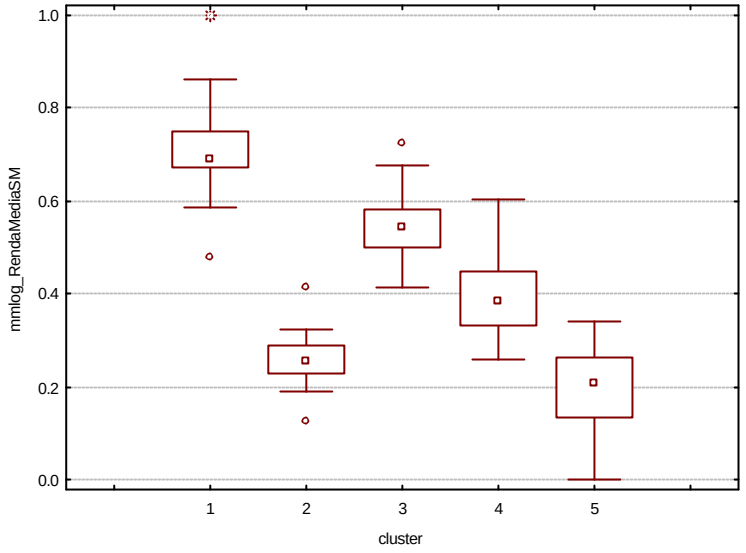

Figura 3-5 - Box plot da variável mmlog_rendaMediaSM para cada cluster.

Verifica-se através do gráfico representado na figura 3.6 que os valores discrepantes e extremos para a variável rendaMediaSM (apresentados na figura 3.2) foram todos agrupados no cluster 1, o de maior renda e de instrução mais elevada.

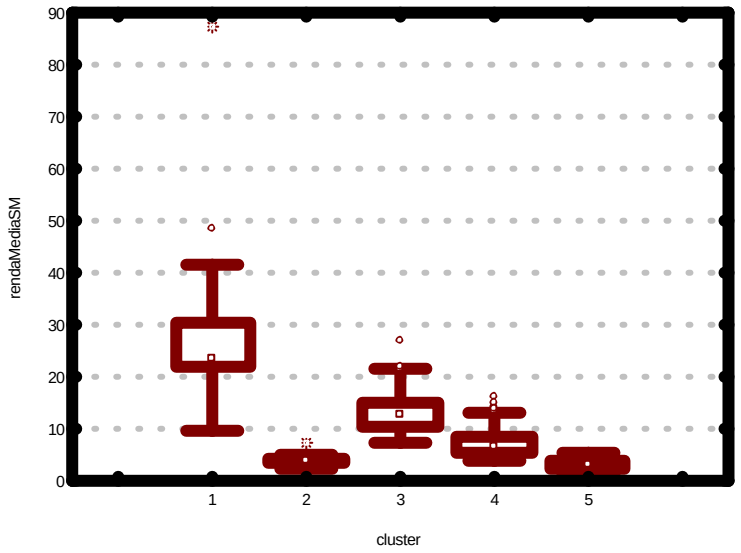

Figura 3-6 - Box plot da variável rendaMediaSM para cada cluster.

Apesar da variável per_semInstrucao ter apresentado grau de importância zero para a formação dos agrupamentos, pode-se verificar na figura 3.7, que o cluster 1, o de maior nível de instrução, é constituído apenas de setores censitários com menores proporções de sem instrução. O contrário é observado no cluster 5, que apresenta os setores censitários com os setores com as maiores proporções de responsáveis sem instrução.

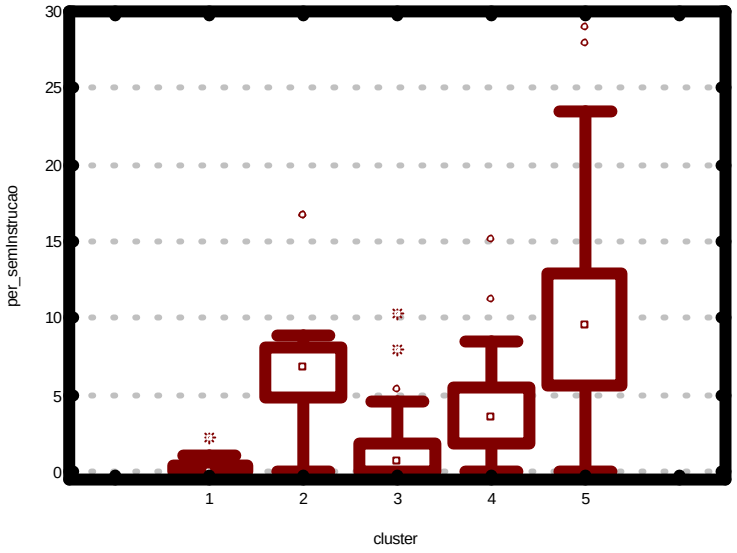

Figura 3-7 - Box plot da variável per_semInstrucao para cada variável.

Nos gráficos das figuras 3.8 e 3.9 destacam-se a distribuição das variáveis per_primeiraQuarta e per_quintaOitava nos clusters 2 e 5, onde o cluster 5 possui os setores com maior proporção de responsáveis com até a quarta série concluída e o grupo 2 possui maior concentração de setores com grandes percentagens de responsáveis com instrução de quinta à oitava séries. Essas diferenças podem ser verificadas através da tabela 3.13.

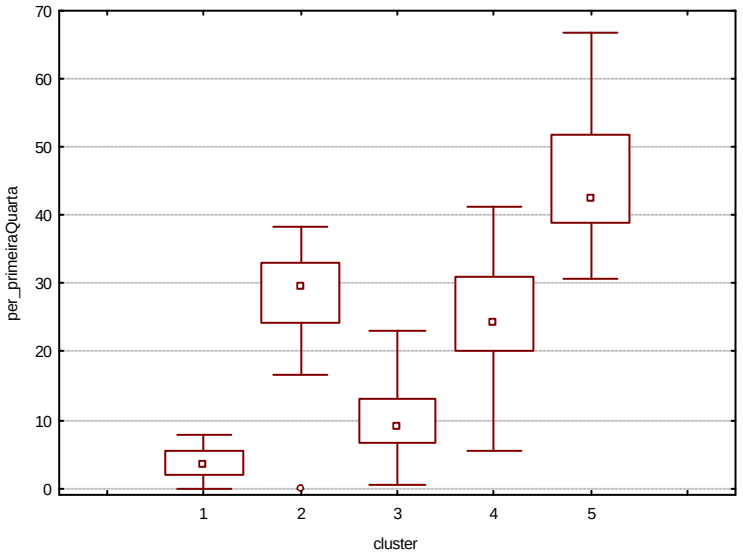

Figura 3-8 - Box plot da variável per_primeiraQuarta

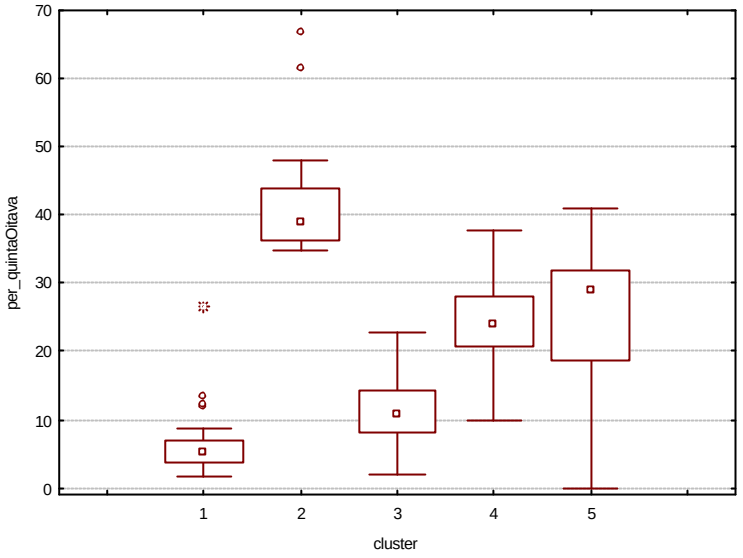

Figura 3-9 - Box plot da variável per_quintaOitava para cada cluster.

A variável per_ensinoMedio possui baixo grau de importância na divisão dos registros em grupos e apresenta o gráfico mais equilibrado de proporções de responsáveis por nível de instrução, como pode ser observado na figura 3.10.

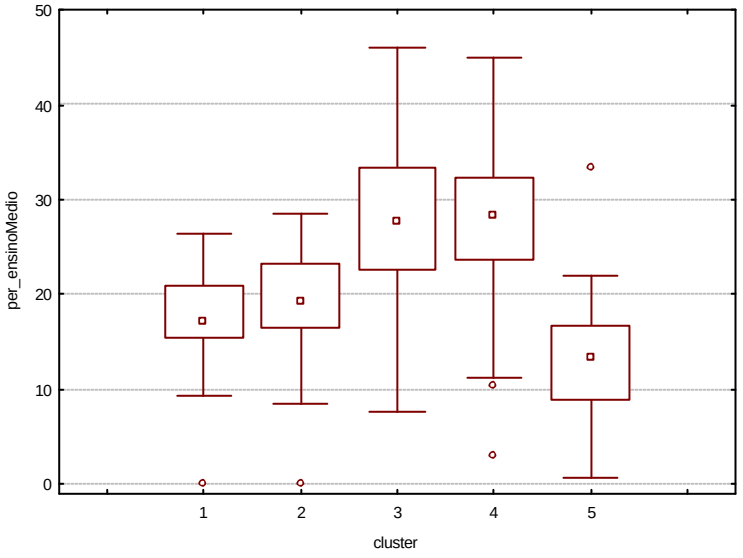

Figura 3-10 - Box plot da variável per_ensinoMedio para cada cluster.

A variável per_ensinoSuperior apresenta o maior grau de importância na formação dos clusters é representada na figura 3.11. Nela pode-se verificar com facilidade a distinção entre os grupos, causadas pela variável per_ensinoSuperior. No gráfico observamos que os setores com menor proporção de responsáveis com ensino superior foram agrupados nos clusters 2 e

5, enquanto que o cluster 1 recebeu setores com maiores proporções de responsáveis com nível superior.

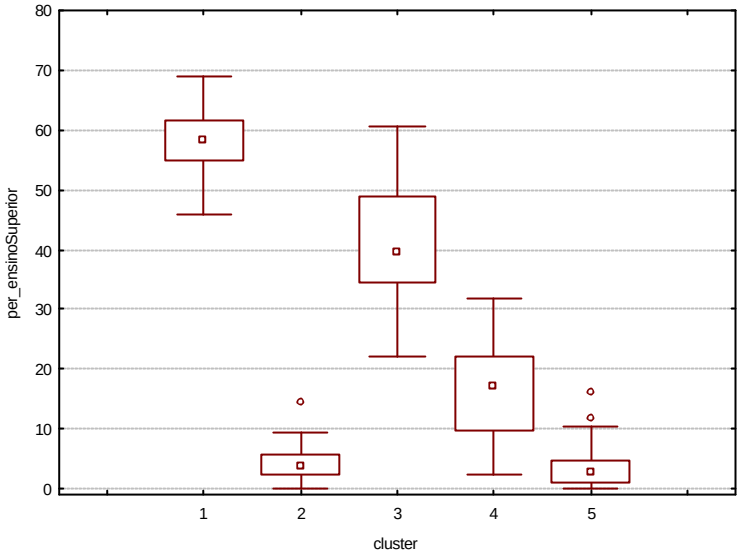

Figura 3-11 - Box plot da variável per_ensinoSuperior para cada cluster.

A figura 3.12 representa o gráfico da variável `per_mestradoDoutorado`. Novamente percebe-se que o grupo 1 e 3 apresentam os setores com maiores proporções de responsáveis por domicílio com mestrado e/ou doutorado. É interessante notar que alguns setores possuem aproximadamente 30% de responsáveis com mestrado e doutorado.

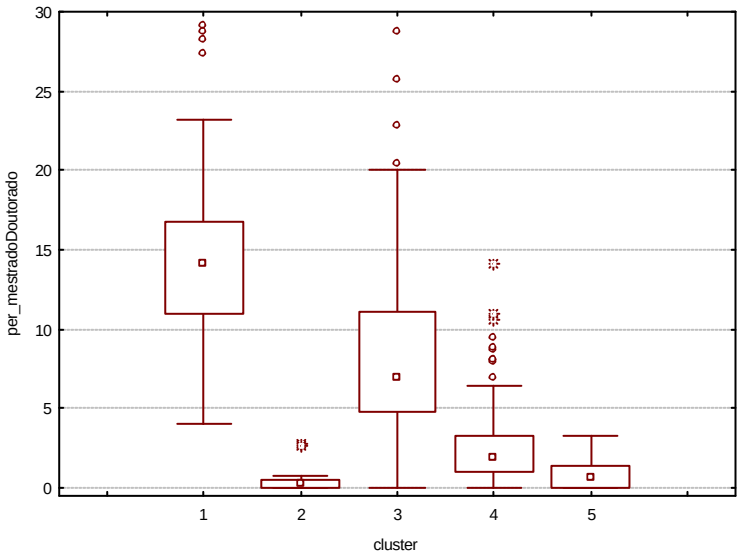

Figura 3-12 - Box plot da variável per_mestradoDoutorado para cada cluster.

3.6 Distribuição espacial dos grupos

A figura 3.15 representa a distribuição espacial no município de Florianópolis dos clusters formados a partir da análise de conglomerados realizada nas sessões anteriores. A figura 3.16 representa a mesma distribuição ampliada para exibir com maior nitidez a região central e continental do município.

Os cartogramas representados nas figuras 3.15 e 3.16 foram construídos com o auxílio do *ArcExplorer* versão 9.1.0. O *ArcExplorer* é um GIS que permite a visualização e consulta em bases de dados geo-referenciadas armazenadas localmente ou na internet. É desenvolvido pela ESRI (2005) e distribuído gratuitamente na internet. O ArcExplorer está apresentado na figura 3.14.

Antes de utilizar essa ferramenta foi necessário realizar o *download* de um produto disponibilizado pelo IBGE, a base de dados geo-referenciada sob o título de *"Malha de setor censitário urbano digital do distrito-sede dos municípios do Brasil 2000"*. Esse produto é disponibilizado gratuitamente, sendo necessário apenas preencher um cadastro simples.

O produto disponibiliza as malhas de sub-distrito, bairro e setor censitário urbano digital do distrito-sede dos municípios do Brasil, situação vigente em 2000, nos formatos *AGF* e *SHAPE* para 1 058 municípios (IBGEa).

Cada município possui três arquivos para realizar a referência geográfica. A descrição dos arquivos é apresentada no tópico 2.4.2.1. Para distribuir espacialmente os *clusters* encontrados na etapa de *data mining* foi necessário editar manualmente os arquivos da base de dados geo-referenciada. Essa edição foi realizada de forma semelhante à realizada na integração dos dados, ou seja, importou-se o arquivo *.dbf* (arquivo que contém os dados) da base geo-referenciada para o *MS SQL Server*. Uma vez importada a tabela bastou-se fazer a união dessa tabela com a tabela utilizada na descrição dos clusters. A junção foi realizada de forma a manter todas as informações originais da base geo-referenciada adicionando-se apenas a informação do cluster a que pertence cada setor censitário. Após a união a tabela resultante foi exportada, sendo necessário sobrescrever o arquivo *.dbf* original por esse, e utilizar o *AcrExplorer* para visualizar as informações.

No software *ArcExplorer* selecionou-se o arquivo *shape* para ser visualizado. Inicialmente é exibida apenas a malha dos setores censitários sem distinção de cores entre eles, pois nenhuma variável foi selecionada para visualização. Como nossa intenção é

apresentar cada setor censitário com uma cor que identifique o cluster a que pertence, através do menu *Layer Properties* alteramos as opções disponíveis na tela conforme a figura 3.13.

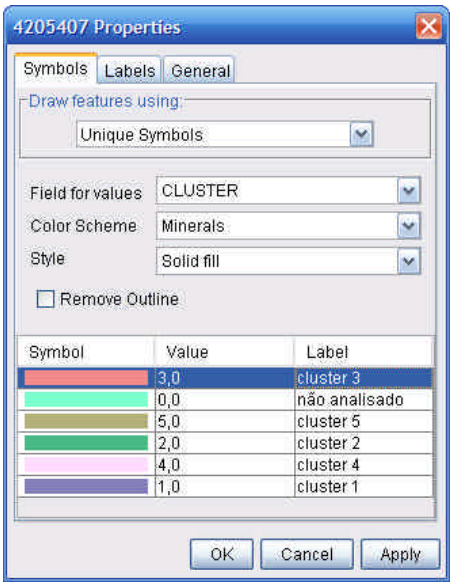

Figura 3-13 - Tela de opções de visualização do cartograma no ArcExplorer

Com as alterações nas opções *Draw features using*, *Field for values*, *Color Scheme* e Label obtivemos a distribuição espacial dos clusters conforme apresentado nas figuras 3.15 e 3.16.

Os setores censitários apresentados na legenda como “não analisado” são os setores censitários que foram removidos das análises por não apresentarem dados de renda e nível de instrução conforme descrito na seção 3.3.

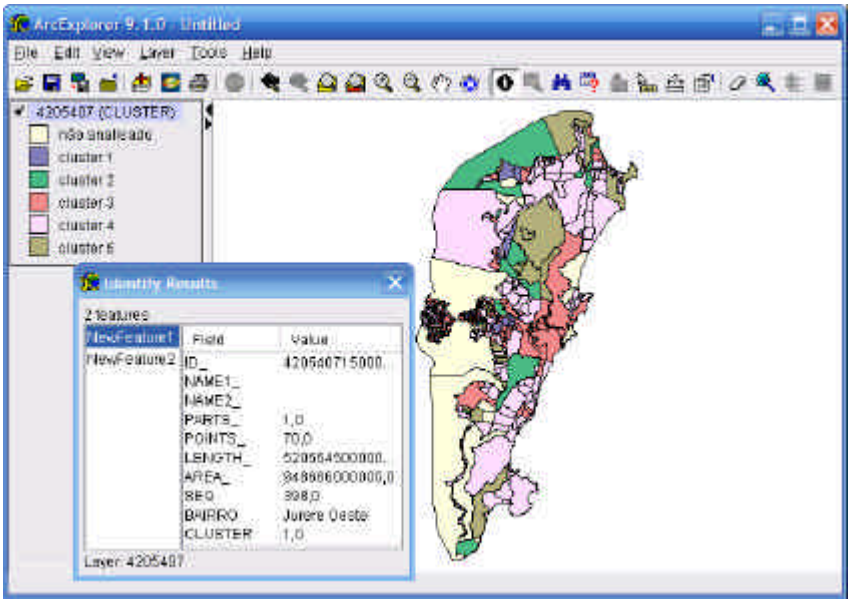

Figura 3-14 – Software ArcExplorer utilizado para a distribuição espacial dos clusters.

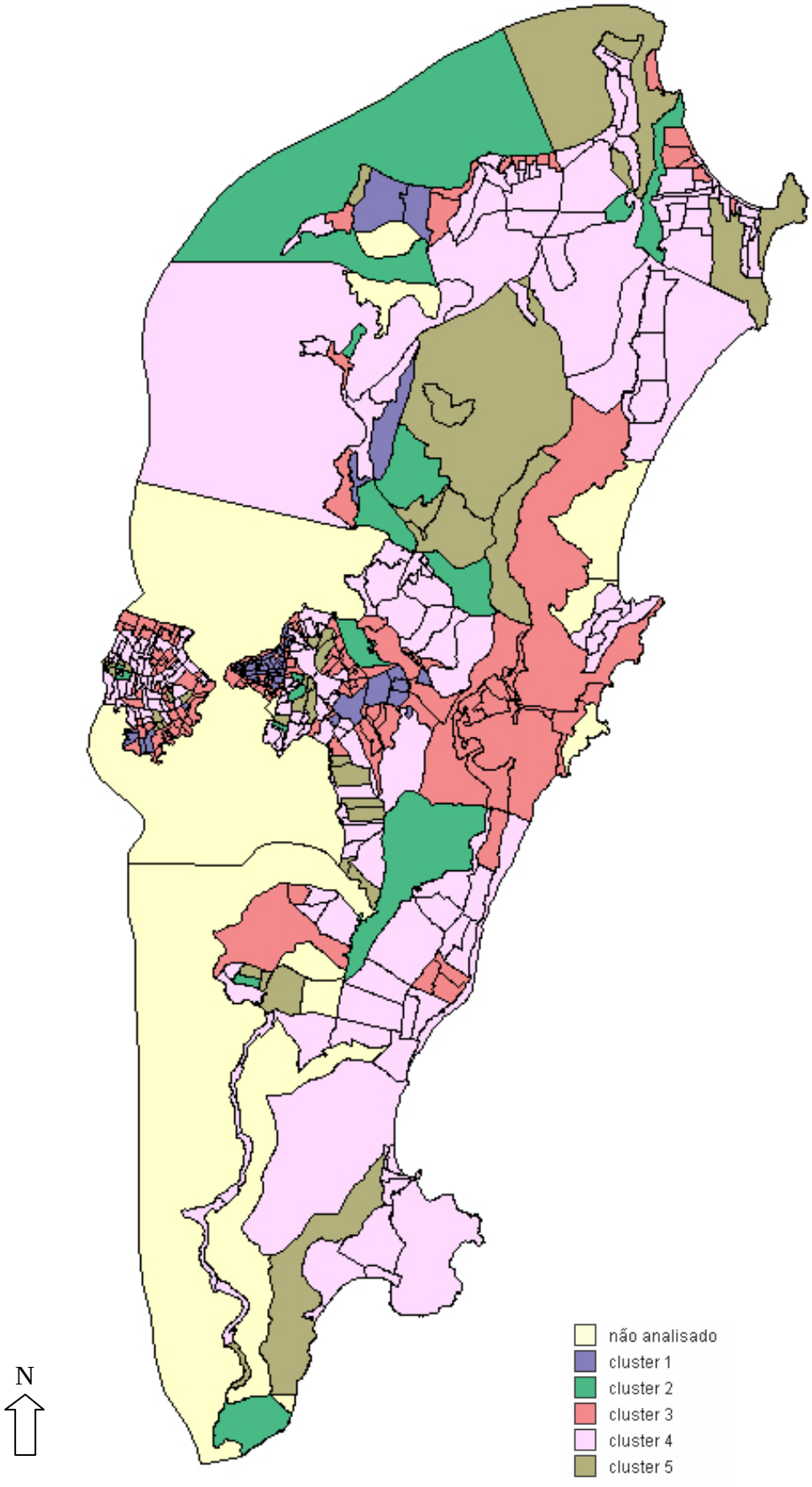

Figura 3-15 - Distribuição espacial dos grupos no município de Florianópolis. Fonte dos dados: IBGE. Elaborado por Rodrigo Benincá Machado

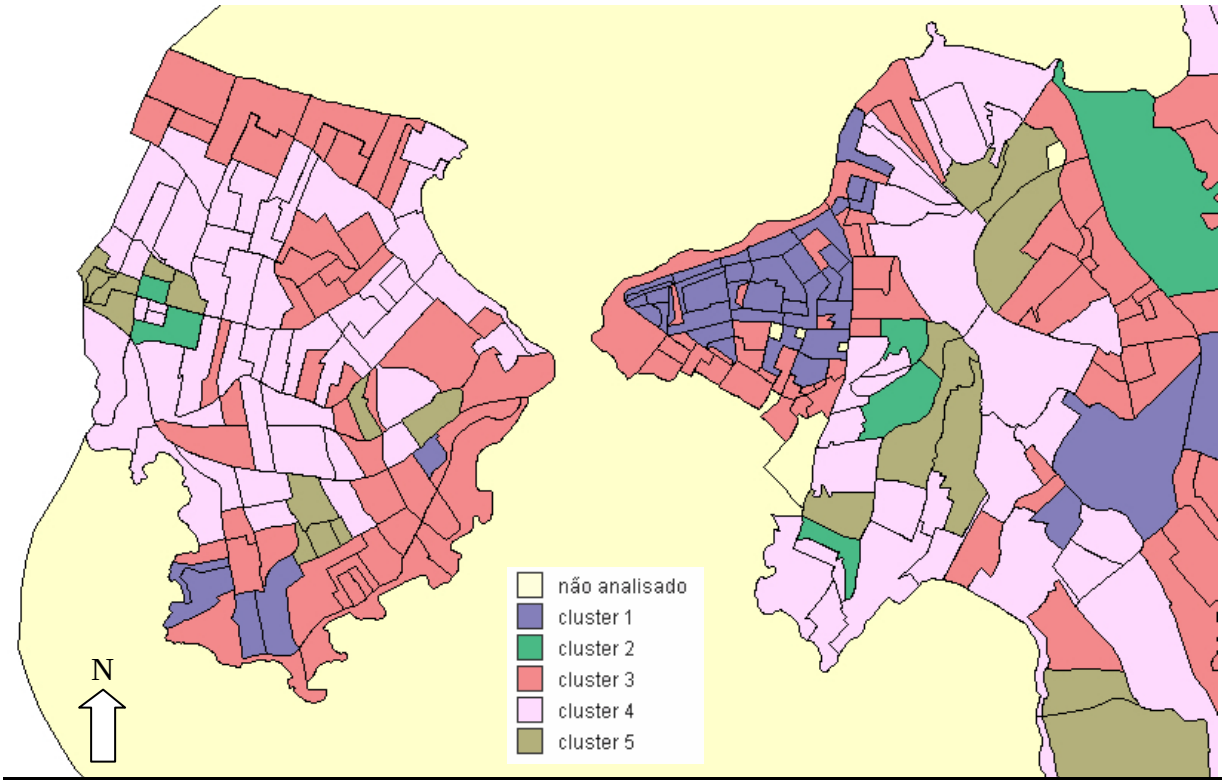

Figura 3-16 - Ampliação da distribuição espacial dos grupos na região central e continental do município de Florianópolis. Fonte dos dados: IBGE. Elaborado por Rodrigo Benincá Machado

Na região central de Florianópolis, conforme figura 3.17, estão destacados os setores censitários que possuem os valores mais elevados de renda média por responsável no município de Florianópolis. Os valores das rendas e média de anos de estudo desses setores podem ser verificados na tabela 3.9.

Figura 3-17 - Distribuição espacial dos setores censitários com os valores mais elevados de renda média por responsável na região central de Florianópolis. Fonte dos dados: IBGE. Elaborado por Rodrigo Benincá Machado

As tabelas 3.15 a 3.19 apresentam a os bairros que constituem cada cluster.

Tabela 3-15 - Bairros e seus setores censitários que constituem o cluster 1.

Bairro	***Quantidade de setores do bairro no cluster***	***Proporção dos setores do bairro no cluster (%)***
Agronômica	4	25
Bom Abrigo	2	100
Centro	31	44,28
Coqueiros	1	5,56
Córrego Grande	1	16,67
Itacorubi	1	6,67
Itaguaçu	2	66,67
Jurere Oeste	2	100
Não existe / não identificado	1	4,55
Saco dos Limões	1	7,14
Santa Mônica	4	80
Trindade	1	5

Tabela 3-16 - Bairros e seus setores censitários que constituem o cluster 2.

Bairro	***Quantidade de setores do bairro no cluster***	***Proporção dos setores do bairro no cluster (%)***
Barra do Sambaqui	1	100
Centro	2	2,86
Itacorubi	1	6,67
José Mendes	1	25
Monte Cristo	2	15,38
Monte Verde	1	14,28
Não existe / não identificado	4	18,18
Saco Grande	2	40
Tapera da Base	1	20
Vargem do Bom Jesus	1	50

Tabela 3-17 - Bairros e seus setores censitários que constituem o cluster 3.

Bairro	***Quantidade de setores do bairro no cluster***	***Proporção dos setores do bairro no cluster (%)***
Abraão	4	50
Agronômica	4	25
Balneário	5	71,42
Base Aérea	1	100
Cacupé	1	100
Campeche Leste	3	100
Canasvieiras	5	38,46
Canto	6	85,71
Canto da Lagoa	1	100
Canto dos Araçás	1	100
Capoeiras	8	32
Carianos	1	33,33
Centro	28	40
Coloninha	1	20
Coqueiros	12	66,67
Córrego Grande	4	66,66
Daniela	1	50
Dunas da Lagoa	1	100
Estreito	3	30
Ingleses Centro	2	33,33
Ingleses Norte	2	100
Ingleses Sul	1	20

Itacorubi	9	60
Itaguaçu	1	33,33
Jardim Atlântico	5	35,71
Jurere Leste	2	66,67
Lagoa	5	100
Não existe / não identificado	3	13,64
Pantanal	3	50
Porto da Lagoa	1	100
Praia Brava	1	100
Praia Mole	1	100
Retiro	1	100
Rio Tavares do Norte	1	100
Saco dos Limões	4	28,57
Sambaqui	1	50
Santa Mônica	1	20
Trindade	15	75

Tabela 3-18 - Bairros e seus setores censitários que constituem o cluster 4.

Bairro	***Quantidade de setores do bairro no cluster***	***Proporção dos setores do bairro no cluster (%)***
Abraão	4	50
Açores	1	100
Agronômica	6	37,5
Alto Ribeirão	1	100
Alto Ribeirão Leste	1	100
Armação	3	100
Autódromo	1	100
Balneário	2	28,58
Barra da Lagoa	5	100
Cachoeira do Bom Jesus	3	100
Cachoeira do Bom Jesus Leste	1	50
Caiacanga	1	100
Campeche Central	1	100
Campeche Norte	2	100
Campeche Sul	2	100
Canasvieiras	8	61,54
Canto	1	14,29
Canto do Lamim	1	100
Capivari	5	100
Capoeiras	15	60
Carianos	2	66,67
Centro	7	10
Coloninha	4	80
Coqueiros	1	5,56
Córrego Grande	1	16,67
Costeira do Pirajubaé	6	54,55
Costeira do Ribeirão	1	100
Daniela	1	50
Estreito	7	70
Ingleses Centro	4	66,67
Ingleses Sul	4	80
Itacorubi	4	26,66
Jardim Atlântico	9	64,29
João Paulo	2	100
José Mendes	3	75
Jurere Leste	1	33,33
Lagoa Pequena	1	100
Moenda	1	100
Monte Cristo	5	38,46

Monte Verde	6	85,72
Morro das Pedras	1	100
Não existe / não identificado	8	36,36
Pantanal	3	50
Pântano do Sul	2	100
Pedrita	1	100
Ponta das Canas	4	100
Recanto dos Açores	1	100
Ressacada	1	100
Ribeirão da Ilha	2	100
Rio das Pacas	1	100
Rio Tavares Central	2	100
Rio Vermelho	3	100
Saco dos Limões	6	42,86
Sambaqui	1	50
Santinho	4	100
Santo Antônio	1	100
Tapera	1	100
Tapera da Base	2	40
Trindade	3	15
Vargem de Fora	1	100
Vargem do Bom Jesus	1	50
Vargem Grande	1	100
Vargem Pequena	1	100

Tabela 3-19 - Bairros e seus setores censitários que constituem o cluster 5.

Bairro	***Quantidade de setores do bairro no cluster***	***Proporção dos setores do bairro no cluster (%)***
Agronômica	2	12,5
Cachoeira do Bom Jesus Leste	1	50
Caieira	1	100
Capoeiras	2	8
Centro	2	2,86
Coqueiros	4	22,22
Costeira do Pirajubaé	5	45,45
Forte	1	100
Lagoinha do Norte	1	100
Monte Cristo	6	46,16
Não existe / não identificado	6	27,27
Pedregal	1	100
Ratones	1	100
Saco dos Limões	3	21,43
Saco Grande	3	60
Tapera da Base	2	40
Trindade	1	5

4 Considerações Finais

4.1 Conclusões

Nesse trabalho de conclusão de curso foram aplicadas as técnicas de *data mining* e um sistema de informações geográficas para obter resultados úteis à tomada de decisões. Inicialmente pretendia-se que esses resultados fossem aproveitados por empresas comercias para a escolha do local mais adequado à sua instalação, como consta no objetivo geral. Entretanto, no desenvolvimento do trabalho, percebeu-se que esses resultados poderiam servir também para tratar de questões sociais.

Neste estudo foram propostos dois objetivos específicos que apresentavam os meios pelos quais o objetivo geral seria alcançado. Para atingir o primeiro objetivo específico foi preciso analisar o processo de descoberta de conhecimento (metodologias, etapas e técnicas) para que fosse obtido o embasamento teórico necessário à sua aplicação.

Ao ser executado o processo de descoberta de conhecimento, este se mostrou como uma ferramenta poderosa se aplicado a dados cuidadosamente trabalhados pois possibilitou através da análise de conglomerados reunir os setores censitários de forma homogênea quanto a renda média e nível de instrução do responsável pelo domicílio. Como resultados dessa análise verificou-se que possuir ensino superior é a variável que mais influencia na formação de grupos seguida pela renda média em salários mínimos do responsável pelo domicílio. E os setores censitários que possuem os valores mais elevados para essas variáveis foram agrupados em apenas dois grupos.

O segundo objetivo específico trata de como distribuir espacialmente no mapa de Florianópolis os grupos encontrados com o uso de um sistema de informações geográficas. O ordenamento espacial permitiu obter uma visão clara da posição geográfica dos *clusters* formados e identificar tanto as áreas que podem ser consideradas mais atraentes comercialmente, como as áreas que necessitam de maior atenção governamental.

4.2 Recomendações e trabalhos futuros

É possível citar algumas sugestões para estudos futuros que poderiam complementar esse trabalho ou mesmo dispor de seus resultados no desenvolvimento de outras pesquisas.

Para complementação e implemento de melhorias no trabalho podem ser sugeridas as seguintes atividades:

- Calcular as densidades demográficas de cada *cluster* encontrado e de cada conjunto isolado de setores (partições dos *clusters*)
- Repetir o processo de descoberta de conhecimento por meio de ferramentas gratuitas, como a linguagem R[3] que permite, dependendo dos pacotes instalados, realizar desde estatísticas simples até execução de algoritmos de *data mining*.
- Repetir a etapa de *data mining* (*cluster analysis*) empregando um método de partição no agrupamento em vez de um método hierárquico e observar as possíveis diferenças encontradas entre os dois métodos.

Os resultados desse trabalho podem motivar as seguintes atividades:

- Utilizar os resultados encontrados para estudar as regiões nas quais poderiam ser oferecidos serviços sociais como classes de alfabetização de adultos ou de programas de alimentação populares como os “Direto do campo” já presentes em Florianópolis.
- Realizar um estudo para encontrar as prováveis razões que levaram a essa segmentação de setores quanto ao nível de instrução e de renda média do responsável pelo domicílio.

[3] Maiores informações sobre a linguagem R pode ser encontradas no site do projeto no endereço http://www.R-project.org.

5 – Referências Bibliográficas

AEDB - Associação Educacional Dom Bosco. Disponível em <http://www.inf.aedb.br/datamining>. Acesso em 08 de março de 2005.

BARBIERI, Carlos – **BI – Business Intelligence, Modelagem e Tecnologia**. Rio de Janeiro: Axcel Books do Brasil Editora, 2001.

CÂMARA, G., MONTEIRO, A. M. V., DRUCK, S., CARVALHO, M. S **– Análise Espacial de Dados Geográficos**. INPE/EMBRAPA/FIOCRUZ/USP. 2000. Disponível em: http://www.dpi.inpe.br/gilberto/livro/analise/index.html. Acessado em 10/05/2005.

CARVALHO, Deborah Ribeiro, et al**. – Ferramenta de Pré e Pós-processamento para data mining**. In: Anais XII SEMINCO, p. 131-140, agosto, 2003.

ELMASRI, Ramez , NAVATHE, Shamkant B. – **Sistemas de Banco de Dados Fundamentos e Aplicações**. 3. ed. Rio de Janeiro: LTC Editora, 2002.

ESRI, http://www.esri.com. Acessado em 01 de maio de 2005.

GUIMARÃES, William Sérgio Azevedo Guimarães – **Data Mining Aplicado ao Serviço Público, Extração de Conhecimento das Ações do Ministério Público Brasileiro**. Dissertação (Mestrado em Ciências da Computação) Universidade Federal de Santa Catarina, Florianópolis, 2000.

IBGE – Instituto Brasileiro de Geografia e Estatística. Disponível em <http://www.ibge.gov.br>. Acessado em 02 de fevereiro de 2005.

IBGEa - **Agregado por Setores Censitários dos Resultados do Universo – 2ª Edição**. CD-ROM.

KAMBER, M., HAN, J. – **Data Mining: Concepts and Techniques**. New York: Editora Morgan Kaufmann Publisher, 2001.

MORAES, André Fabiano de – **Um modelo representativo de conhecimento para aplicação da mineração de dados no cadastro técnico urbano**. Dissertação (Mestrado em Engenharia de Produção) Universidade Federal de Santa Catarina, 2003. Disponível em <http://aspro02.npd.ufsc.br/arquivos/195000/197100/18_197129.htm>

OGLIARI, P. J. **Disciplina INE5644 – Data Mining**. Disponível em <http://www.inf.ufsc.br/~ogliari/cursodedatamining>. Acesso em 9 de novembro de 2004.

PACHECO, Juliano Anderson – **Métodos Estatísticos Espaciais no Planejamento da Prestação de Serviços**. Dissertação (Mestrado em Ciências da Computação) Universidade Federal de Santa Catarina, Florianópolis, 2005.

PRASS, Fernando Sarturi – **Estudo comparativo entre algoritmos de análise de agrupamentos em data mining**. Dissertação (Mestrado em Ciências da Computação) Universidade Federal de Santa Catarina, Florianópolis, 2004.

Anexo I – ARTIGO

Uso de Data Mining e Sistemas de Informações Geográficas no Apoio a Tomada de Decisões

Rodrigo Benincá Machado
rodrigobeninca@yahoo.com.br

Resumo

Esse trabalho de conclusão de curso objetivou desenvolver um documento que contivesse o resultado da distribuição espacial de setores censitários do município de Florianópolis semelhantes quanto ao nível de instrução e renda média do responsável pelo domicílio. O alcance do objetivo acima se deu por meio do estudo e da aplicação do processo de descoberta de conhecimento (KDD) e do emprego de um sistema de informações geográficas. A aplicação de todas as etapas do processo de descoberta de conhecimento se deu com o uso da técnica de *cluster analysis* que possibilitou a formação dos grupos de setores censitários semelhantes. A distribuição espacial dos grupos encontrados pelo KDD foi obtida pelo emprego de um sistema de informações geográficas.

Como resultado desse estudo, podemos concluir que a aplicação do processo de descoberta de conhecimento permitiu reunir os setores censitários de forma homogênea quanto à renda média e nível de instrução do responsável pelo domicílio. Além disso, o uso do sistema de informações geográficas permitiu obter uma visão clara da posição espacial dos *clusters* formados e identificar tanto as áreas que podem ser consideradas mais atraentes comercialmente, como as áreas que necessitam de maior atenção governamental.

Palavras-chaves: análise de conglomerados, mineração de dados, processo de descoberta de conhecimento, sistemas de informações geográficas.

1.Introdução

A evolução tecnológica que ocorre nas bases materiais de nossa sociedade disponibiliza às instituições privadas e públicas a capacidade de produzir e armazenar grandes quantidades de dados referentes aos seus respectivos negócios. Esses dados podem ser utilizados para que as transações nessas empresas se tornem mais eficazes.

Nas duas últimas décadas houve um grande crescimento na quantidade de dados produzidos e armazenados em meio eletrônico. O valor desses dados está relacionado à capacidade de extrair informações úteis ao suporte de decisões operacionais, táticas e estratégicas. É possível que existam ainda, padrões ou tendências úteis que, se descobertos, podem ser empregados, por exemplo, para auxiliar em um processo de decisão em uma empresa (AEDB, 2005). Pode ser citada como exemplo de fonte de dados para análises a que se encontra nas bases de dados do censo demográfico realizado pelo IBGE.

A cada 10 anos o IBGE (Instituto Brasileiro de Geografia e Estatística) realiza o censo da população pesquisando de maneira completa variáveis demográficas, níveis de nupcialidade e fecundidade, condições de trabalho, educação e renda e características dos domicílios. O censo é efetivado em dois estágios, dos quais o primeiro se dá por meio da aplicação de um conjunto de questões básicas a toda a população, e o segundo, por meio da abordagem de uma amostra através de um questionário mais abrangente. Em todo o território nacional foram selecionados 5.304.711 domicílios para responder ao questionário da amostra, o que significou uma fração amostral da ordem de 11,7%.

É possível se extrair informações proveitosas a partir de uma base de dados como a do IBGE através de técnicas de *data mining*. O *data mining* é parte de um processo maior, chamado de processo de descoberta de conhecimento, que tem o objetivo de otimizar e automatizar a descrição das tendências e padrões presentes nos dados (OGLIARI, 2004). A informação e conhecimento obtidos nessa operação podem ser aproveitados tanto em aplicações de *business management*, controle de produção e análise de mercado como em aplicações em projetos de engenharia e exploração científica.

Esse trabalho de conclusão de curso faz uso da técnica de *data mining* conhecida como *cluster analysis* e de um sistema de informações geográficas para atingir os objetivos que serão apresentados a seguir.

O tópico 2 desse artigo expõe os objetivos do trabalho, enquanto o terceiro tópico faz um breve relato da metodologia empregada.O tópico 4 apresenta os principais resultados. No tópico 5 estão presentes as considerações finais e no 6 a bibliografia utilizada.

2.Objetivos

O objetivo geral desse trabalho é desenvolver um documento que contenha o resultado da distribuição espacial de setores censitários que sejam semelhantes quanto ao nível de instrução e renda média do responsável pelo domicílio. Estes resultados servirão como artefato no processo de escolha do local mais adequado para instalação de empresas comerciais.

Os objetivos específicos são:

- Aplicar o processo de descoberta de conhecimento para definir os grupos de setores censitários semelhantes quanto ao nível de instrução e renda média do responsável pelo domicílio.
- Empregar um sistema de informações geográficas para distribuir espacialmente os grupos de setores censitários formados no processo de descoberta de conhecimento.

3.Metodologia utilizada

A metodologia de desenvolvimento desse trabalho pode ser dividida em três partes: a fundamentação teórica, a aplicação do processo de descoberta de conhecimento para formação dos grupos e a distribuição espacial dos grupos formados por meio de um sistema de informações geográficas.

A primeira seção, fundamentação teórica introduz os conceitos de cada etapa do processo de descoberta de conhecimento, enfatizando a técnica de *cluster analysis* e os conceitos de sistema de informações geográficas.

Após a conclusão da fundamentação teórica foi dado início às atividades do processo de descoberta de conhecimento por meio da análise da base de dados e da seleção, preparação e transformação das variáveis a serem consideradas na etapa de *data mining*. Estas variáveis, selecionadas na base de dados do censo demográfico, devem estar relacionadas ao nível de instrução e à renda dos responsáveis por domicílios no município de Florianópolis.

Para finalizar as atividades aqui propostas, fez-se necessário que os grupos encontrados no processo de descoberta de conhecimento fossem distribuídos espacialmente no mapa do município de Florianópolis com uso de um sistema de informações geográficas.

A execução das três fases descritas acima permitiram o alcance dos objetivos expostos na seção 1.1 desse documento.

4.Principais resultados

O processo de descoberta de conhecimento é um processo composto por diversas etapas, envolvendo metodologias e técnicas de data mining. O seu objetivo é o de "otimizar e automatizar o processo de descrição das tendências e dos padrões contidos nesse processo, potencialmente úteis e interpretáveis" (OGLIARI, 2004). Para PRASS (2004), o processo de descoberta de conhecimento compreende todo o ciclo que os dados percorrem até virar informação. O processo de descoberta de conhecimento é formado por duas partes principais, a preparação dos dados e o *data mining*. A preparação dos dados divide-se nas etapas de seleção dos dados, pré-processamento dos dados e transformação e integração dos dados.

A seleção dos dados é primeira fase do processo de descoberta de conhecimento onde é selecionado um conjunto de dados contendo as variáveis que serão utilizadas em análises nas fases posteriores. Essas variáveis são selecionadas de acordo com o objetivo em questão e deve ser realizada com auxílio de um especialista no assunto (OGLIARI, 2004).

No pré-processamento dos dados são identificados e corrigidos problemas presentes nos dados selecionados. Esses problemas podem ser dados inconsistentes, dados faltantes (*missing*) ou valores discrepantes (*outliers*) (KAMBER, 2001) e geralmente são detectados através da análise exploratória dos dados. A análise exploratória também tem o objetivo de estudar / conhecer o conjunto de dados que será utilizado na finalização da preparação dos dados bem como testar suposições do modelo.

Nessa etapa do KDD pode-se destacar os resultados obtidos na análise exploratória dos dados onde observou-se que a média de anos de estudos dos setores censitários é correspondente a primeira série do ensino médio e a variável de renda média em salários mínimos destaca-se por apresentar valor máximo em torno de 87 salários

vezes maior que seu valor mínimo (1,28). Analisando-se agora as variáveis de nível de instrução observa-se que, em média, os setores censitários possuem em torno de 28% de seus responsáveis por domicílio com ensino superior. Nota-se também que a distribuição das variáveis de proporção de responsáveis por domicílio sem instrução e de responsáveis com mestrado ou doutorado são semelhantes, ou seja, da mesma maneira que existem setores censitários com, aproximadamente, 30% de seus responsáveis sem instrução, há a presença de setores com aproximadamente a mesma proporção de responsáveis com mestrado ou doutorado.

Finaliza-se a fase de preparação dos dados com a etapa de integração e transformação dos dados onde os dados são formatados e armazenados adequadamente para que os algoritmos de *data mining* possam ser aplicados.

A fase de mineração de dados é composta basicamente pela etapa de data mining e a avaliação dos padrões gerados.

A tarefa de *data mining* utilizada nesse trabalho corresponde a tarefa de agrupamento que visa agrupar os registros mais semelhantes entre si de acordo com as variáveis selecionadas para a análise.

A tarefa de agrupamento segue a metodologia não supervisionada de conhecimento e a técnica que a implementa é a *cluster analysis* (ou análise de conglomerados).

O software utilizado para a mineração de dados foi o *The SAS System for Windows* versão 8.2, mais especificamente seu módulo *SAS Enterprise Miner*.

Como resultado da análise de conglomerados obteve-se a formação de cinco grupos de setores censitários distintos quanto a renda média em salários mínimos e do nível de instrução dos responsáveis por domicílios. A variável com maior grau de importância na formação dos *clusters* foi a variável de proporção de responsáveis com nível superior no setor seguida pela variável que representa a renda média em salários mínimos no setor.

Nesse artigo será dado destaque ao grupo que representa as maiores médias de renda e escolaridade (*cluster* 1) e ao grupo com as menores médias para as mesmas variáveis (*cluster* 5).

A renda média em salários mínimos do grupo 1 chama a atenção por ser duas vezes e meia superior a média apresentada no município e sua média de anos de estudo ser 1,4 vezes maior que a média municipal. O grupo 5, ao contrário, apresenta uma renda 3,5 vezes inferior a de Florianópolis e os anos de estudo pouco superior a metade.

O *cluster* 1 possui quase três quartos de seus responsáveis por domicílio com ensino superior ou mestrado/doutorado, contrastando com o *cluster* 5 onde apenas 4,94% de seus responsáveis possuem ensino superior ou mestrado/doutorado e, 82,56% possuem até a oitava série do ensino fundamental concluída.

Com a formação e avaliação dos grupos formados na fase de mineração de dados foi necessário realizar a distribuição espacial desses setores censitários no mapa do município de Florianópolis utilizando um sistema de informações geográficas.

Os sistemas de informações geográficas podem ser definidos como:

> *Sistemas que realizam o tratamento computacional de dados geográficos e manipulam a geometria e os atributos dos dados que estão geo-referenciados, ou seja, localizados na superfície terrestre e representados numa projeção cartográfica (Câmara et al., 2000).*

> Um sistema de suporte à decisão que integra dados referenciados espacialmente num ambiente de respostas a problemas (Cowen in CÂMARA et. al (2000).

O sistema de informações geográficas foi utilizado para criar um cartograma utilizando o atributo correspondente ao cluster que cada setor censitário está associado.

5.Considerações finais

Nesse trabalho de conclusão de curso foram aplicadas as técnicas de *data mining* e um sistema de informações geográficas para obter resultados úteis à tomada de decisões. Inicialmente pretendia-se que esses resultados fossem aproveitados por empresas comercias para a escolha do local mais adequado à sua instalação, como consta no objetivo geral. Entretanto, no desenvolvimento do trabalho, percebeu-se que esses resultados poderiam servir também para tratar de questões sociais.

Neste estudo foram propostos dois objetivos específicos que apresentavam os meios pelos quais o objetivo geral seria alcançado. Para atingir o primeiro objetivo

descoberta de conhecimento (metodologias, etapas e técnicas) para que fosse obtido o embasamento teórico necessário à sua aplicação.

Ao ser executado o processo de descoberta de conhecimento, este se mostrou como uma ferramenta poderosa se aplicado a dados cuidadosamente trabalhados pois possibilitou através da análise de conglomerados reunir os setores censitários de forma homogênea quanto a renda média e nível de instrução do responsável pelo domicílio. Como resultados dessa análise verificou-se que possuir ensino superior é a variável que mais influencia na formação de grupos seguida pela renda média em salários mínimos do responsável pelo domicílio. E os setores censitários que possuem os valores mais elevados para essas variáveis foram agrupados em apenas dois grupos.

O segundo objetivo específico trata de como distribuir espacialmente no mapa de Florianópolis os grupos encontrados com o uso de um sistema de informações geográficas. O ordenamento espacial permitiu obter uma visão clara da posição geográfica dos *clusters* formados e identificar tanto as áreas que podem ser consideradas mais atraentes comercialmente, como as áreas que necessitam de maior atenção governamental.

6.Bibliografia

AEDB - **Associação Educacional Dom Bosco**. Disponível em <http://www.inf.aedb.br/datamining>. Acesso em 08 de março de 2005.

CÂMARA, G., MONTEIRO, A. M. V., DRUCK, S., CARVALHO, M. S **– Análise Espacial de Dados Geográficos**. 2000. Disponível em: http://www.dpi.inpe.br/gilberto/livro/analise/index.html. Acessado em 10/05/2005.

KAMBER, M., HAN, J. – **Data Mining: Concepts and Techniques**. New York: Editora Morgan Kaufmann Publisher, 2001.

OGLIARI, P. J. **Disciplina INE5644 – Data Mining**. Disponível em <http://www.inf.ufsc.br/~ogliari/cursodedatamining>. Acesso em 9 de novembro de 2004.

PRASS, Fernando Sarturi – **Estudo comparativo entre algoritmos de análise de agrupamentos em data mining**. Dissertação (Mestrado em Ciências da Computação) Universidade Federal de Santa Catarina, Florianópolis, 2004.