UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de São José do Rio Preto

Camila Alves de Medeiros

Extração de conhecimento em bases de dados espaciais: Algoritmo CHSMST^{+}

São José do Rio Preto
2014

Camila Alves de Medeiros

Extração de conhecimento em bases de dados espaciais: Algoritmo CHSMST^{+}

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, Área de Concentração – Computação Aplicada, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Orientador: Prof. Dr. Carlos Roberto Valêncio

São José do Rio Preto
2014

Medeiros, Camila Alves de.
Extração de conhecimento em bases de dados espaciais : algoritmo CHSMST^{+} / Camila Alves de Medeiros. -- São José do Rio Preto, 2013
121 f. : il., tabs.

Orientador: Carlos Roberto Valêncio
Dissertação (mestrado) – Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências, Letras e Ciências Exatas

1. Computação. 2. Sistemas de informação geográfica. 3. Sistemas de dados espaciais. 4. Banco de dados. 5. Análise espacial (Estatística) I. Valêncio, Carlos Roberto. II. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. III. Título.

CDU – 518.72:91

Ficha catalográfica elaborada pela Biblioteca do IBILCE
UNESP - Câmpus de São José do Rio Preto

Camila Alves de Medeiros

Extração de conhecimento em bases de dados espaciais:

Algoritmo CHSMST^{+}

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação, junto ao Programa de Pós-Graduação em Ciência da Computação, Área de Concentração – Computação Aplicada, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de São José do Rio Preto.

Banca Examinadora

Prof. Dr. Carlos Roberto Valêncio
UNESP – São José do Rio Preto
Orientador

Prof. Dr. Adriano Mauro Cansian
UNESP – São José do Rio Preto

Profa. Dra. Elaine Parros Machado de Sousa
USP – São Carlos

São José do Rio Preto
28 de março de 2014

AGRADECIMENTOS

Agradeço primeiramente a Deus, que sempre me deu forças e que tornou possível o desafio de fazer esse curso, mesmo diante das adversidades.

Aos meus pais, Angela e Hélio, que torceram e se sacrificaram, fazendo além do que estava ao alcance deles para que eu pudesse estar aqui.

Ao meu namorado, Marcos, pelo amor e carinho, pela paciência e apoio para que eu prosseguisse com os meus objetivos.

Ao meu orientador Prof. Dr. Carlos Roberto Valêncio, por acreditar em meu potencial e me incentivar a prosseguir nos meus estudos.

A toda a minha família e amigos que sempre torceram por mim.

Aos amigos do GBD, Matheus, Thatiane, Paulo, Diogo, Guilherme que sempre me ajudaram.

A todos os docentes e colegas, que contribuíram com o meu crescimento e aprendizado.

À minha família.

A verdadeira sabedoria consiste em saber como aumentar o bem-estar do mundo.
Benjamin Franklin

RESUMO

O desenvolvimento de tecnologias de coleta de informações espaciais resultou no armazenamento de um grande volume de dados, que devido à complexidade dos dados e dos respectivos relacionamentos torna-se impraticável a aplicação de técnicas tradicionais para prospecção em bases de dados espaciais. Nesse sentido, diversos algoritmos vêm sendo propostos, sendo que os algoritmos de agrupamento de dados espaciais são os que mais se destacam devido a sua alta aplicabilidade em diversas áreas. No entanto, tais algoritmos ainda necessitam superar vários desafios para que encontrem resultados satisfatórios em tempo hábil. Com o propósito de contribuir neste sentido, neste trabalho é apresentado um novo algoritmo, denominado CHSMST^{+}, que realiza o agrupamento de dados considerando tanto a distância quanto a similaridade, o que possibilita correlacionar atributos espaciais e não espaciais. Tais tarefas são realizadas sem parâmetros de entrada e interação com usuário, o que elimina a dependência da interpretação do usuário para geração dos agrupamentos, bem como possibilita a obtenção de agrupamentos mais eficientes uma vez que os cálculos realizados pelo algoritmo são mais precisos que uma análise visual dos mesmos. Além destas técnicas, é utilizada a abordagem *multithreading*, que possibilitou uma redução média de 38,52% no tempo de processamento. O algoritmo CHSMST^{+} foi aplicado em bases de dados espaciais da área da saúde e meio ambiente, mostrando a capacidade de utilizá-lo em diferentes contextos, o que torna ainda mais relevante o trabalho realizado.

Palavras-chave: prospecção de dados espaciais; agrupamento de dados espaciais*; Clustering algorithm based on Hyper Surface and Minimum Spanning Tree* - CHSMST.

ABSTRACT

The development of technologies for collecting spatial information has resulted in a large volume of stored data, which makes inappropriate the use of conventional data mining techniques for knowledge extraction in spatial databases, due to the high complexity of these data and its relationships. Therefore, several algorithms have been proposed, and the spatial clustering ones stand out due to their high applicability in many fields. However, these algorithms still need to overcome many challenges to reach satisfactory results in a timely manner. In this work, we present a new algorithm, namely CHSMST^{+}, which works with spatial clustering considering both distance and similarity, allowing to correlate spatial and non-spatial attributes. These tasks are performed without input parameters and user interaction, eliminating the dependence of the user interpretation for cluster generation and enabling the achievement of cluster in a more efficient way, since the calculations performed by the algorithm are more accurate than visual analysis of them. Together with these techniques, we use a multithreading approach, which allowed an average reduction of 38,52% in processing time. The CHSMST^{+} algorithm was applied in spatial databases of health and environment, showing the ability to apply it in different contexts, which makes this work even more relevant.

Keywords: spatial data mining; spatial clustering; clustering algorithm based on Hyper Surface and Minimum Spanning Tree - CHSMST.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADACLUS	*ADAptive spatial CLUStering algorithm based on delaunay triangulation*
AGNES	*AGglomerativeNESting*
AUTOCLUST	*Automatic Clustering via Boundary Extraction for Mining Massive Point-DataSets*
AMOEBA	*A Multidirectional Optimal Ecotope-Based Algorithm*
API	*Application Program Interface*
BIRCH	*Balanced Interative Reducing and Clustering using Hierarchies*
CEREST	Centro de Referência de Saúde do Trabalhador
CHS	*Clustering based on Hyper Surface*
CHSMST	*Clustering algorithm based on Hyper Surface and Minimum Spanning Tree*
CLARA	*ClusteringLargeApplications*
CLARANS	*Clustering Large Applications based on Randomized Search*
CLIQUE	*Clustering In Quest*
CURE	*Clustering Using Representatives*
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
DENCLUE	*DensityBasedClustering*
DIANA	*DIvisive ANAlysis*
EPS	*episolon*
FADE	Ferramenta para Agrupamento de Dados Espaciais
FEHIDRO	Fundo Estadual de Recursos Hídricos
GFS	*Google File System*
GBD	Grupo de Banco de Dados
GPS	*Global Positioning System*
HDFS	*Hadoop Distributed File System*
IPTU	Imposto Predial e Territorial Urbano
KDD	*Knowledge Discovery in Databases*

KDSD	*Knowledge Discovery in Spatial Databases*
MinPts	Número mínimo de pontos
MST	*Minimum Spanning Tree*
OPTICS	*Ordering Points to Identify the Clustering Structure*
PaaS	*Plataform-as-a-Service*
PVM	*Parallel Virtual Machine*
SGBDs	Sistemas Gerenciadores de Banco de Dados
SGBDR	Sistema Gerenciador de Banco de Dados Relacionais
GRH	Sistema de Gerenciamento sobre Recursos Hídricos e Ambientais
SIG	Sistema de Informação Geográfica
SIVAT	Sistema de Informação e Vigilância de Acidentes do Trabalho
SQL	*Structured Query Language*
STING	*Statical Information Grid-based method*

SUMÁRIO

CAPÍTULO 1
Introdução

1.1 Considerações Iniciais

O imenso volume de dados armazenados, em diversos domínios, inviabiliza a análise dos mesmos de forma manual. Para possibilitar a extração de conhecimento, surgiu a prospecção de dados, que consiste em um conjunto de técnicas automáticas utilizadas para descobrir padrões e relações em grandes bases de dados em que tal tarefa não é trivial ao ser humano (HAN; KAMBER, 2011).

Com o desenvolvimento de tecnologias de coleta de informações espaciais, um novo tipo de dado passou a ser registrado, sendo que a complexidade desses dados e de seus relacionamentos torna inviável o uso de métodos convencionais de prospecção de dados (MILLER; HAN, 2009), o que motivou o desenvolvimento do processo de prospecção de dados espaciais com o objetivo de possibilitar a extração de conhecimento em base de dados espaciais (JIN; MIAO, 2010). Para isso, diversas técnicas foram desenvolvidas, sendo destaque a técnica de agrupamento de dados espaciais, uma vez que pode ser empregada em diversos domínios de aplicações e é um método não supervisionado, visto que é capaz de realizar a prospecção de dados espaciais sem necessitar de uma predefinição dos agrupamentos a serem gerados (PATTABIRAMAN et al., 2009, DENG et al., 2011).

1.2 Motivação e Escopo

Muitos pesquisadores têm investido esforços para o desenvolvimento de novos métodos e algoritmos de agrupamento de dados espaciais com o objetivo de extrair conhecimento útil das bases de dados espaciais de forma mais eficiente e eficaz. No entanto, devido à complexidade na manipulação e análise dos dados espaciais, vários algoritmos apresentam dificuldades em relação a:

- **Agrupamentos de formatos variados**[1] **-** Muitos algoritmos não são capazes de detectar agrupamentos com formatos variados. Outros conseguem detectar tais agrupamentos, mas não são eficazes quando aplicados em conjunto de dados cujos agrupamentos não estão bem distribuídos (DENG et al., 2011);
- **Agrupamentos com densidade variada:** A maioria dos algoritmos de agrupamento de dados espaciais não é escalável quando aplicado em conjuntos de dados com grande variedade de densidade (DENG et al., 2011);
- **Sensibilidade à ordem de entrada dos dados:** Os resultados dos algoritmos de agrupamento podem ser influenciados pela ordem de entrada dos dados (XIAO-PING; ZHENG-YUAN; JAIN-HUA, 2010);
- **Parâmetros de entrada –** A maioria dos algoritmos requer que o usuário defina parâmetros que influenciarão na geração dos agrupamentos, como por exemplo, número de agrupamentos, número mínimo de vizinhos a serem analisados, número mínimo de pontos em cada agrupamento e etc;
- **Presença de ruídos** – Alguns algoritmos, ao realizar o agrupamento de dados espaciais possibilitam que determinados objetos não pertençam a nenhum agrupamento, distribuindo-se de forma aleatória pelo conjunto de dados, ou realizem a conexão de dois agrupamentos separados (DENG et al., 2011);
- **Presença de pontos discrepantes** – No conjunto de dados analisado, podem existir objetos cujos atributos não espaciais sejam muito diferentes em relação aos demais objetos (LIU; LU; CHEN, 2010);
- **Problemas com alta dimensionalidade de dados:** A maioria dos algoritmos não é eficiente para altas dimensões de dados (XIAO-PING; ZHENG-YUAN; JAIN-HUA, 2010).

Por meio de estudo comparativo com vários algoritmos de agrupamentos de dados espaciais verificou-se que nenhum desses algoritmos consegue superar todas as dificuldades descritas. Outro problema detectado, é que tais algoritmos frequentemente apresentam dificuldades relacionadas ao armazenamento e processamento devido à complexidade dos dados espaciais e de seus relacionamentos, o que dificulta a correlação de padrões espaciais (MENNIS; GUO, 2009). Considerando estes elementos, encontrou-se a motivação para inves-

[1]Alguns algoritmos são capazes gerar apenas agrupamentos de formatos esféricos, sendo complicado gerar agrupamentos de outros tipos de formatos tais como: côncavos, com furo, agrupamento dentro de agrupamento, em formato caracol, etc. (SHEIKHOLESLAMI; CHATTERJEE; ZHANG, 1998).

tir esforços com o propósito de desenvolver um algoritmo mais eficiente e eficaz, que possa superar tais dificuldades.

1.3 Objetivos e Metodologia

O objetivo deste trabalho é contribuir com a área de prospecção de dados espaciais por meio de proposição de um algoritmo de agrupamento de dados espaciais. A partir do estudo comparativo realizado, foi verificado que o algoritmo *Clustering algorithm based on Hyper-Surface and Minimum Spanning Tree* - CHSMST (HE; ZHAO; SHI, 2011) é capaz de superar a maior parte das dificuldades encontradas nos algoritmos de agrupamento de dados espaciais: geração de agrupamentos com formatos e densidades variadas, robusto a ruído e insensibilidade à ordem de entrada dos dados, além de ser recente. Diante disso, o referido algoritmo foi selecionado para ser a base deste trabalho.

Na comparação com vários algoritmos, o CHSMST apresentou como características negativas o fato de não considerar a similaridade na geração dos agrupamentos e a exigência de parâmetros de entrada. Além disso, durante a análise do algoritmo foi verificado que o mesmo exige uma iteração do usuário para definir quais agrupamentos devem ser gerados. Esta também é uma característica negativa, uma vez que o ideal é que os algoritmos de prospecção de dados espaciais necessitem do mínimo de interação possível para revelar padrões úteis e inéditos. Neste contexto, este trabalho tem o propósito de atacar os problemas encontrados no algoritmo CHSMST e reduzir o tempo de processamento do algoritmo por meio da técnica de *multithreading*, sendo que com as modificações realizadas, surge um novo algoritmo, denominado CHSMST^{+}.

Este trabalho também contempla, como um subproduto, o desenvolvimento de uma ferramenta para suporte de agrupamento de dados espaciais, que possibilita a aplicação do algoritmo desenvolvido em bases de dados de diferentes áreas: saúde, meio ambiente, entre outras. Com isso, tem-se mais uma contribuição, visto que, normalmente, as ferramentas de prospecção de dados espaciais são desenvolvidas para atender a um domínio específico, o que limita o conhecimento que pode ser descoberto.

Com o propósito de validar a aplicabilidade do algoritmo desenvolvido em diferentes contextos, o algoritmo CHSMST^{+} foi aplicado em uma base de dados relacionada a acidentes de trabalho que contempla mais de 20 mil registros georreferenciados e em uma base de dados que armazena dados relacionados a pontos de captação de recursos hídricos e respectivos usuários, com 699 pontos georreferenciados. Além disso, para validar o desempenho do algorit-

mo, o mesmo foi aplicado em uma base de dados do *Academy of Natural Sciences*, que contempla dados georreferenciados sobre espécies de animais.

1.4 Organização da Monografia

Nesta seção é apresentada a estrutura deste trabalho.

- **Capítulo 2 – Levantamento Bibliográfico:** contêm os principais conceitos relacionados ao processo *Knowledge Discovery in Databases* - KDD, dados espaciais e prospecção de dados espaciais.
- **Capítulo 3 – Agrupamento de Dados Espaciais:** aborda os principais conceitos e algoritmos de agrupamento de dados espaciais.
- **Capítulo 4 – Extração de conhecimento em bases de dados espaciais: Algoritmo CHSMST^{+}:** apresenta o trabalho desenvolvido.
- **Capítulo 5 – Experimentos e Resultados:** contempla os experimentos realizados em bases de dados reais para verificar a aplicabilidade do algoritmo em diferentes contextos, a eficácia da estratégia de similaridade, comparativo do algoritmo com outro bem conceituado e análise do desempenho do algoritmo desenvolvido com e sem a técnica de *multithreading*.
- **Capítulo 6 – Conclusões:** apresenta a discussão final sobre o trabalho desenvolvido e as sugestões de trabalhos futuros.

CAPÍTULO 2
Levantamento Bibliográfico

2.1 Considerações Iniciais

O desenvolvimento de tecnologias de coleta de informações espaciais, tais como: satélites, sistemas de monitoramento remoto e *Global Positioning System* - GPS, aliado à incorporação de tais tecnologias em dispositivos móveis, veículos e à Internet, resultaram no armazenamento de uma grande quantidade de dados espaciais (YUESHUN; WEI, 2009, MILLER; HAN, 2009).

No entanto, a complexidade desses dados e respectivos relacionamentos tornam as estruturas convencionais de armazenamento e extração de conhecimento inadequadas para serem aplicadas em bases de dados espaciais (MILLER; HAN, 2009, SHEKHAR; ZHANG; HUANG, 2003). Assim, surge o processo de prospecção de dados espaciais com o objetivo de possibilitar a extração de conhecimento em base de dados espaciais.

Neste capítulo são apresentados os principais conceitos relacionados ao processo *Knowledge Discovery in Databases* - KDD, dados espaciais e prospecção de dados espaciais.

2.2 O Processo KDD

O KDD pode ser entendido como um processo não trivial de extração de conhecimento útil a partir de uma base de dados. O processo é composto das etapas: pré-processamento, prospecção de dados e pós-processamento (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, PIATETSKY-SHAPIRO, 2012). Na Figura 1 o processo KDD é ilustrado.

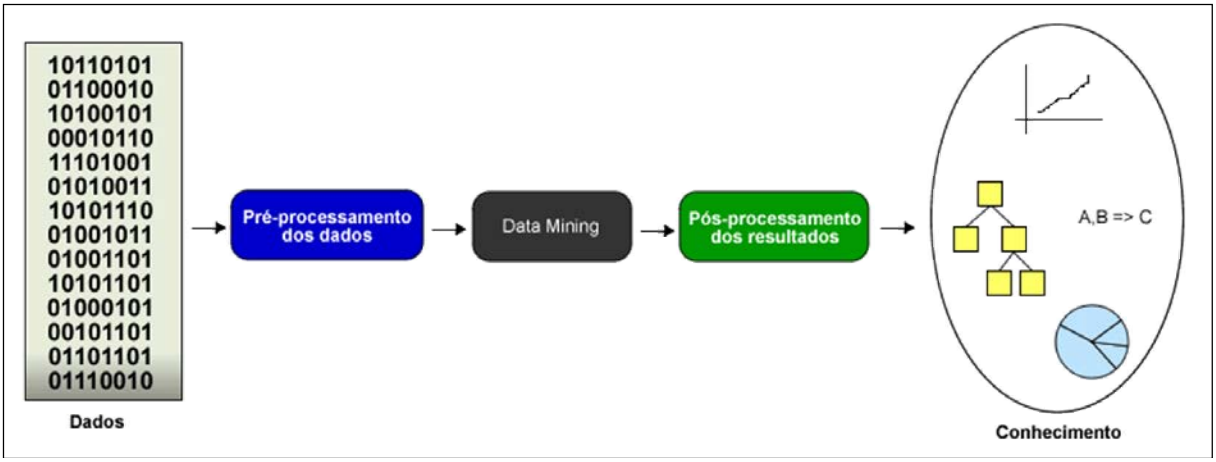

Figura 1: Processo KDD (CARREIRA, 2008, PIATETSKY-SHAPIRO, 2012).

2.2.1 Pré-processamento dos dados

Frequentemente as organizações têm a necessidade de reunir dados provenientes de fontes distintas para suprir uma necessidade de um negócio, suportar a elaboração de estratégias governamentais, bem como agregar dados de diversas pesquisas científicas para alcançar um objetivo comum (JARKE et al., 2003). No entanto, se as fontes de dados foram construídas de forma independentemente uma das outras, as bases de dados serão heterogêneas tanto em nível de esquema quanto em nível de semântica, o que torna a integração das mesmas uma tarefa complexa (JAVED; NAWAZ, 2010). Para contornar este problema, o processo KDD prevê a integração dos dados, uma vez que esta permite uma visão global sobre diversas fontes de dados, o que possibilita a recuperação de informações relevantes a um determinado contexto e tornam homogêneos os dados provenientes de fontes distintas (HALEVY; RAJARAMAN; ORDILLE, 2006).

Em repositórios de dados é comum encontrar inconsistências e problemas causados por diversos fatores, tais como: esquemas mal projetados, alimentação da base com dados incorretos e anomalias geradas durante o processo de integração de bases heterogêneas (ELMAGARMID; IPEIROTIS; VERYKIOS, 2007). Para evitar que tais problemas comprometam a qualidade dos dados e acarretem em decisões equivocadas, o processo KDD estabelece que antes da etapa de prospecção de dados seja realizada a etapa de *data cleaning*, que consiste na detecção e na eliminação de erros nas bases de dados para garantir a qualidade dos dados armazenados em grandes repositórios (RAHM; DO, 2000).

O pré-processamento também envolve a transformação dos dados, em que os dados são transformados e consolidados por meio de técnicas como a normalização, com o propósito de torna-los adequados a etapa de prospecção de dados, bem como melhorar a precisão e o de-

sempenho do algoritmo. Outra técnica frequentemente aplicada nesta etapa é a redução de dados, que por meio de técnicas de agregação e agrupamento é capaz de reduzir o volume de dados com a eliminação de características redundantes, o que também contribui para alcançar um melhor desempenho na etapa de prospecção de dados (HAN; KAMBER, 2011).

Com isso, é possível melhorar a qualidade dos dados armazenados e consequentemente a confiabilidade dos resultados alcançados na etapa de prospecção de dados.

2.2.2 Prospecção de dados

O grande volume de dados armazenados em diversas áreas de conhecimento tem exigido o desenvolvimento de técnicas e ferramentas no sentido de apoiar a extração de conhecimento a partir desses dados, uma vez que o ser humano é incapaz de analisar e entender uma grande quantidade de dados manualmente: tais circunstâncias remetem a um cenário onde existem dados valiosos, mas há pouca informação. A etapa de prospecção de dados pode ser entendida como o conjunto de técnicas automáticas utilizadas para descobrir padrões e relações em grandes bases de dados em que tal tarefa não é trivial ao ser humano. As técnicas mais utilizadas para a descoberta de conhecimento são: classificação e regressão; regras de associação; agrupamento; caracterização e discriminação; e análise de exceções (HAN; KAMBER, 2011).

2.2.3 Pós-processamento dos dados

A etapa de prospecção de dados pode gerar um número muito elevado de resultados, porém nem todos apresentarão conhecimento interessante e inovador. Além disso, o volume de resultados obtidos pode tornar a sua análise impraticável. Assim, a etapa de pós-processamento tem o propósito de validar, interpretar e apresentar o conhecimento extraído na etapa de prospecção de dados de forma que possa ser compreensível ao ser humano (HAN; KAMBER, 2011).

Neste cenário, a visualização dos dados se mostra uma importante ferramenta da etapa de pós-processamento, já que é capaz de representar graficamente os resultados obtidos na etapa de prospecção de dados e revelar padrões e relações implícitas, o que facilita o processo de descoberta de conhecimento (QIANG; WEI; HANFEI, 2010, JIA; LIU, 2009).

2.3 Dados Espaciais

Os dados espaciais são formados por atributos estruturados e não estruturados. Os atributos estruturados correspondem aos atributos convencionais, normalmente especificados por tipo alfanumérico, enquanto que os atributos não estruturados referem-se aos atributos espaciais que pode ser um ponto, uma reta ou um polígono (HAN; KAMBER, 2011, CÂMARA, 1995). Assim, o tamanho, o formato e o contorno dos atributos espaciais influenciam no processo de análise de tais dados.

O tipo mais comum de dado espacial é a localização, normalmente representada a partir dos sistemas de coordenadas (MAMOULIS, 2011), sendo que na maioria das definições sobre dados espaciais, o conceito de localização está relacionado à posição que o objeto está em relação à superfície da Terra (YEUNG; HALL, 2007). No entanto, também são considerados dados espaciais, os objetos relacionados à posição dos genomas responsáveis por determinada função biológica no DNA ou RNA (ROQUEIRO; FRASOR; DAI, 2010), bem como a posição das estrelas, luas e planetas no espaço sideral. Assim, em uma definição mais ampla, um dado espacial pode ser considerado qualquer dado que contemple um objeto geométrico, tais como: pontos multidimensionais, linhas, retângulos, polígonos e cubos (RAMAKRISHNAN; GEHRKE, 2008), conforme é ilustrado na Figura 2.

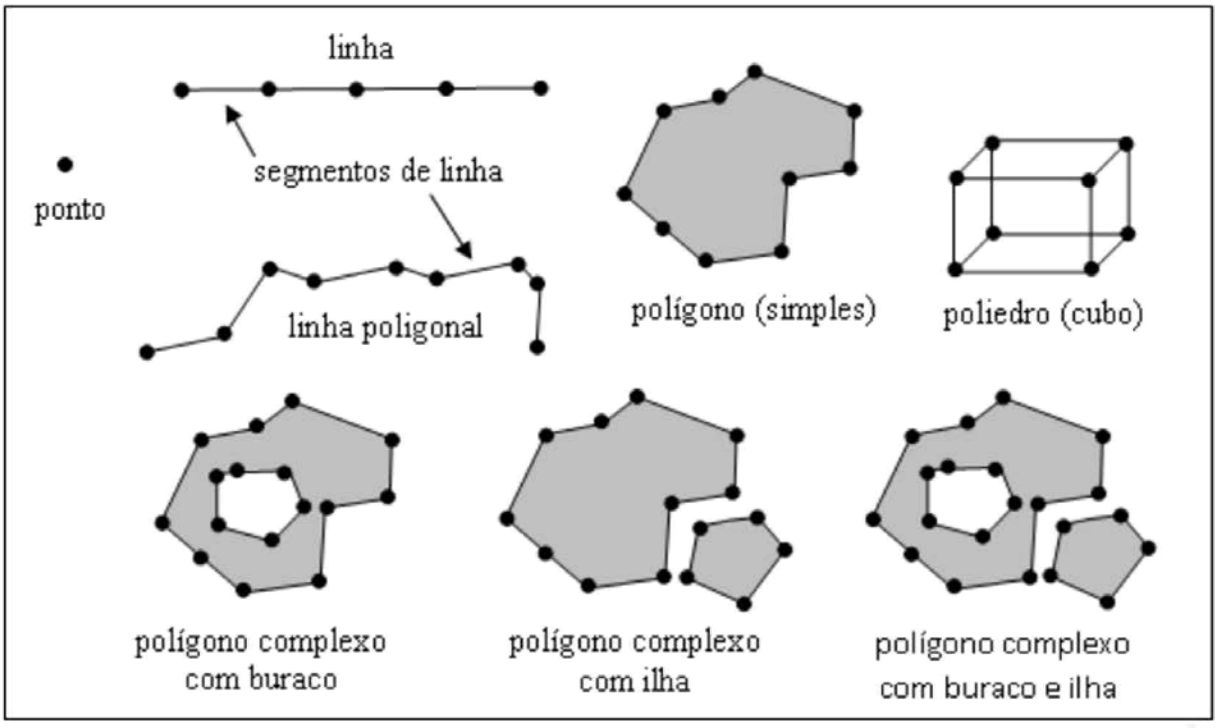

Figura 2: Tipos de dados espaciais (YEUNG; HALL, 2007).

Os Sistemas Gerenciadores de Banco de Dados Relacionais - SGBDRs não são capazes de manipular os dados espaciais com eficiência, já que para lidar com atributos não estruturados seria necessário utilizar o formato de arquivo, em que tais atributos seriam armazenados separadamente dos atributos estruturados, o que tornaria impraticável a extração de conhecimento (JIEHAI; WEI, 2010). Com o objetivo de possibilitar a manipulação de dados espaciais, muitos bancos de dados relacionais criaram extensões específicas para esta tarefa, como o PostGIS (REFRACTIONS RESEARCH, 2013) que é uma extensão do PostgreSQL (POSTGRESQL, 2013) para manipulação de dados espaciais, que fornece funções de: topologias, validação de dados e transformação de coordenadas (OBRADOVIC et al., 2011). Dentre os bancos de dados espaciais, destacam-se: banco de dados geográficos, banco de dados de imagens médicas e banco de dados de imagens de satélite (HAN; KAMBER, 2011).

Os bancos de dados espaciais armazenam informações espaço-relacionadas, que geralmente são armazenados na forma de vetor - utilizando a representação geométrica dos objetos por meio de pontos, linhas, polígonos ou *raster* – por meio de estruturas n-dimensionais formadas, por exemplo, com mapas de *bits* ou mapas de *pixels*, para representação de imagens de satélite (MAMOULIS, 2011). Na Figura 3 é apresentado um exemplo de como um mesmo objeto é representado na forma de vetor e *raster* para ser armazenado em uma base de dados espaciais.

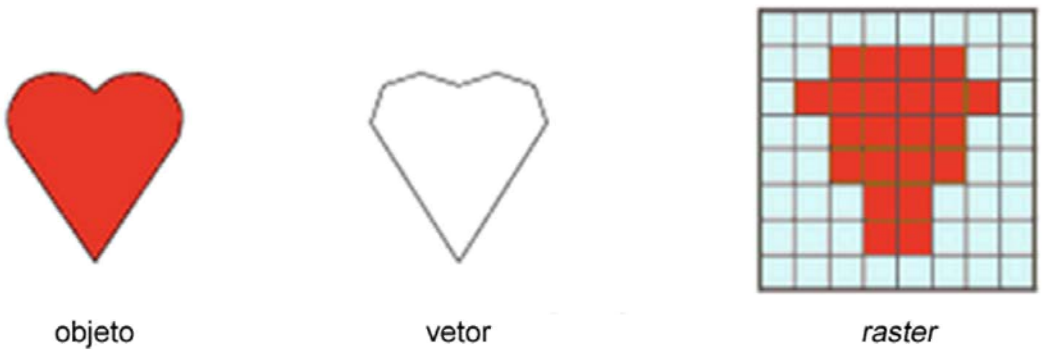

Figura 3: Representação de um objeto no formato vector e *raster* adaptado de (MAMOULIS, 2011).

Nas consultas em base de dados espaciais frequentemente é necessário analisar a posição de um objeto no espaço. Assim, obter as relações espaciais é uma importante tarefa na análise de dados espaciais, ao passo que estes são espacialmente dependentes, ou seja, existe uma relação entre a localização destes dados, sendo que a direção, conectividade, distância e outros atributos podem influenciar nesta dependência (MILLER; HAN, 2009). Uma das relações espaciais é a topologia que consiste em estabelecer a relação do objeto e seus vizinhos de

acordo com um conjunto de relações entre a borda, o interior e o exterior de cada objeto (YEUNG; HALL, 2007, ABDUL-RAHMAN; PILOUK, 2007). Na Figura 4 são apresentadas as principais relações topológicas entre os objetos.

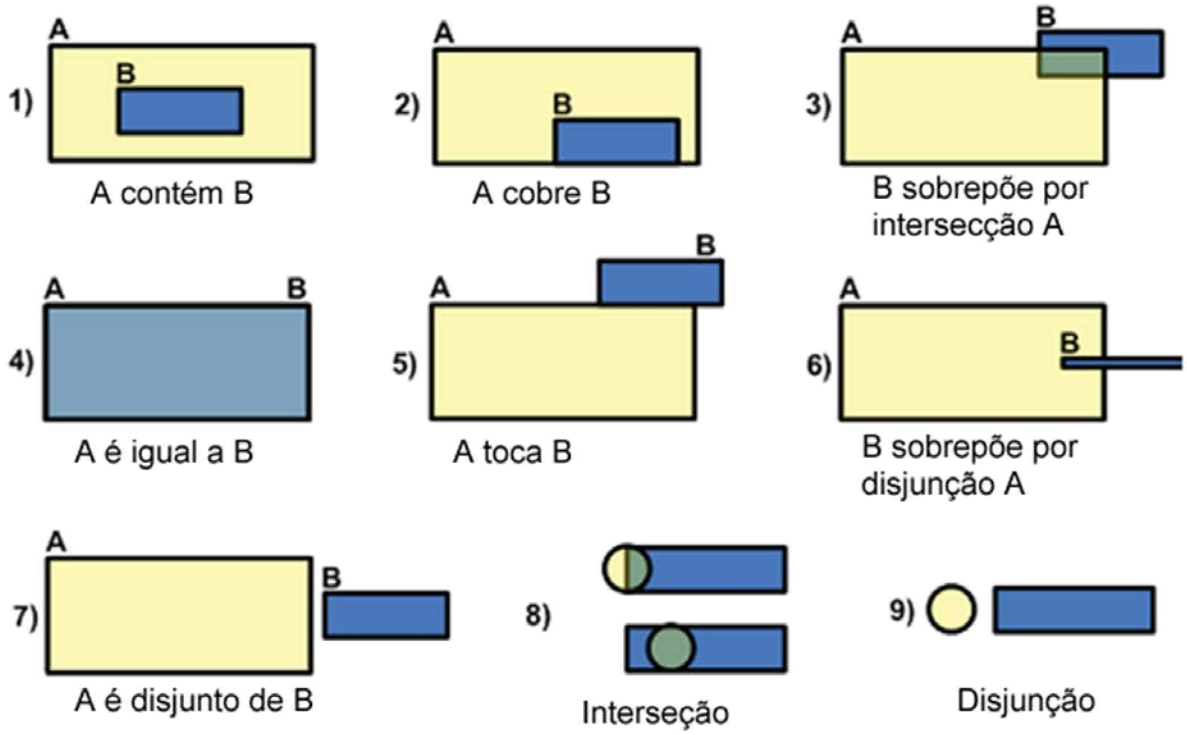

Figura 4: Principais relações topológicas adaptadas de (ROQUEIRO; FRASOR; DAI, 2010).

Outra relação espacial bastante utilizada é a distância entre objetos, que normalmente é calculada pela distância Euclidiana ou Manhattan. A relação de distância é amplamente utilizada em algoritmos de prospecção de dados espaciais, ao passo que estabelece quanto um objeto está próximo de outro. Por fim, existe também a relação de orientação, que pode definir se um objeto está posicionado à esquerda, à direita, abaixo ou acima de outro. No contexto de dados geográficos, a relação de orientação também pode ser utilizada para verificar a localização de um objeto em relação ao sistema de referência global, no qual um objeto pode ser localizado a norte, sul, leste e oeste (MAMOULIS, 2011).

As relações topológicas são verificadas a partir de funções booleanas disponibilizadas pelas extensões espaciais dos Sistemas Gerenciadores de Banco de Dados - SGBD's e utilizadas como predicados nas consultas, por meio dos operadores: *Equals*, *Disjoint*, *Intersects*, *Touches*, *Crosses*, *Within*, *Contains*, *Overlaps*, *Beyond*, *DWithin* e *BBOX*. Para obter outras propriedades do objeto geométrico, como por exemplo, formato e limite, são utilizadas as

operações de análise espacial: *Distance*, *Buffer*, *ConvexHull*, *Intersection*, *Union*, *Difference* e *SymDifference*(YEUNG; HALL, 2007).

As características dos dados espaciais aliadas à grande heterogeneidade dos dados espaciais, múltiplas dimensões relacionadas e ao grande volume de dados armazenados, tornam impraticável a utilização de métodos convencionais de prospecção de dados, bem como de métodos tradicionais de análise espacial - como estatística espacial, cartografia analítica e análise exploratória de dados espaciais - para a extração de conhecimento em base de dados espaciais (MENNIS; GUO, 2009).

2.4 Aplicações de Dados Espaciais

Os dados espaciais são amplamente utilizados em sistemas de informação geográfica, uma vez que estes apresentam recursos para a manipulação e análise de dados geográficos, que são dados espaciais cuja localização é referente à posição do objeto em relação à superfície da Terra, normalmente tais dados são representados por pontos, linhas e regiões bidimensionais e tridimensionais (RAMAKRISHNAN; GEHRKE, 2008).

Outra possibilidade de aplicação dos dados espaciais é em bancos de dados de imagens médicas, nos quais são armazenadas imagens provenientes de equipamentos de raios-X, ressonância magnética, entre outros, em que a recuperação de informações é baseada no conteúdo das imagens e poderiam ser simplificadas se ao invés das imagens, nas consultas fossem utilizados os dados espaciais para representá-las (RAMAKRISHNAN; GEHRKE, 2008).

Os dados espaciais também podem ser utilizados no reconhecimento de padrões, como por exemplo, para verificar se a topologia de um determinado gene no genoma está presente em outro mapa de característica de sequência na base de dados (ELMASRI; NAVATHE, 2011).

2.5 Prospecção de Dados Espaciais

Com o propósito de resolver as necessidades sobre bases de dados espaciais foi proposta a prospecção de dados espaciais ou *Knowledge Discovery in Spatial Databases* - KDSD, que pode ser entendido como o processo de extração de dados a partir de banco de dados espaciais com conhecimento e padrões implícitos e relações espaciais (JIN; MIAO, 2010). Tal processo surgiu recentemente, a partir de métodos tradicionais de análise espacial e das técnicas convencionais de prospecção de dados, e consiste em descobrir a relação entre os

atributos espaciais e não espaciais (MENNIS; GUO, 2009, CÂMARA, 1995). A seguir as técnicas de prospecção de dados espaciais são apresentadas.

2.5.1 Classificação espacial

No contexto convencional, tal técnica consiste em classificar um conjunto de dados de acordo com determinados conceitos pré-estabelecidos, sendo, portanto um método supervisionado. Assim, é gerado um modelo, capaz de representar o comportamento dos dados em geral, que pode ser representado por regras se-então, árvores de decisão, fórmulas matemáticas ou redes neurais. Os algoritmos de classificação são bastante utilizados para elaboração de estratégias de *marketing*, diagnósticos médicos, indústria e detecção de fraudes (JIEHAI; WEI, 2010).

Na classificação espacial o objetivo é definir um modelo que relacione as características espaciais e não espaciais dos objetos e seus vizinhos, com a distribuição dos objetos em classes pré-definidas, baseadas em um conjunto de atributos espaciais e suas relações espaciais (MENNIS; GUO, 2009, JIEHAI; WEI, 2010). Assim, esta técnica pode ser utilizada para entender o relacionamento entre os objetos e prever o comportamento de outros objetos, como por exemplo, a classificação de uma cidade em rica ou pobre, baseada na renda média das famílias (SANTOS; AMARAL, 2004). Outro exemplo nesse sentido, é utilizar dados sobre durabilidade da vegetação e profundidade da água para descobrir os locais de ninhos em um pântano (ELMASRI; NAVATHE, 2011).

Na área de sensoriamento remoto, a classificação espacial tem sido bastante aplicada para classificar os *píxels* das imagens, uma vez que estes são formados por valores similares e possuem tamanho, formato e relação espacial compatíveis com o cenário do mundo real que modela (CLEVE, 2008, MENNIS; GUO, 2009). Um exemplo nesse sentido é a classificação dos objetos no espaço sideral, tais como: estrelas, galáxias, luas e planetas em uma imagem. Também é possível utilizar a classificação espacial para identificar nas imagens, os locais em que são encontrados vulcões na superfície de Vênus (KOPERSKI; HAN, 1997). Outro exemplo é apresentado na Figura 5, em que foram utilizados os *píxels* de uma imagem para classificar a vegetação de uma região geográfica.

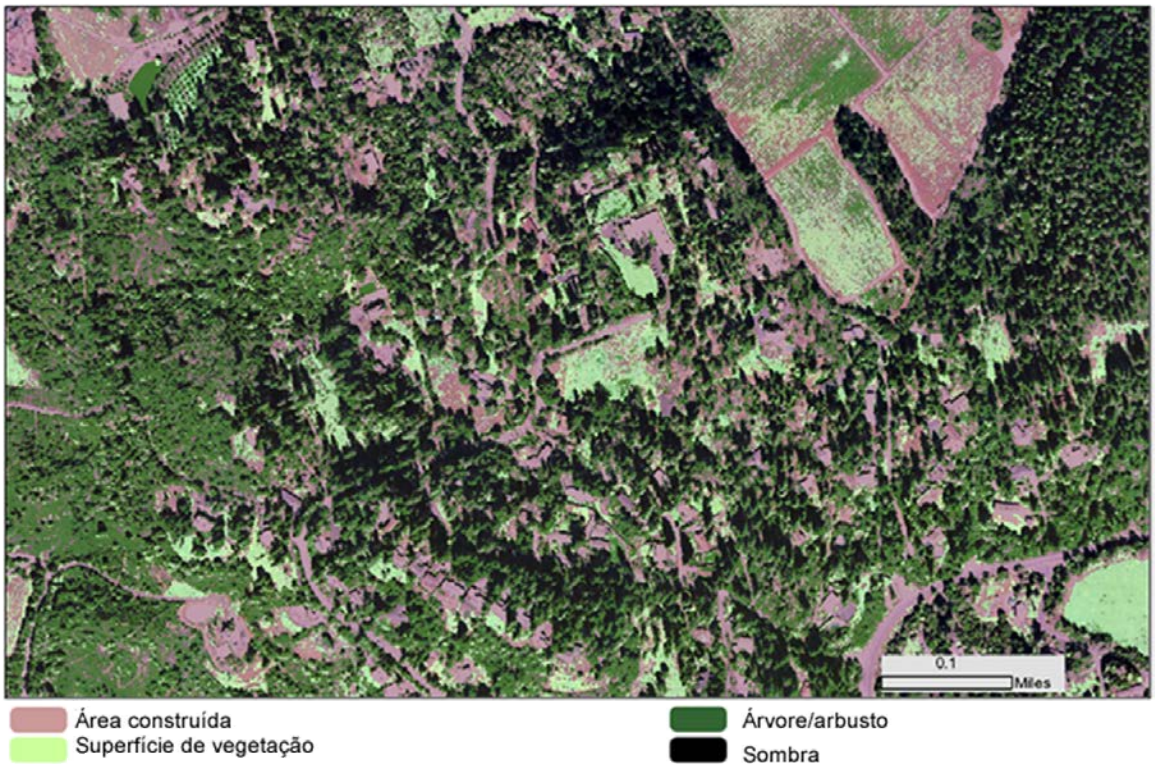

Figura 5: Classificação de uma região geográfica com base em *pixels* adaptado de (CLEVE et al., 2008).

2.5.2 Regras de associação espacial

Para realizar a descoberta de conhecimento em bases de dados, uma técnica bastante utilizada é a prospecção por meio de regras de associação. Tal técnica permite criar regras que refletem o quanto determinado item está correlacionado ou associado a outros e quais são os padrões frequentes nas bases de dados. Para verificar se uma regra de associação é útil, são analisados o suporte e a confiança da regra que devem ser superiores aos limites definidos pelo usuário ou pelo próprio algoritmo. O suporte representa o percentual do conjunto de dados processado para qual a regra é válida e a confiança mostra a frequência que a regra é válida para a base de dados (HAN; KAMBER, 2011).

A associação espacial busca regras de associação para identificar padrões e a frequência dos mesmos em um conjunto de dados por meio de predicados espaciais ou não espaciais (SANTOS; AMARAL, 2004). Assim, no contexto espacial uma regra de associação pode ser definida por (MILLER; HAN, 2009):

$$S \rightarrow N \text{ [Suporte = X\%, Confiança = Y\%] ou}$$

$$N \rightarrow S \text{ [Suporte = X\%, Confiança = Y\%]}$$

Em que S representa o conjunto de predicados espaciais, enquanto que N representa o conjunto de predicados não espaciais (JIN; MIAO, 2010). Dentre os predicados espaciais podem-se relacionar os predicados topológicos: disjunto, toque, sobreposição, cobrir, conter, dentro, bem como distância e direção (MILLER; HAN, 2009). Diante da complexidade dos dados espaciais e dos relacionamentos que devem ser tratados na associação espacial, geralmente são utilizadas estratégias para otimizar o processo de extração de conhecimento por meio de regras de associação (JIN; MIAO, 2010).

As regras de associação espacial possibilitam obter, por exemplo, que países adjacentes ao Mar Mediterrâneo são produtores de vinho com determinado suporte e confiança (ELMASRI; NAVATHE, 2011). Também é possível aplicá-las para obter padrões sobre os crimes ocorridos em São José do Rio Preto, que resultaria em regras do tipo: "a maioria dos crimes ocorridos no ano de 2012 estão próximos à av. Bady Bassit". Outro exemplo de aplicação das regras de associação espacial é apresentado na Figura 6, no qual foram geradas regras a partir da relação entre lotes, valores de Imposto Predial e Territorial Urbano - IPTU, escolas e relação topológica de proximidade. Neste exemplo, a regra destacada mostra que 14,5% dos lotes particulares próximos às escolas estaduais, não são isentos a IPTU e que 99,7% dos lotes particulares também não são isentos de IPTU.

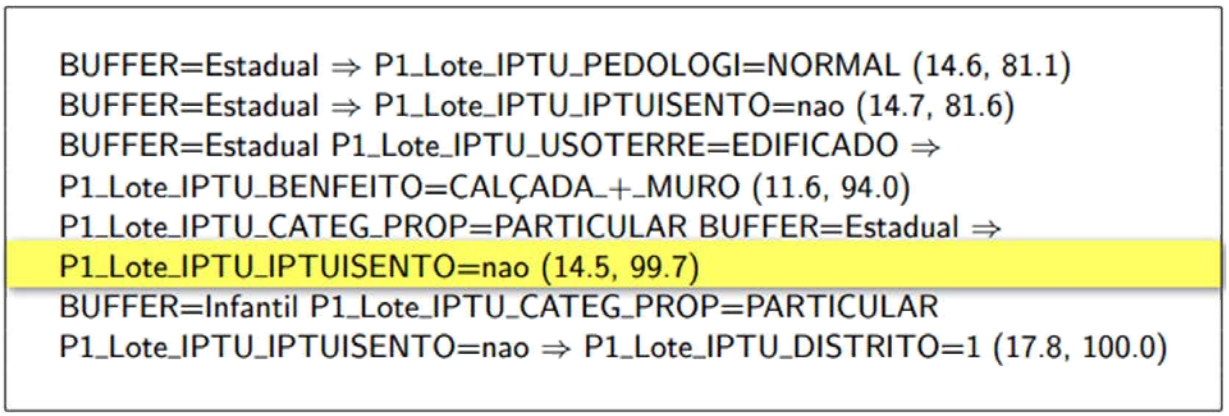

Figura 6: Regras de associação de lotes e escolas em Jaboticabal adaptado de (PIVATO, 2006).

2.5.3 Tendência espacial

A análise de tendência, no contexto convencional, refere-se à análise dos dados por meio de linhas e curvas, tais como análise linear ou análise de regressão logística, que são fáceis e rápidas de estimar (MILLER; HAN, 2009). Para isso, tem-se um modelo integrado com componentes e movimentos para caracterizar os dados de séries temporais em relação à (HAN; KAMBER, 2011):

- **Tendências ou movimentos em longo prazo –** indica a direção geral que a série temporal do gráfico se movimentará ao longo do tempo;
- **Ciclo de movimentos –** indica a oscilação em longo prazo da tendência;
- **Variação sazonal –** indica padrões que aparecem na série temporal durante vários períodos em anos sucessivos;
- **Movimento randômico –** caracteriza mudanças esporádicas durante um acontecimento.

A técnica de tendência espacial encontra padrões de mudança e de tendência dos atributos não espaciais a partir do movimento de um objeto (JIN; MIAO, 2010). Para isso, a mudança de localização do objeto é modelada por meio da análise dos vizinhos, em seguida é analisada a regressão dos atributos do objeto para determinar a tendência espacial, sendo que a distância do objeto aos vizinhos é utilizada como variável independente e a diferença dos atributos do objeto e dos vizinhos é a variável dependente da regressão (ESTER et al., 1998). Na Figura 7 são apresentadas as três situações que podem ser obtidas com a aplicação da regressão linear: tendência positiva, tendência negativa e sem tendência.

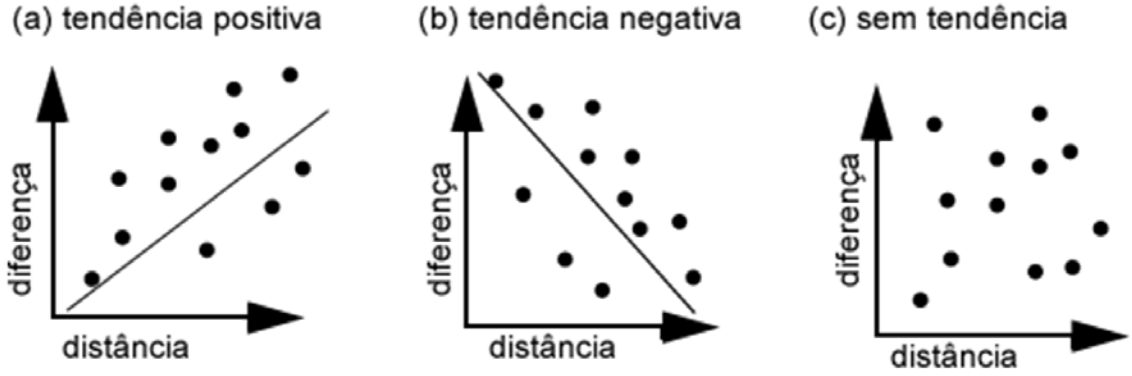

Figura 7: Situações de tendência espacial adaptado de (ESTER; KRIEGEL; SANDER, 2001).

A tendência espacial pode ser aplicada para avaliar o poder econômico de regiões próximas, analisando, por exemplo, a renda média da população em cada região. Um exemplo disso é apresentado na Figura 8, na qual se observa uma queda regular na renda média da população ao passo que a região encontra-se mais distante de Munique.

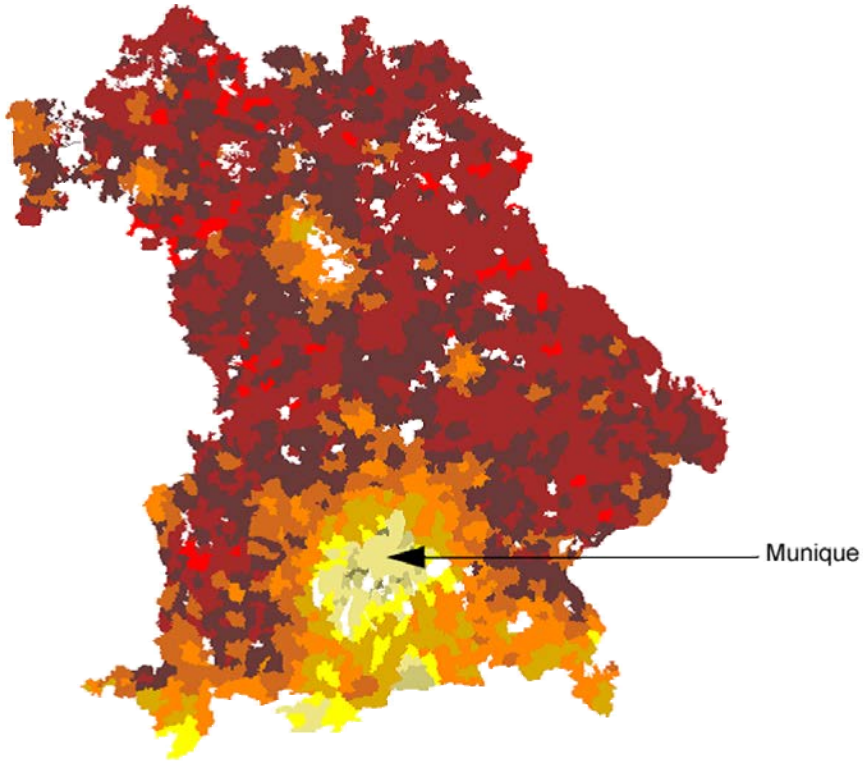

Figura 8: Renda média das comunidades da Baviera adaptado de (ESTER; KRIEGEL; SANDER, 1997).

2.5.4 Caracterização espacial

A caracterização de dados no contexto convencional consiste em resumir as propriedades de um conjunto alvo de dados (HAN; KAMBER, 2011). Já a caracterização espacial consiste na descrição de predicados espaciais ou não espaciais, que permite identificar características de um grupo de dados específico. Para isso, são analisadas as propriedades do grupo e também de seus vizinhos (MILLER; HAN, 2009, HASAN et al., 2008). Os algoritmos de caracterização espacial normalmente executam os seguintes passos (ESTER; KRIEGEL; SANDER, 2001):

1. Seleção de um conjunto de dados que possua uma determinada característica não espacial.
2. Seleção dos vizinhos do conjunto de dados, cuja frequência relativa possua uma diferença significativa do restante da base.
3. Geração da regra de caracterização que descreve o conjunto de dados selecionado no passo 1.

Um exemplo de caracterização espacial é apresentado na Figura 9(a), na qual são selecionadas regiões alvo que contemplem um alto índice de aposentados. Em seguida, são selecionadas as regiões próximas às regiões alvo e verificadas as características destas que

diferem das encontradas em toda a base de dados, conforme é exibido na Figura 9(b). Ao final são geradas as regras que caracterizam as regiões alvo, como por exemplo, “idosos preferem viver em áreas rurais próximas a montanhas” (ESTER; KRIEGEL; SANDER, 2001).

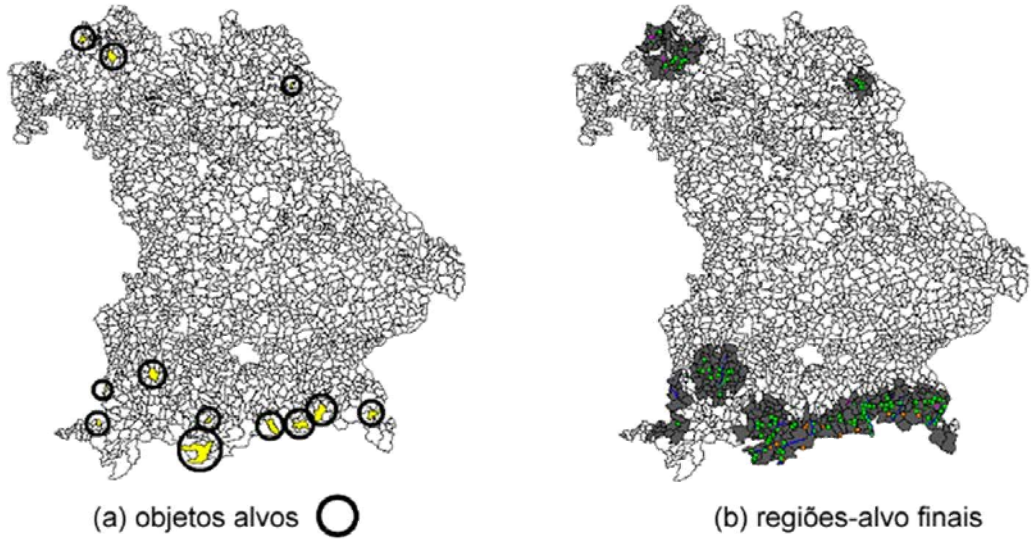

Figura 9: Caracterização da proporção de idosos adaptado de (ESTER; KRIEGEL; SANDER, 2001).

2.5.5 Agrupamento de dados espaciais

O agrupamento de dados convencional particiona os objetos em grupos, reunindo em um mesmo agrupamento aqueles que são semelhantes. Tal técnica é capaz de identificar características inesperadas nos agrupamentos de dados, além de ser bastante adaptável às mudanças na base (HAN; KAMBER, 2011).

O agrupamento de dados espaciais reúne os objetos espaciais semelhantes em um mesmo agrupamento para encontrar a distribuição de padrões, sendo considerada a localização espacial dos objetos (FAN, 2009). Tal técnica é detalhada no Capítulo 3.

2.6 Desafios da Prospecção de Dados Espaciais

O grande volume de dados espaciais que vem sendo coletado oferece a oportunidade de descoberta de conhecimento em diversas áreas, tais como: saúde, meio ambiente, arqueologia, extração mineral, gerenciamento urbano, *marketing* e meteorologia (MENNIS; GUO, 2009). No entanto, para que a prospecção seja realizada em tempo hábil e que de fato traga resultados significativos à área em que é aplicada, a prospecção de dados espaciais necessita superar alguns desafios (JIN; MIAO, 2010):

- **Complexidade –** As bases de dados espaciais normalmente são muito maiores que as bases de dados convencionais, por conta da complexidade das estruturas dos dados espaciais e de seus relacionamentos. Assim, problemas relacionados a armazenamento e processamento são frequentemente encontrados;
- **Escalabilidade –** Os algoritmos em geral são pouco eficientes e não refinam os padrões encontrados. Com isso, os resultados apresentam muitos padrões incorretos, que dependem de uma análise humana para serem identificados e removidos;
- **Padronização –** Os bancos de dados espaciais ainda não possuem uma linguagem específica de consulta que facilite a varredura em estruturas espaciais. Assim, as consultas em banco de dados espaciais utilizam a mesma linguagem de consulta para dados convencionais, como por exemplo, a *Structured Query Language* - SQL. A padronização de uma linguagem para dados espaciais poderia tornar mais eficiente o processo de extração de conhecimento;
- **Interação –** As técnicas de prospecção de dados espaciais ainda são muito dependentes do conhecimento humano para revelar padrões úteis e inéditos;
- **Generalização –** Normalmente, as ferramentas de prospecção de dados espaciais são desenvolvidas para atender a um domínio específico, o que limita o conhecimento que pode ser descoberto.

2.7 Estratégias para Melhorar o Desempenho dos Algoritmos

Uma estratégia para minimizar os problemas relacionados a processamento na aplicação de algoritmos de prospecção de dados espaciais é utilizar todo o potencial disponibilizado na arquitetura *multicore* para reduzir o tempo de processamento dos algoritmos. No entanto, para atingir este objetivo é necessário que os algoritmos sejam desenvolvidos utilizando a programação paralela (GENG; YANG, 2013), (EICHINGER; PANKRATIUS; BÖHM, 2014), (NATARAJAN et al., 2012).

A técnica mais utilizada para a implementação da programação paralela é a *multithreading,* que pode ser entendida como a execução simultânea de várias *threads* independentes, ou seja, tarefas independentes que pertencem a um mesmo processo e que podem ser executadas em paralelo ou de forma intercalada no tempo (KYUEUN; GAUDIOT, 2010). Na Figura 10 é exibida uma representação gráfica da técnica *multithreading.*

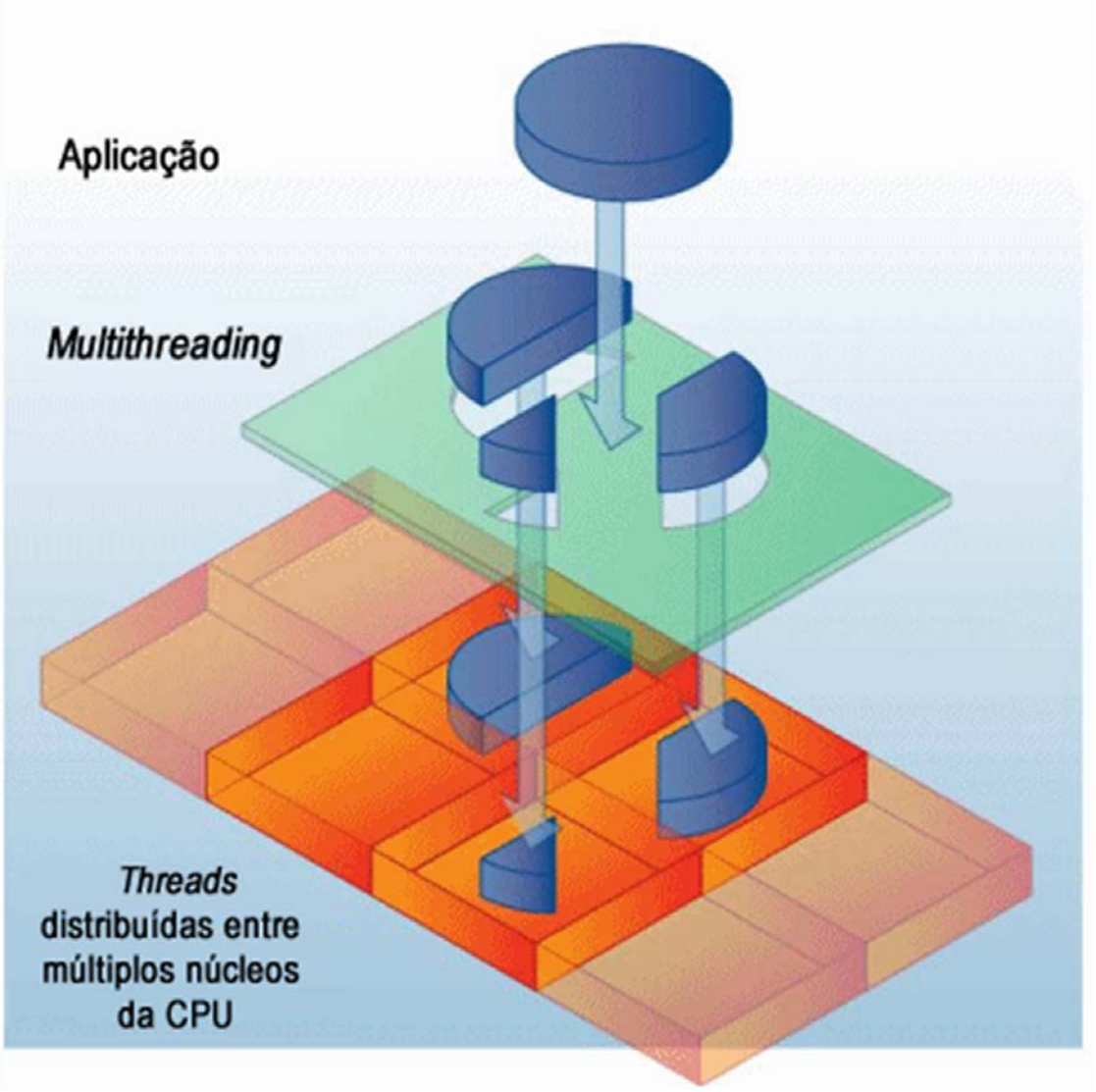

Figura 10: Representação da técnica de *multithreading* adaptada de (ANWER, 2007).

Os principais benefícios da técnica *multithreading* são (ORACLE, 2014, NEMIROVSKY; TULLSEN, 2013):

- Paralelizar os processos com um baixo custo, pois as *threads* impõem um impacto mínimo sobre os recursos do sistema;
- Otimizar o rendimento do algoritmo por meio da execução de múltiplas operações simultaneamente e solicitação de E/S em um único processo;
- Utilizar múltiplos processadores simultaneamente para computação de I/O;
- Melhorar a capacidade de resposta do servidor, uma vez que solicitações grandes ou complexas não bloqueiam outras solicitações de serviço;
- Melhorar a eficiência energética, ao passo que reduz o vazamento estático, que ocorre quando o processador fica ocioso;

- Simplificar a estrutura de aplicações complexas, uma vez que podem ser construídos métodos simples para cada atividade, facilitando a construção de sistemas complexos;
- Melhorar a comunicação entre processos por meio de funções de sincronização.

Outra forma de melhorar o desempenho dos algoritmos é utilizar a abordagem *MapReduce*, que consiste em um modelo de programação desenvolvido pelo Google no qual os algoritmos são construídos em termos das funções *Map* e *Reduce*, que são baseadas nas primitivas *map* e *reduce* das linguagens funcionais, como por exemplo, a LISP. Os dados são organizados em pares (chave, valor), em que o trabalho é distribuído em pequenas tarefas, e os valores intermediários são calculados por meio da função *Map*, após esta etapa a função *Reduce* é aplicada, na qual as listas de valores intermediários com a mesma chave são combinadas para obter o resultado final. A principal vantagem da abordagem *MapReduce* é que os detalhes de tolerância a falhas, dados de distribuição e balanceamento de carga são transparentes ao usuário (DEAN; GHEMAWAT, 2008, YANG ET AL., 2007, HE et al., 2011). Na Figura 11 é exibida uma representação da abordagem *MapReduce*.

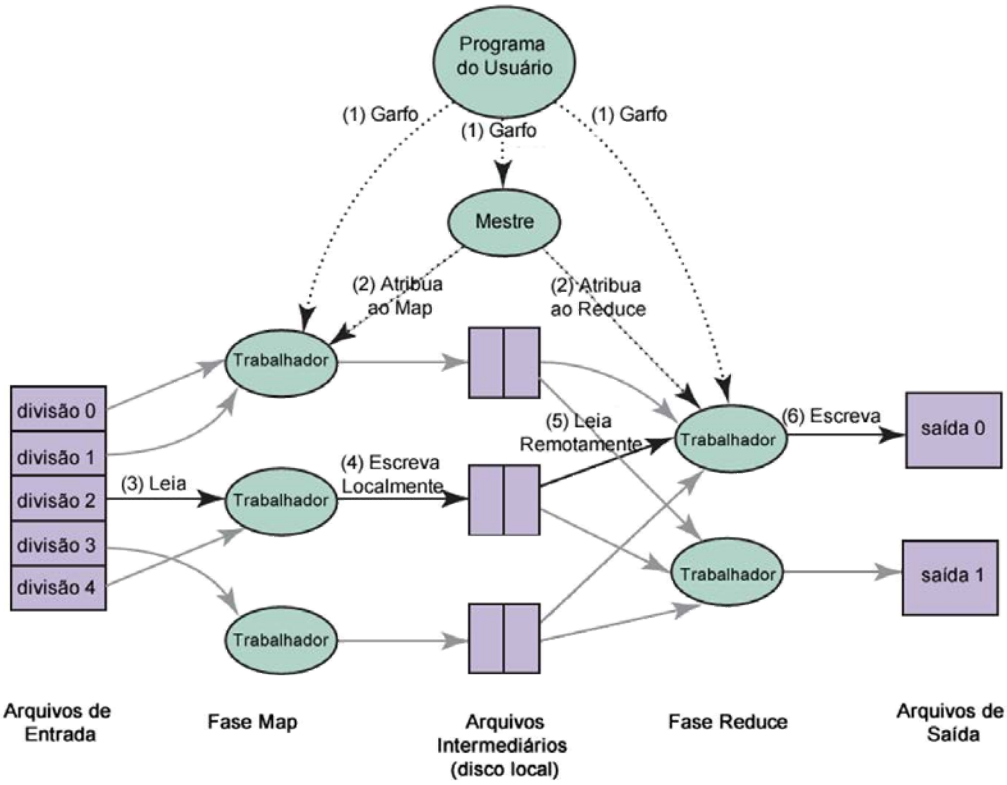

Figura 11: Visão Geral da Abordagem *MapReduce* adaptada de(DEAN; GHEMAWAT, 2008) .

Na Tabela 1, as características das abordagens *multithreading* e *MapReduce* são sumarizadas.

Tabela 1: *Multithreading* versus *MapReduce* (DEAN; GHEMAWAT, 2008, KAFURA, 2014).

Características	*Multithreading*	*MapReduce*
O que é	Uma técnica para programação paralela	Um paradigma de programação e seu sistema de execução associado
Execução	Paralela	Sequencial, já que a função *Reduce* só pode ser executada após o término da função *Map*
Dado a ser processado	Irrestrito	Par <chave, valor>
Operações	Irrestritas	*Map* e *Reduce.*
Usabilidade	As tarefas devem ser independentes	Conceitos simples podem ser difíceis de otimizar
Principal benefício	Aproveitamento dos recursos disponibilizados na arquitetura multicore com um baixo custo	Detalhes de tolerância a falhas, dados de distribuição e balanceamento de carga são transparentes ao usuário

2.8 Considerações Finais

O volume de dados espaciais que tem sido coletado motivou o desenvolvimento de diversas técnicas para a prospecção de dados espaciais, que vem sendo aplicadas em diversos contextos, contribuindo com o avanço da ciência de forma geral. No entanto, vários desafios ainda precisam ser superados para que a prospecção de dados espaciais possa ser aplicada de forma mais eficiente e eficaz.

Este trabalho tem o objetivo de contribuir neste sentido, utilizando para isso o agrupamento de dados espaciais, uma técnica que tem motivado o desenvolvimento de várias pesquisas devido à capacidade de gerar resultados sem requerer um conhecimento prévio do usuário e ser aplicável em diversos domínios de aplicações. Com o propósito de facilitar o entendimento do trabalho proposto, no Capítulo 3 é apresentada a fundamentação teórica referente a técnica de agrupamento de dados espaciais, bem como é apresentado um estudo comparativo realizado com diversos algoritmos que implementam a referida técnica.

CAPÍTULO 3
Agrupamento de Dados Espaciais

3.1 Considerações Iniciais

O agrupamento de dados espaciais reúne os objetos espaciais semelhantes em um mesmo agrupamento para encontrar a distribuição de padrões, levando em consideração a localização dos objetos (FAN, 2009). Tal técnica tem se mostrado um método importante da tarefa de prospecção de dados espaciais, visto que não é necessário ter um conhecimento prévio para executar seu trabalho, tal como a classificação espacial, e pode ser empregado em diversos domínios de aplicações, tais como: análise de imagens de satélite, sistemas de informações geográficas, análise de imagens médicas, *marketing* e etc. (PATTABIRAMAN et al., 2009). Com isso, nos últimos anos, muitos pesquisadores têm investido esforços no desenvolvimento de novas técnicas e algoritmos de agrupamento de dados espaciais com o objetivo de extrair conhecimento útil das bases de dados espaciais com eficácia e eficiência.

Para aplicar o agrupamento dos dados espaciais é possível utilizar os seguintes métodos: hierárquico, particionamento, baseado em densidade, baseado em grade e baseado em grafos. Tais métodos são detalhados nas próximas subseções, bem como os principais algoritmos utilizados. É importante salientar que apesar de alguns algoritmos não serem aplicados exclusivamente para a prospecção de dados espaciais, estes são frequentemente citados como trabalhos correlatos e utilizados como base para a confecção de novos algoritmos na área, o que caracteriza a importância de descrevê-los neste trabalho.

3.2 Método Hierárquico

O método hierárquico agrupa os objetos em uma estrutura de árvore e a cada iteração o espaço da busca é reduzido (MILLER; HAN, 2009). Cada nó desta estrutura armazena conjuntos de agrupamentos filhos, sendo que cada nó filho possui objetos em comum com o nó pai. Assim, a prospecção de dados pode ser realizada em diferentes níveis de granularidade. Este método pode ser aplicado de duas formas, a saber (MILLER; HAN, 2009, XI, 2009):

- **Aglomerativo -** Neste método a construção da árvore é realizada por meio da técnica *bottom-up*, em que vários pequenos agrupamentos são unidos em agrupamentos maiores até atingir algum parâmetro de parada;
- **Divisivo -** A construção da árvore inicia-se por um grande agrupamento, composto por todo o conjunto de objetivos, que vão sucessivamente sendo divididos em agrupamentos menores por meio da técnica *top-down*.

Na Figura 12 são ilustradas as duas abordagens do método hierárquico.

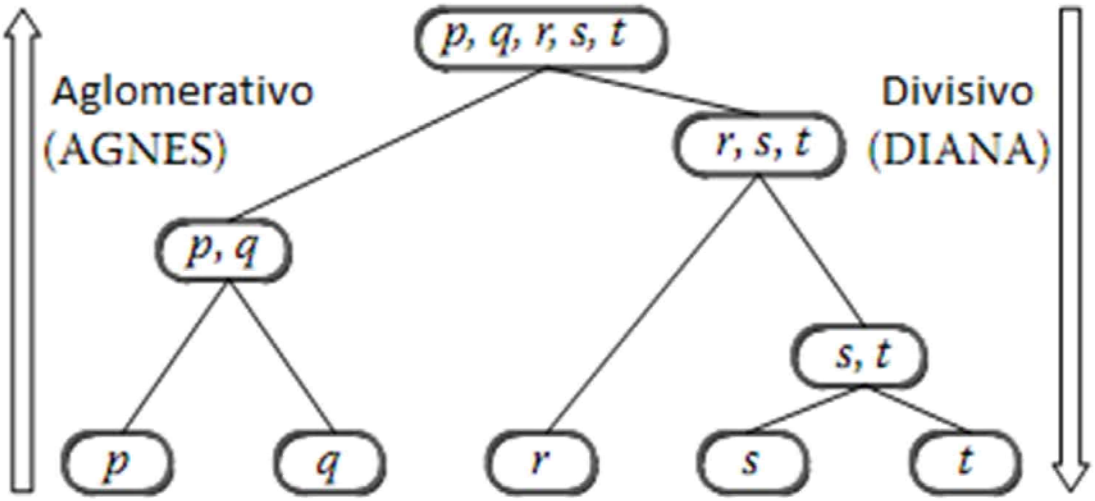

Figura 12: Agrupamento de dados por meio dos métodos aglomerativo e divisivo adaptado de (MILLER; HAN, 2009).

Alguns dos algoritmos hierárquicos encontrados na literatura são: *AGglomerative NESting* - AGNES (KAUFMAN; ROUSSEEUW, 1990), *DIvisive ANAlysis* - DIANA (KAUFMAN; ROUSSEEUW, 1990); *Single*, *Average* e *Complete-Linkage* (JAIN, 1988); *Balanced Interative Reducing and Clustering using Hierarchies* - BIRCH (ZHANG; RAMAKRISHNAN; LIVNY, 1996).

O algoritmo AGNES define que cada objeto é um agrupamento. Em seguida, é iniciado o processo de junção, em que os agrupamentos iniciais são mesclados para formar um único agrupamento, que será cada vez maior, até que todos os objetos estejam no mesmo agrupamento. Já o algoritmo DIANA realiza o oposto, inicia com um único agrupamento que reúne todos os objetos, e posteriormente o divide em agrupamentos menores até atingir um ponto de parada (MILLER; HAN, 2009).

O algoritmo *Single-Linkage* determina a distância mínima entre dois agrupamentos. Para isso é selecionado um objeto de um agrupamento que esteja mais próximo a um objeto

do outro agrupamento. Já o algoritmo *Complete-Linkage* determina a distância máxima entre dois agrupamentos, selecionando o objeto mais distante de cada agrupamento. O *Average-Linkage* não possui o problema de encadeamento como o *Single-Linkage* e consegue tratar conjunto de dados com grande número de pontos discrepantes (*outliers*), diferente do que ocorre com o *Complete-Linkage*. Por outro lado, o *Average-Linkage* possui um custo computacional mais elevado que os demais algoritmos (ANDREOPOULOS et al., 2009).

O algoritmo BIRCH é capaz de agrupar grandes volumes de dados numéricos, por meio da utilização das técnicas de agrupamento hierárquico e particionamento. O algoritmo implementa o agrupamento de características - que sumarizam as estatísticas sobre o agrupamento, como por exemplo a distância média entre os pontos e a soma dos quadrados dos pontos de dados - e constrói a árvore de características (*CF tree*) para otimizar a representação dos agrupamentos, o que torna o algoritmo escalável em grandes bases de dados (HAN; KAMBER, 2011).

A principal vantagem dos algoritmos hierárquicos é a construção de uma árvore que encadeia os agrupamentos gerados e pode revelar informações significativas que os outros métodos não seriam capazes de realizar (ZHONG; MIAO; FRÄNTI, 2011). Por outro lado, os algoritmos hierárquicos podem não selecionar corretamente os objetos que serão agrupados ou que serão dispersos em vários agrupamentos, já que a estrutura hierárquica de tais algoritmos permite a propagação de erros dos nós pais para os nós filhos. Com isso, o resultado da extração de conhecimento poderá não corresponder à realidade. Outra dificuldade encontrada é que tais algoritmos necessitam verificar um grande número de objetos e agrupamentos, o que torna a maioria desses algoritmos não escalável (MILLER; HAN, 2009).

3.3 Método de Particionamento

O método de particionamento divide iterativamente o conjunto de dados em subconjuntos, de forma que cada subconjunto seja composto por dados que são semelhantes em determinados atributos (NEVES; FREITAS; CÂMARA, 2001). Tal método possui algoritmos clássicos como o *k-means* (LLOYD, 1982) e o *k-medoid* (KAUFMAN; ROUSSEEUW, 1990).

O *k-means* utiliza o valor médio dos objetos do agrupamento como ponto de referência (centroide) para definir se um objeto pertence ou não a esse agrupamento de dados. Esse algoritmo é simples de implementar e relativamente eficiente no processo de agrupamento em grandes bases de dados. Entretanto, os resultados obtidos por esse algoritmo podem ser incor-

retos se a média tiver sido influenciada pela presença de ruídos e pontos discrepantes na base de dados (FAN, 2009).

O algoritmo *k-medoid* utiliza o objeto mais central do agrupamento como ponto de referência para definir se um objeto pertence ou não a esse agrupamento de dados. Os resultados obtidos com esse algoritmo são menos influenciados por ruídos e pontos discrepantes que os resultados obtidos com o *k-means*. Entretanto, o *k-medoid* exige um tempo de processamento muito maior (FAN, 2009). Os principais algoritmos baseados no *k-medoid* são: *Particioning Around Medoids* - PAM (KAUFMAN; ROUSSEEUW, 1990), *Clustering Large Applications* - CLARA (KAUFMAN; ROUSSEEUW, 1990) e o *Clustering Large Applications based on Randomized Search* -CLARANS (NG; HAN, 1994).

O algoritmo PAM define arbitrariamente os *k-medoids* iniciais e aloca cada objeto ao agrupamento que possui o *medoid* mais semelhante e/ou mais próximo ao objeto. A partir dos agrupamentos iniciais, o algoritmo realiza uma busca exaustiva até formar agrupamentos que sejam o mais homogêneo possível. Esse algoritmo não é eficiente para médio e grande volume de dados porque verifica todas as possibilidades de *medoids* e agrupamentos formados (NG; HAN, 1994).

O algoritmo CLARA tem custo computacional inferior ao PAM, visto que define um conjunto amostral dos dados e aplica o algoritmo PAM sobre esse conjunto com o objetivo de encontrar os *k-medoids* ótimos. Como a qualidade dos resultados do algoritmo CLARA depende da qualidade da amostra definida, os autores sugerem que sejam testadas cinco amostras de dados e escolhida a melhor para obter resultados satisfatórios com o CLARA (NG; HAN, 1994).

O algoritmo CLARANS busca os *k-medoids* ótimos de forma contínua e aleatória em subconjuntos da base de dados. Dessa forma os resultados satisfatórios são atingidos mais rapidamente que ao utilizar o algoritmo CLARA. O CLARANS foi demonstrado por seus autores, como sendo mais eficiente, em relação ao PAM e ao CLARA, para pequenos e grandes volumes de dados (NG; HAN, 2002).

O agrupamento de dados espaciais por meio do método de particionamento tem resultado satisfatório apenas para identificar agrupamentos esféricos e de tamanhos parecidos (DENG et al., 2011).

3.4 Método Baseado em Densidade

O método baseado em densidade encontra agrupamentos de qualquer formato baseado

na densidade dos dados e de seus vizinhos (CHOWDHURY, 2010). Alguns dos algoritmos deste método são: *Density-Based Spatial Clustering of Applications with Noise* - DBSCAN (ESTER et al., 1996), *Ordering Points to Identify the Clustering Structure* - OPTICS (ANKERST et al., 1999), o *Density Based Clustering* - DENCLUE (HINNEBURG; HINNEBURG; KEIM, 1998), *Varied Density Based Spatial Clustering of Applications with Noise* - VBDCAN (LIU; ZHOU; WU, 2007) e o ADACLUS – *ADAptive CLUStering* (NOSOVSKIY; LIU; SOURINA, 2008).

O DBSCAN utiliza dois parâmetros para formar os agrupamentos, a saber: *episolon* - EPS e o número mínimo de pontos - MinPts. O algoritmo inicia com a seleção arbitrária do número de pontos. Em seguida, verifica o número de vizinhos de cada um destes pontos, que no máximo estão a uma distância EPS do ponto e caso este número seja maior igual ao MinPts, o ponto e seus vizinhos são agrupados, caso contrário o ponto é marcado como ruído (*noise*). O processo é repetido até que todos os pontos e seus vizinhos tenham sido visitados (PATTABIRAMAN et al., 2009).

O OPTICS gera agrupamentos de densidades variadas, ordenando os objetos pela distância para representar a hierarquia intrínseca dos agrupamentos, além de permitir uma análise automática e interativa dos agrupamentos. Com isso é possível obter algumas informações tais como os centros dos agrupamentos e formatos variados (HAN; KAMBER, 2011, PETER; ANTONYSAMY, 2010).

O algoritmo DENCLUE é baseado em funções de distribuição de densidade, que estimam o local com densidade máxima. Este local é utilizado como base para a geração dos agrupamentos. Se o local for considerado pequeno, os objetos deste agrupamento serão classificados como ruídos (PETER; ANTONYSAMY, 2010). O DENCLUE generaliza os métodos de particionamento, hierárquico e de densidade, o que permite gerar agrupamentos de formatos variados e o torna eficiente para bases de dados com grande número de ruídos. Por outro lado, a escolha dos parâmetros de entrada pode influenciar significativamente nos resultados obtidos (HAN; KAMBER, 2011).

Os algoritmos de agrupamento baseados em densidade normalmente não são capazes de gerar agrupamentos em base de dados com densidade variada. Para contornar esse problema foi proposto o *Varied Density Based Spatial Clustering of Applications with Noise* - VBDCAN (LIU; ZHOU; WU, 2007). O algoritmo constrói o gráfico *k-dist set*, que contempla um conjunto de pontos que representam as distâncias medidas de cada objeto até seu k-ésimo vizinho. Um exemplo de gráfico *k-dist* é ilustrado na Figura 13, no qual as marcações indicam os possíveis valores do Eps. A partir desse gráfico são obtidos os valores Eps. Para cada Eps o

VDBSCAN executa o algoritmo tradicional DBSCAN, o que possibilita encontrar agrupamentos em densidades variadas. O algoritmo VDBSCAN é robusto a ruído e consegue gerar agrupamento de formatos variados, no entanto requer como parâmetro de entrada o valor k (LIU; ZHOU; WU, 2007, CHOWDHURY, 2010).

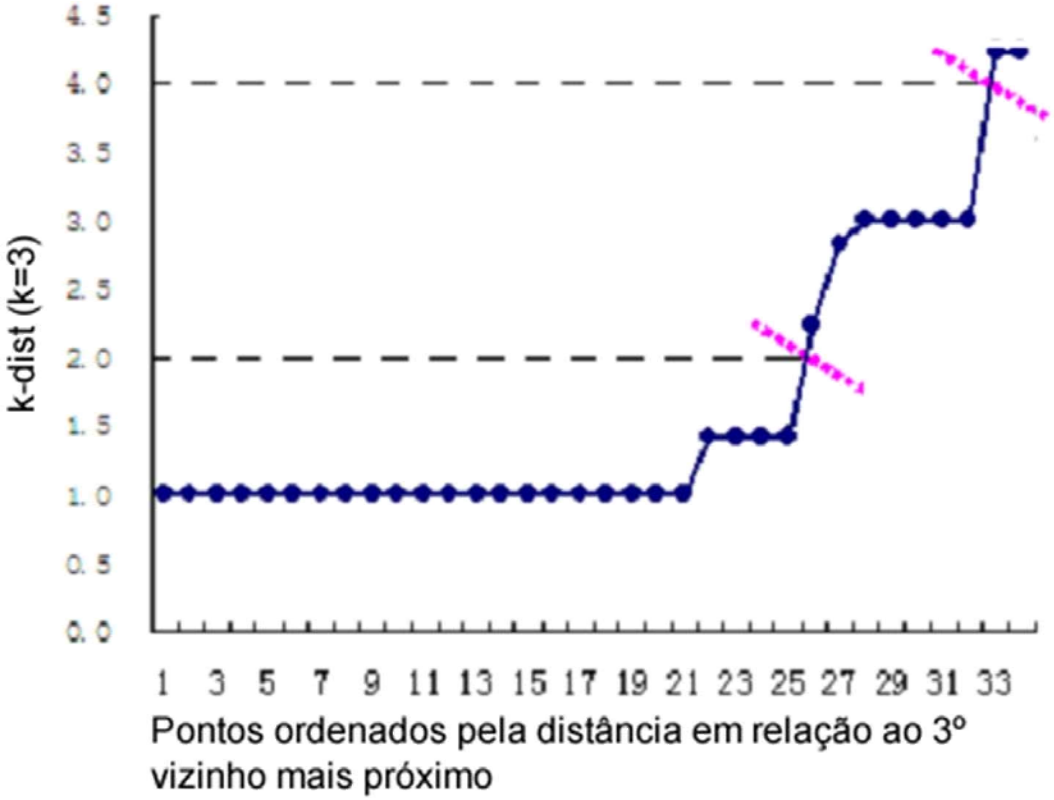

Figura 13: Gráfico *k-dist* adaptado de (LIU; ZHOU; WU, 2007).

O ADACLUS – *ADAptive CLUStering* é um algoritmo baseado em densidade, que constrói funções de influência adaptativas - funções de probabilidade para adaptar a agrupamentos de formatos e tamanhos variados e tratar pontos discrepantes. O algoritmo também possibilita descobrir agrupamentos de densidade variada, é insensível à ordem de entrada dos dados e é robusto a ruído. Possui três parâmetros intrínsecos: Smin, Smax e c. Os dois primeiros são utilizados para simular o comportamento do olho humano, que quanto mais pontos estão presentes, menos detalhes particulares são notados. Já o terceiro parâmetro é utilizado para manter a complexidade do algoritmo em O(n), uma vez que o número médio de pontos analisados durante o processo de adaptação é restrito em 100 pontos. Segundo os autores esses parâmetros possuem pouca influência na execução do algoritmo. No entanto, o algoritmo se mostra mais eficiente quando Smin=200 e Smax=2000 (NOSOVSKIY; LIU; SOURINA, 2008).

Os algoritmos de agrupamento baseado em densidade são eficientes apenas para conjunto de dados em que os agrupamentos são bastante separados e possuem densidades semelhantes. Além disso, tais algoritmos têm resultados satisfatórios apenas se os ruídos estiverem isolados dos outros dados espaciais (DENG et al., 2011).

3.5 Método Baseado em Grade

O método baseado em grade divide o espaço em células e gera os agrupamentos baseados nessa estrutura (MILLER; HAN, 2009). Para o método baseado em grade alguns dos algoritmos encontrados são: *Statical Information Grid-based method* - STING (WANG; YANG; MUNTZ, 1997), *WaveCluster* (SHEIKHOLESLAMI; CHATTERJEE; ZHANG, 1998) e o CLIQUE (AGRAWAL et al., 1998).

O algoritmo STING divide o espaço em células retangulares, formando uma estrutura hierárquica em que cada célula pode ser divida em sub-células, tal como é ilustrado na Figura 14.

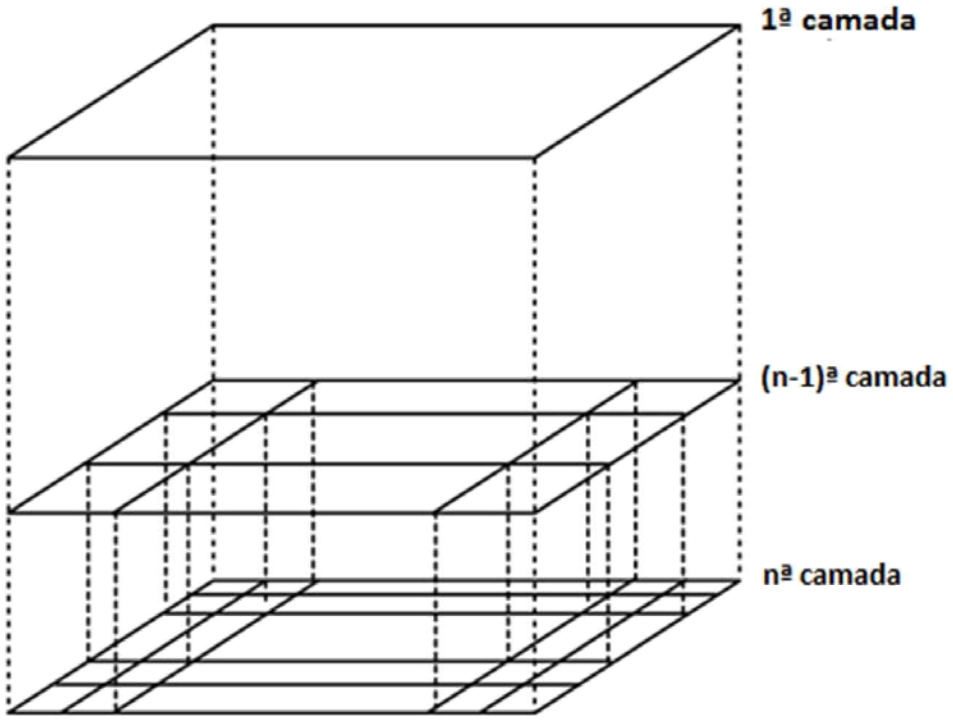

Figura 14: Divisão do espaço pelo método baseado em grade adaptado de (HAN; KAMBER, 2011).

O algoritmo STING é independe de consulta, pois cada célula armazena informações estatísticas sobre os objetos, tais como: média, valor mínimo e máximo para cada atributo. Além disso, tal algoritmo é favorável ao processamento paralelo e a atualização incremental (HAN; KAMBER, 2011).

O *WaveCluster* é um algoritmo baseado no método grade e densidade, que gera uma estrutura multidimensional no espaço de dados com o propósito de sumarizar as informações dos objetos. Em seguida, utiliza-se a transformação *wavelet* para identificar a densidade das regiões no espaço transformado. Tal algoritmo é eficiente para grandes bases de dados, encontra agrupamentos de formatos variados, não requer parâmetros de entrada, não é sensível à ordem de entrada dos dados e trata pontos discrepantes eficientemente. No entanto não é eficiente para conjuntos de dados com densidade variada (HAN; KAMBER, 2011, LEI et al., 2011).

O algoritmo CLIQUE (AGRAWAL et al., 1998) é baseado em grade e em densidade, e é capaz de obter agrupamentos em conjunto de dados com alta dimensionalidade. Para isso, o algoritmo identifica as grades densas de uma dimensão, posteriormente de duas dimensões, e assim por diante até que todas as grades de dimensão k sejam obtidas. Este algoritmo é capaz de gerar agrupamentos de formato e densidade variada, é insensível à ordem de entrada dos dados e é eficiente para o processamento de grandes dimensões de dados espaciais. No entanto, o CLIQUE requer o valor k como parâmetro de entrada, não é robusto a ruído e as grades podem comprometer os formatos dos agrupamentos. Além disso, quando o algoritmo constrói agrupamentos em todas as dimensões, muitas unidades podem ser geradas, comprometendo a eficiência e eficácia do algoritmo (AGRAWAL et al., 1998).

A principal vantagem do método baseado em grade é o rápido processamento dos dados, sendo que a velocidade não depende do número de objetos a serem processados e sim do número de células em que cada dimensão foi dividida (HAN; KAMBER, 2011).

3.6 Método Baseado em Grafos

Os grafos vêm sendo bastante utilizados para modelar estruturas complexas em diversas áreas. Assim, com o propósito de construir uma representação mais adequada dos dados, os grafos têm sido incluídos nos métodos de agrupamento de dados (HAN; KAMBER, 2011, ZHONG; MIAO; WANG, 2010).

O algoritmo AMOEBA (ESTIVILL-CASTRO; LEE, 2000a) estabelece uma área inicial, em que os vizinhos estão interligados e novas áreas são adicionadas até que a inclusão

não aumente a magnitude dos grafos locais. O procedimento é realizado para todas as áreas, e ao final são eliminadas as sobreposições de áreas, obtendo-se os agrupamentos espaciais. O AMOEBA não é eficiente para grandes bases de dados, uma vez que o tempo de processamento aumenta consideravelmente com o aumento dos dados a serem processados e o tamanho dos agrupamentos gerados (DUQUE et al., 2011).

O *Automatic Clustering via Boundary Extration for Mining Massive Point-Data Sets* - AUTOCLUST (ESTIVILL-CASTRO; LEE, 2000b) é semelhante ao AMOEBA, sendo que este incorpora variação local e global, o que permite tratar eficientemente ruídos, pontos discrepantes, pontes e distribuição variada. Além disso, esse algoritmo é capaz de gerar agrupamentos de formatos, tamanhos e densidades variadas e não requer parâmetros de entrada. No entanto, apresenta um tempo de processamento superior ao AMOEBA (ESTIVILL-CASTRO; LEE, 2002).

O algoritmo *Hierarchical Clustering Algorithm Using Dynamic Modeling* - Chameleon (GUHA; RASTOGI; SHIM, 1999) constrói um grafo com os k vizinhos mais próximos, e por meio da proximidade e interconectividade dos objetos realiza o agrupamento dos dados. Este algoritmo pode ser aplicado a qualquer tipo de dado, sendo possível aplicar funções de similaridade para considerar as características não espaciais na formação dos agrupamentos (HAN; KAMBER, 2011, ZHONG; MIAO; WANG, 2010). No entanto, o custo de processamento pode chegar a $O(n^2)$ no pior caso.

O algoritmo CHSMST é composto pelos algoritmos *Clustering based on Hyper Surface* - CHS e *Minimum Spanning Tree* - MST. O primeiro é um algoritmo baseado na classificação de hipersuperfícies e tem o propósito de realizar um pré-agrupamento dos dados. Já o segundo algoritmo é baseado no método grafos e realiza o refinamento dos agrupamentos encontrados (HE; ZHAO; SHI, 2011).

A classificação de hipersuperfícies é baseada no teorema das curvas de Jordan, cujo objetivo geral é dividir o plano em uma região limitada e outra ilimitada, em que a curva é a fronteira entre as duas regiões (HE; SHI; REN, 2002). De acordo com o teorema das curvas de Jordan, a base do algoritmo de classificação de hipersuperfícies pode ser representada conforme é ilustrado na Figura 15.

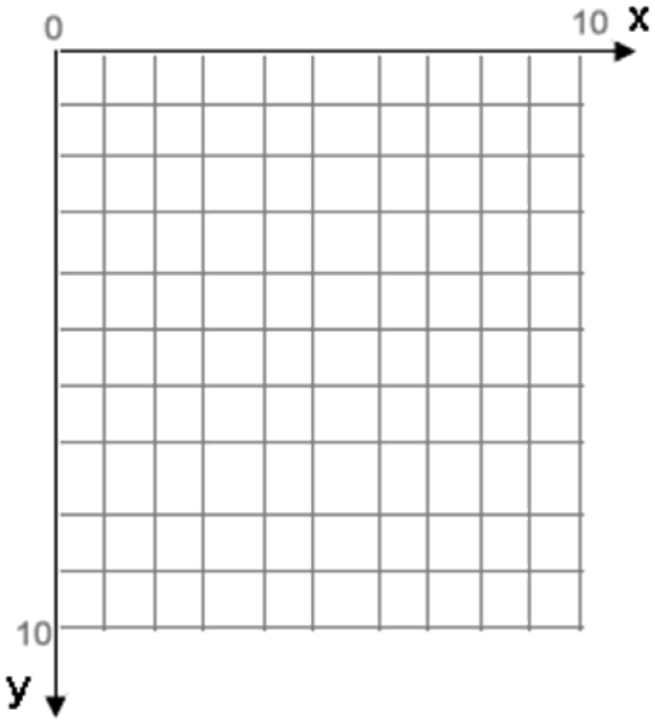

Figura 15: Representação da construção de hipersuperfícies (HE; SHI; REN, 2002).

O algoritmo MST destaca-se entre os algoritmos baseados em grafos devido a sua facilidade e eficiência na representação dos dados, bem como na aplicação do algoritmo em diversas áreas, tais como: análise de dados biológicos, processamento de imagens, reconhecimento de padrões entre outros.

Dado um grafo conexo ponderado G, uma árvore geradora mínima T consiste em um subgrafo conexo acíclico com o menor peso de G, que contém todos os vértices do grafo G, cujo peso da árvore é definido pela soma dos pesos das arestas pertencentes à mesma.

O CHSMST foi comparado com os algoritmos CLARANS, DBSCAN e CLIQUE, que são representantes dos métodos de particionamento, densidade e grade, respectivamente. Diferentemente do CLARANS, este algoritmo é capaz de gerar agrupamentos de formatos variados; agrupamentos em densidade variada; e é insensível à ordem de entrada dos dados. Já em relação ao CLIQUE, o algoritmo se mostrou superior em relação à robustez à presença de ruídos e para o DBSCAN foi superior em relação à insensibilidade à ordem de entrada dos dados. Os algoritmos CLARANS, DBSCAN e CLIQUE possuem complexidade $O(n^2)$ enquanto que o algoritmo CHSMST possui complexidade $O(n)$, além disso, os autores apresentam vários testes que mostram que o referido algoritmo é executado em tempo inferior que os demais. Neste cenário o CHSMST se mostra superior tanto em termos de eficácia quanto de eficiência em relação aos algoritmos CLARANS, DBSCAN e CLIQUE (HE; ZHAO; SHI, 2011).

3.7 Estudo Comparativo entre os Algoritmos de Agrupamentos de Dados Espaciais

Com parte deste trabalho, foi realizado um comparativo sobre os algoritmos de agrupamento de dados espaciais com mais de 30 algoritmos, que envolveu trabalhos do estado da arte, bem como algoritmos consagrados na literatura. Durante este levantamento, foi verificado que algumas características eram frequentemente citadas como sendo importantes para que os algoritmos fossem capazes de gerar resultados satisfatórios, sendo que algumas foram mencionadas em todos os trabalhos comparativos, a saber: gerar agrupamentos com formatos variados; robusto a ruídos; gerar agrupamentos com densidade variada; e não necessitar de parâmetros de entrada. A partir desse levantamento, foram selecionados os 14 algoritmos mais relevantes em relação às características analisadas e o resultado é apresentado na Tabela 2.

Tabela 2: Comparativo de algoritmos (PATTABIRAMAN et al., 2009, DENG et al., 2011, XIAO-PING; ZHENG-YUAN; JAIN-HUA, 2010, XI, 2009, ANDREOPOULOS et al., 2009, PETER; ANTONYSAMY, 2010, LIU et al., 2012, SONG; JIN; SHEN, 2011, NOSOVSKIY; LIU; SOURINA, 2008, PRASAD; SARMAH, 2011).

Algoritmo	CHSMST	ADACLUS	VDBSCAN	WAVECLUSTER	STING	CLIQUE	DENCLUE	OPTICS	MST	DBSCAN	BIRCH	CURE	SINGLE LINK	CLARANS
Formatos Variados	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗
Densidade Variada	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Robusto a Ruído	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓	✓	✓	✗	✓
Considera a Similaridade	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Sem Parâmetros de Entrada	✗	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗
Insensível à Ordem dos Dados	✓	✓	✗	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✗
Suporta a Alta Dimensionalidade	✓	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗

A análise dos algoritmos apresentados na Tabela 2 revela que nenhum desses algoritmos contempla todo o conjunto de características analisadas. A maioria consegue tratar agrupamentos de formatos variados, mas somente CHSMST, ADACLUS, VDBSCAN e CLIQUE

conseguem resolver completamente o problema das densidades variadas. E por fim, a maioria requer a interação do usuário por meio de parâmetros de entrada. Diante disso, é possível verificar que o algoritmo CHSMST e o ADACLUS atendem a maioria das características analisadas, no entanto como o CHSMST é mais recente, este algoritmo foi selecionado para ser a base deste trabalho.

3.8 Considerações Finais

Nesse capítulo foram apresentados os principais conceitos no contexto de agrupamento de dados espaciais. Foi realizado um estudo comparativo entre vários algoritmos de agrupamentos de dados espaciais, que resultou na seleção do algoritmo CHSMST para ser à base deste trabalho. A partir disso, foram elaboradas estratégias para aprimorar os resultados e o desempenho do mesmo, o que resultou em um novo algoritmo denominado CHSMST^{+}, que é apresentado no Capítulo 4.

CAPÍTULO 4
Extração de Conhecimento em Bases de Dados Espaciais: Algoritmo CHSMST^{+}

4.1 Considerações Iniciais

Neste capítulo, o algoritmo desenvolvido, denominado CHSMST^{+}, é descrito desde a sua concepção à sua implementação, que consiste no aprimoramento dos resultados e desempenho realizado no algoritmo CHSMST, selecionado no estudo comparativo por atender a maioria das características analisadas e ser o mais recente entre os algoritmos estudados. Para isso, incialmente o funcionamento do algoritmo CHSMST é detalhado por meio de pseudocódigos e fluxograma. Em seguida, as estratégias para aprimoramento dos resultados e desempenho do algoritmo são detalhadas, sendo apresentados os pseudocódigos dos principais métodos desenvolvidos no CHSMST^{+} e com o auxílio de um fluxograma é possível identificar as modificações realizadas no algoritmo original. Ao final, é apresentada a estratégia para verificar a eficiência dos agrupamentos gerados e a ferramenta desenvolvida para suportar a aplicação do algoritmo desenvolvido em diferentes bases de dados.

4.2 O Algoritmo CHSMST

Conforme apresentado na Seção 3.6, o algoritmo CHSMST é composto pelos algoritmos CHS e MST. O algoritmo CHS inicialmente considera todos os pontos pertencentes a uma mesma classe. Posteriormente, o espaço é divido em hipersuperfícies, sendo que cada hipersuperfície contempla um conjunto de amostras, que correspondem aos agrupamentos iniciais. Apesar do algoritmo não requerer um conhecimento prévio para definir o número de agrupamentos, é necessário estabelecer o valor d, utilizado para definir o número de divisões que serão aplicadas ao espaço. O algoritmo CHS possui os seguintes passos (HE; ZHAO; SHI, 2011):

1. Realizar o pré-processamento de todas as amostras com d divisões;

2. Distribuir as amostras fornecidas no interior de um retângulo;

3. Dividir a região em $\overbrace{10 \times 10 \times \cdots \times 10}^{d}$ regiões menores, denominadas unidades;

4. Para cada unidade, se houver uma ou mais amostras, armazenar a fronteira da região como uma *string;*

5. Combinar as unidades adjacentes dos quais existem uma ou mais amostras e reorganizar as fronteiras das regiões;

6. Definir as amostras delimitadas pela mesma fronteira como um agrupamento.

O algoritmo MST pode ser construído de diferentes formas, sendo que as implementações mais utilizadas são baseadas no algoritmo de Prim (PRIM, 1957) ou no algoritmo de Kruskal (GOODMAN; KRUSKAL, 1954). Uma forma de implementar o algoritmo MST é descrita a seguir (HE; ZHAO; SHI, 2011):

1. Construir um grafo completo não direcionado ponderado G = (V, E), onde V é o conjunto de amostras em D e o peso de cada aresta E é a distância Euclidiana entre os dois vértices da aresta;

2. Aplicar o algoritmo Prim para obter a árvore geradora mínima T;

3. Excluir a (k-1) aresta com maior peso em T, e então tem-se um gráfico desconexo G';

4. Buscar na primeira profundidade de G, para obter k ramos conexos de G';

5. A amostra em cada ramo conexo forma um agrupamento, que são exatamente os agrupamentos contidos em D.

O fluxograma do algoritmo CHSMST é apresentado na Figura 16, no qual observa-se que após a aplicação do CHS, é exigida uma interação com o usuário, em que o mesmo deve identificar visualmente se os agrupamentos iniciais gerados pelo CHS necessitam ser subdivididos e quantas subdivisões são necessárias. Em caso positivo, o algoritmo MST é executado com o número de subdivisões estabelecido.

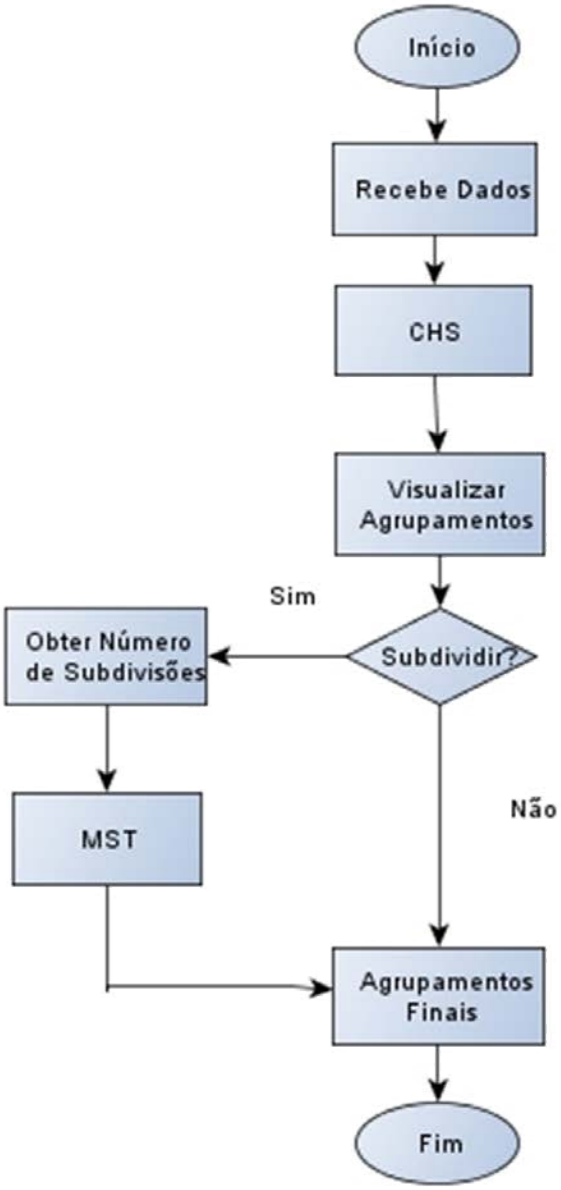

Figura 16: Fluxograma do algoritmo CHSMST.

Na Figura 17(a) é apresentado um conjunto de dados espaciais, no qual foi aplicado o algoritmo CHSMST. O resultado obtido após a fase de aplicação do algoritmo CHS é mostrado na Figura 17(b), em que é possível observar que foram gerados três agrupamentos de dados espaciais, no entanto visualmente é identificado que o correto seria gerar cinco agrupamentos de dados espaciais. Somente após a aplicação do MST, o algoritmo conseguiu identificar corretamente os agrupamentos de dados espaciais, conforme é verificado na Figura 17 (c).

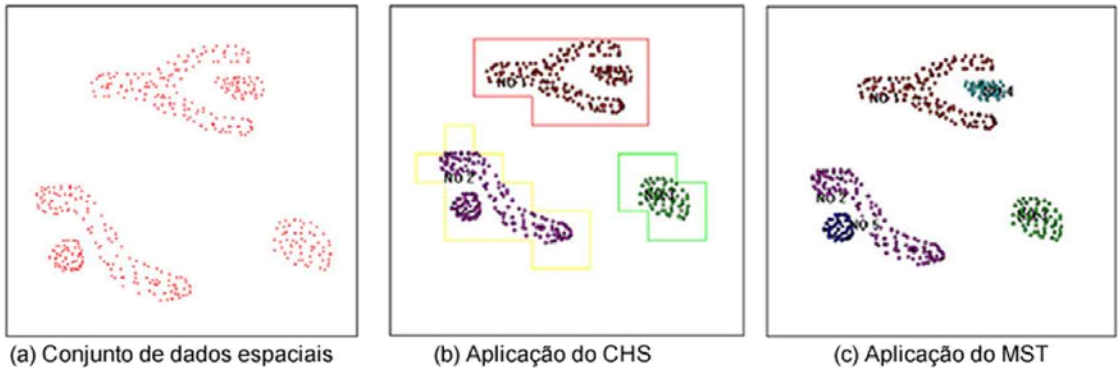

Figura 17: Agrupamentos de dados espaciais gerados com a aplicação do algoritmo CHSMST adaptado de (HE; ZHAO; SHI, 2011).

4.3 O Trabalho Desenvolvido

O algoritmo CHSMST requer parâmetro de entrada, o que implica que os resultados do algoritmo serão influenciados pelo parâmetro que o usuário definirá, sendo que a dificuldade está em estabelecer uma métrica para definir qual parâmetro deve ser utilizado para que o algoritmo possa obter os melhores resultados. Além disso, o algoritmo exige que o usuário visualize os pré-agrupamentos e defina quais agrupamentos devem ser gerados, o que amplia a dependência do conhecimento humano na geração dos resultados.

Outro problema do CHSMST é não considerar a similaridade dos atributos não espaciais na geração dos agrupamentos, assim apenas os atributos espaciais são considerados, fazendo que a análise dos agrupamentos seja restringida apenas à questão de proximidade entre os objetos, ignorando outras características que os mesmos possam apresentar em comum.

Neste contexto este trabalho visa atacar os problemas encontrados no CHSMST para aprimorar os resultados do algoritmo, ao mesmo tempo em que busca ampliar o desempenho do CHSMST. Com as melhorias realizadas o algoritmo passa a ser denominado CHSMST^{+}, sendo que modificações realizadas são detalhadas a seguir.

4.3.1 A Estratégia para aprimoramento dos resultados

A primeira iniciativa, no sentido de promover algumas facilidades junto ao algoritmo CHSMST, foi eliminar o parâmetro de entrada. No CHSMST, as unidades são construídas e o usuário tem que decidir se é necessário subdividi-las em unidades menores por meio da visualização dos resultados. Para contornar esta etapa, foi introduzido o cálculo das distâncias local e global, no qual é calculada a distância media dos objetos em cada unidade, e posteriormente

obtém-se a média dessas distâncias para todas as unidades. Assim, o algoritmo subdividirá uma unidade se a distância local for maior que a distância global, o que permite reduzir a espacialidade dos dados dentro das unidades.

A distância entre os objetos é calculada por meio das coordenadas x e y. Essas coordenadas são consideradas variáveis de escala-intervalo, sendo assim pode-se calcular a distância entre dois objetos por meio da distância Euclidiana, Manhattan ou Minkowski (HAN; KAMBER, 2011). No algoritmo CHSMST é utilizada a distância Euclidiana, apresentada na Equação (1).

$$d(pi, pj) = \sqrt{(pix - pjx)^2 + (piy - pjy)^2} \quad (1)$$

É importante salientar que as coordenadas x e y normalmente são referentes à latitude e longitude, no entanto a distância Euclidiana também pode ser calculada para outros sistemas de coordenadas. Um exemplo disso é um trabalho que calcula a posição dos genomas responsáveis por determinada função biológica no DNA ou RNA, em que a coordenada x representa a posição no cromossomo, enquanto que a coordenada y é referente ao número do cromossomo (ROQUEIRO; FRASOR; DAI, 2010). Assim, verifica-se que o algoritmo CHSMST$^+$ não é restrito a aplicações em dados geográficos.

Para obter o agrupamento dos dados apenas a localização espacial dos objetos é considerada no CHSMST. Existem aplicações que com apenas a proximidade entre os objetos é possível obter informações relevantes para os pesquisadores, como por exemplo, na detecção de agrupamentos de terremotos, no qual a partir da ocorrência de vários terremotos de baixa intensidade é possível prever a ocorrência de um terremoto de maior intensidade próximo à ocorrência dos demais. No entanto, para a maioria das aplicações reais verifica-se a necessidade de se obter também o nível de similaridade entre os objetos dentro de um agrupamento para extrair conhecimento útil (HAN; KAMBER, 2011). Diante deste cenário, no algoritmo desenvolvido optou-se em contemplar tanto a proximidade quanto a similaridade entre os objetos para a geração dos agrupamentos.

Um agrupamento de dados pode conter objetos com vários tipos de atributos: inteiros, reais, booleanos e cadeias de caracteres. Por outro lado, os valores dos atributos podem ser do tipo: escala-intervalo, tais como latitude, longitude, peso e altura; valores binários; valores categóricos, por exemplo, cores, classes e etnias; valores ordinários: que podem ser 1, 2, 3 ou ouro, prata e bronze; valores de escala-variável tais como crescimento da população de bacté-

rias e queda de radioatividade. Existem fórmulas específicas para o cálculo de cada um destes tipos de valores, como também existe uma fórmula para o cálculo de tipos mistos de valores, apresentada a seguir (HAN; KAMBER, 2011):

$$d(i, j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}, \tag{2}$$

Em que $\delta_{ij}^{(f)} = 0$ se o valor da variável x do objeto i ou do objeto j é nula, se a variável f é uma binária assimétrica ou se ambas são iguais a 0, caso contrário o $\delta_{ij}^{(f)} = 1$. Já o cálculo de $d_{ij}^{(f)}$ depende do tipo de valor da variável, para aplicar a fórmula específica para valores: escala-intervalo, categóricos, binários, ordinários e de escala-variável (HAN; KAMBER, 2011). Assim, a utilização da Equação (2) exige que o algoritmo seja aplicado em situações específicas ou seria necessária a interação humana para definir o tipo de valor para cada atributo.

O cálculo da similaridade entre valores escala-intervalo pode ser feita por meio da Euclidiana, Manhattan ou Minkowski. Para a aplicação destas equações para qualquer valor de atributo seria necessário discretizar os valores dos atributos para tipo inteiro ou real. Como exemplo, tem-se a discretização para um atributo etnia que contempla os valores: branco, pardo, amarelo e negro, que poderia resultar em: 1, 2, 3 e 4 respectivamente. Para um conjunto de dados com vários atributos, a realização desta tarefa dificultaria a análise dos resultados pelos usuários, uma vez que seria necessário fazer a correspondência dos resultados obtidos com os valores originais dos atributos.

Além dos problemas apresentados para a aplicação de forma genérica das equações de valores escala-intervalo, nos casos de valores ordinários e de escala-variável esta estratégia é ainda menos vantajosa, uma vez que estes tipos de valores não são muito frequentes na maioria das bases de dados.

Por outro lado, no cálculo da similaridade para valores categóricos ou binários, são verificadas as coincidências dos atributos dos objetos i e j. Na Equação (3) é apresentada uma equação para o cálculo da similaridade para valores categóricos ou binários.

$$s(i, j) = 1 - \frac{p - m}{p} \tag{3}$$

Onde **p** é o número total de atributos não espaciais e **m** é o número de coincidências. Assim a Equação (3) pode ser aplicada para qualquer tipo de dado (inteiros, reais, booleanos e cadeias de caracteres), sem a necessidade de discretizá-los.

Diante do exposto, no algoritmo CHSMST$^+$ optou-se pela utilização da Equação (3) para o cálculo da similaridade durante o cálculo da árvore geradora mínima, bem como durante o processo de poda da árvore. Com esta modificação, a similaridade entre dois objetos passa a influenciar a construção da árvore geradora mínima tanto quanto a distância entre as extremidades da aresta. Já na etapa de poda, as arestas com a maior distância entre os pontos da árvore ou com similaridade inferior à mínima são removidas, com o propósito de gerar agrupamentos com objetos mais similares, o que facilita na identificação do padrão das características apresentadas pelos registros pertencentes à região do agrupamento.

4.3.2 Otimização do algoritmo CHSMST$^+$

No sentido de proporcionar avanços na capacidade de processamento do algoritmo CHSMST$^+$, é proposta alterações que o capacita na execução com o suporte da técnica *multithreading*. Tal estratégia foi possível porque a construção de hipersuperfícies pelo algoritmo CHS facilita a incorporação de tal técnica, uma vez que ao dividir o espaço em unidades, todo o processamento em cada unidade é realizado de forma independente. Além disso, como as unidades também podem ser subdivididas em regiões menores, a própria subdivisão pôde ser realizada paralelamente.

4.3.3 O algoritmo desenvolvido CHSMST$^+$

Com as modificações realizadas, o algoritmo passa a requer apenas o conjunto de dados, não necessitando de nenhum parâmetro definido pelo o usuário para execução do algoritmo, como pode ser observado a seguir.

As etapas de distribuição dos pontos no retângulo inicial e subdivisão do mesmo em unidades menores no CHSMST$^+$ são realizadas pelo algoritmo CHS, apresentado a seguir.

Como é possível verificar a divisão das unidades em regiões menores é realizada por meio de *threads* de forma assíncrona, sendo que o número de subdivisões é definido automaticamente por meio do cálculo da distância local e global. Ao final da execução das *threads* são obtidas as unidades iniciais, e então realiza-se o processo de junção das unidades adjacen-

tes com o propósito de contemplar em uma única unidade as respectivas amostras. Após esta etapa as amostras pertencentes a unidades adjacentes são agrupadas e uma nova unidade é gerada contemplando as fronteiras das unidades anteriores. A seguir o algoritmo *Combinar_Adjacentes* é apresentado.

Algoritmo CHSMST^{+}

Entradas: D – Conjunto de dados.

Procedimento:

Receber o conjunto de dados *D*

Distribuir *D* em um retângulo *R*
Aplicar *CHS* para o retângulo *R*
Aplicar *Combinar_Adjacentes* para *Unidades*
Para cada unidade *i* no conjunto unidades *U*
    Aplicar o algoritmo MST para unidade *i*

Saídas: Agrupamentos finais *AF*

Algoritmo CHS

Entradas: R – retângulo.

Procedimento:

Dividir o retângulo *R* em 10 unidades

Calcular a distância global dos pontos *dist_global*

Para cada *i* unidade pertencente ao conjunto de unidades *uni*

    Para cada *p* processador

        Buscar processador livre

        Se encontrar processador livre

            Disparar *thread*

                Calcular a distância local dos pontos *dist_local*

                Se *dist_local*>*dist_global*

                  Aplicar *SubRetangulos* para dividi-lo em 10 unidades

            Libera *thread*

Saídas: Conjunto de unidades *uni*

AlgoritmoCombinar_Adjacentes

Entradas: *uni* – conjunto de unidades

Procedimento:

Para cada *i* unidade pertencente a *uni*

Se existir ao menos um vértice igual a algum vértice de *v*

Combinar as unidades adjacentes

Rearranjar as fronteiras das unidades

Saídas: Conjunto de unidades rearranjadas *uni_adj*

Após a distribuição dos dados em unidades, realiza-se a aplicação do algoritmo MST, no qual a árvore MST é obtida para cada unidade por meio do algoritmo Calcular_MST. As subdivisões dentro da unidade são realizadas por meio do processo de poda, por meio do algoritmo *Podar_MST*. A poda consiste em eliminar as arestas da MST que possuam as maiores distâncias ou similaridade seja inferior à similaridade mínima. Com isso, são geradas subárvores que correspondem aos agrupamentos presentes em cada unidade.

Algoritmo MST

Entradas: *uni* – conjunto de unidades

Procedimento:

Para cada *i* unidade pertencentes a *uni*

Para cada *p* processador

Buscar processador livre

Se encontrar processador livre

Disparar *thread*

Aplicar *Calcular_MST* para obter a MST da unidade i

Aplicar *Podar_MST* para gerar as subárvores

Rearranjar as subárvores

Libera *thread*

Saídas: Conjunto de agrupamentos finais *agrupa_final*

Algoritmo Calcular_MST

Entradas: *C* – Conjunto de Pontos

Procedimento:

Para cada ponto *p1* pertencente a C

    Para cada ponto *p2* pertencente a C

        dist=Distancia(p1,p2);

        simi=Similaridade(p1,p2);

        Se *dist<menor_dist*

            menor_dist= dist

            p=p2;

        Se *simi>maior_simi*

            maior_simi=simi

            p=p2;

    a= Criar_Aresta(p1,p);

    Add_MST(a);

Saídas: Retorna a árvore *MST*

Algoritmo Podar_MST

Entradas: *MST* – árvore MST

Procedimento:

Calcular a *dist_max* distância máxima entre os vértices da *MST*

Calcular a *simi_min* similaridade mínima entre os vértices da *MST*

Para cada aresta *a* pertencente a *MST*

    Se o peso.distância de *a* > *dist_max* ou peso.similaridadede *a* >*simi_min*

        Remover a aresta *a*

        Definir uma nova subárvore *subarv* para um dos vértices

Saídas: *MST* podada

Com o objetivo de possibilitar um melhor entendimento sobre o CHMST^{+}, na Figura 18 é apresentado o respectivo fluxograma, em que as áreas destacadas em amarelo representam a inclusão ou modificação relevante de um processo com relação ao CHSMST original.

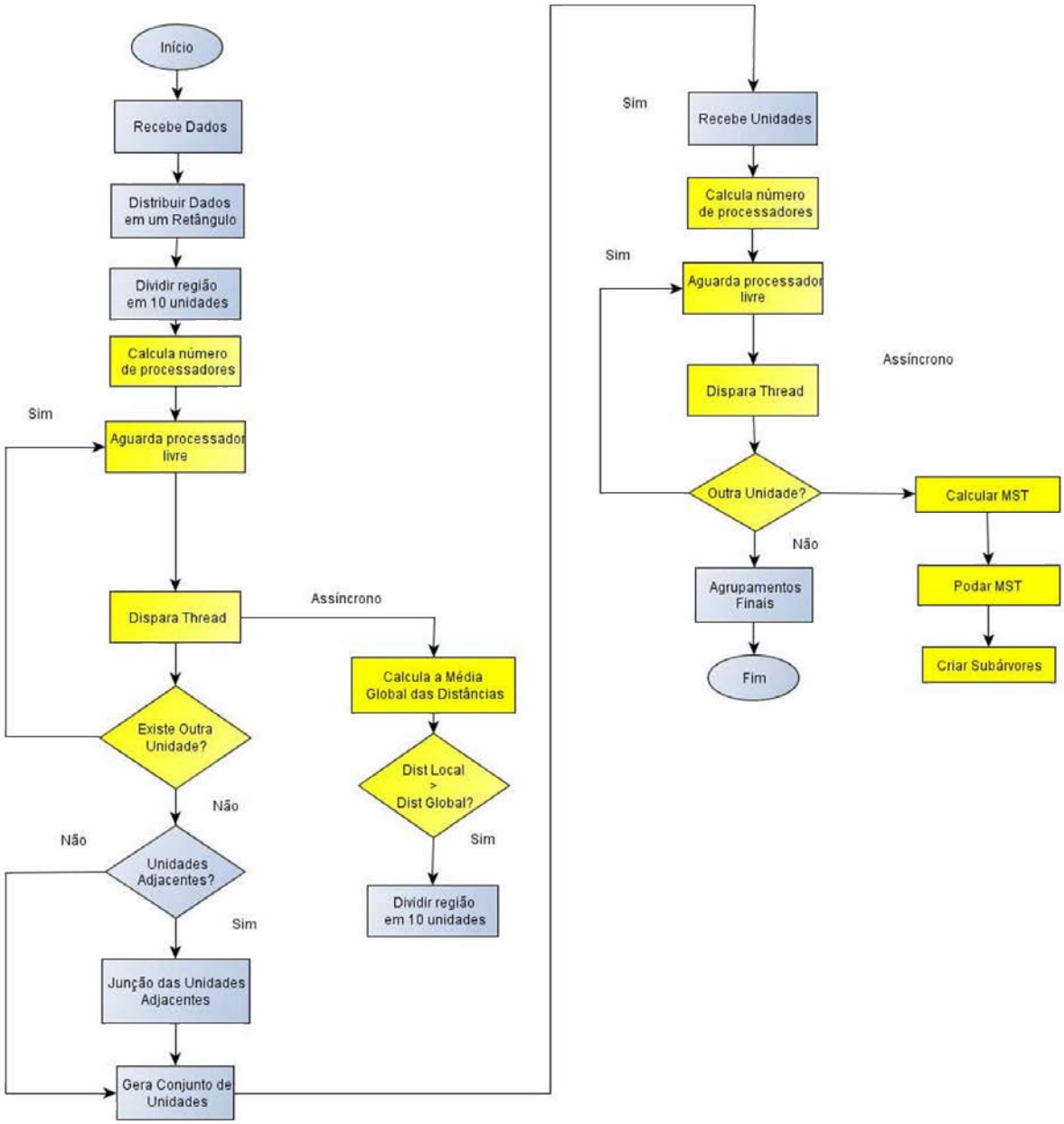

Figura 18: Fluxograma do algoritmo CHSMST^{+}.

4.3.4 Qualidade dos agrupamentos

Uma estratégia para medir a qualidade dos agrupamentos gerados e verificar se os resultados obtidos são satisfatórios é submeter os agrupamentos à análise de um especialista da área correspondente ao conjunto de dados analisados, uma vez que este estará habilitado para julgar se os resultados obtidos são relevantes para a respectiva área. Este trabalho pode ser realizado com o suporte de diversos métodos denominados como extrínsecos, que medem o quanto um agrupamento está próximo do agrupamento ideal, definido pelo especialista. No entanto, quando não é possível obter o suporte de um especialista para verificar a eficiência dos agrupamentos deve-se utilizar os métodos intrínsecos, que baseado na premissa de que os

agrupamentos devem conter dados homogêneos e devem ser heterogêneos entre si, estes métodos medem a compactação de cada agrupamento e sua dissimilaridade em relação aos demais (HAN; KAMBER, 2011).

Um dos métodos intrínsecos mais utilizados é o coeficiente de silhueta que consiste em calcular para cada objeto **o** do conjunto de dados: **a(o)**, que corresponde à média da distância do objeto **o** para os demais objetos pertencentes ao agrupamento $\mathbf{C_i}$, sendo que $C_i \in C=\{ C_1, C_2, C_3, C_k\}$ e **b(o)**, que representa a distância média mínima de **o** para todos os outros agrupamentos, conforme verificado nas Equação (4) e na Equação (5). Finalmente, coeficiente de silhueta é definido por **s(o)** apresentado pela Equação (6) (HAN; KAMBER, 2011).

$$\boldsymbol{a(o)} = \frac{\sum_{o' \in Ci\ o \neq o'} \boldsymbol{dist(o,o')}}{\boldsymbol{|C_i|-1}} \tag{4}$$

$$\mathbf{b}(\boldsymbol{o}) = min_{C_{j:1 \leq j \leq k, i \neq j}} \left\{ \frac{\sum_{o' \in Cj} \boldsymbol{dist(o,o')}}{\boldsymbol{|C_j|}} \right\} \tag{5}$$

$$\boldsymbol{s(o)} = \frac{b(o)-a(o)}{max\{a(o),b(o)\}} \tag{6}$$

O item **a(o)** é referente à compactação do agrupamento $\mathbf{C_i}$ e o item **b(o)** corresponde à dissimilaridade de $\mathbf{C_i}$ para os demais agrupamentos. Os valores do coeficiente de silhueta **s(o)** para cada agrupamento podem variar entre -1 e 1. Quando o valor de **s(o)** é negativo significa que o agrupamento $\mathbf{C_i}$ é similar aos demais agrupamentos obtidos, enquanto que se o valor de **s(o)** é próximo de 1 significa que o agrupamento $\mathbf{C_i}$ é compacto e dissimilar em relação aos demais agrupamentos, e portanto a qualidade do agrupamento neste caso é alta (HAN; KAMBER, 2011). Neste trabalho o coeficiente de silhueta foi utilizado para medir a qualidade dos agrupamentos gerados tanto em termos de distância e similaridade, sendo que a distância é calculada por meio da Equação Euclidiana e a similaridade por meio da Equação (3).

4.3.5 Ferramenta para agrupamento de dados espaciais

Para suportar a extração de conhecimento foi construída uma ferramenta *Web* denominada Ferramenta para Agrupamento de Dados Espaciais - FADE para execução do algoritmo $CHSMST^{+}$, bem como para visualização dos resultados obtidos. Nesta seção são apresentadas as tecnologias utilizadas para construção da ferramenta.

4.3.5.1 Tecnologias

Para o desenvolvimento da ferramenta FADE foi utilizada a linguagem Java devido à sua portabilidade e capacidade de conexão com os principais SGBD's (ORACLE, 2013a), *Java Server Page* - JSP e *Servlets* (ORACLE, 2013b). Também foi utilizado o framework ExtJS, que disponibiliza um conjunto de componentes, tais como *grids,* menus, árvores e formulários, que facilitaram o desenvolvimento da interface *Web* (SENCHA, 2013). Como ambiente de desenvolvimento foi utilizado o NetBeans 7.

O desenvolvimento de sistemas computacionais que envolvem a manipulação e análise de dados espaciais normalmente é realizado segundo nos padrões definidos pelo *Open Geospatial Consortium* - OGC, por meio da utilização dos serviços: *Web Map Service* - WMS, que gera mapas como imagens ou no formato vetorial, permitindo incorporá-los na interface *Web* e *Web Feature Service* - WFS, que permite a manipulação de dados espaciais por meio de imagens de satélite, fotos aéreas, dados digitais de elevação entre outros (BOYD; FOODY, 2011, HEN; LI; ZHAN, 2012, ERL, 2005). Para isso, é necessário o suporte de um conjunto de tecnologias, que no caso ferramenta FADE foram: *Geoserver,* um servidor de código aberto, que é baseado nos padrões OGC, e possui uma interface gráfica que permite projetar os dados espaciais com facilidade (GEOSERVER, 2013); *Openlayers*, uma biblioteca *JavaScript* de código aberto para a visualização de mapas (OPENLAYERS, 2013) e Google *Maps* (GOOGLE, 2013), uma *Application Program Interface* - API *JavaScript* para geração de mapas (ZHANG et al., 2010).

Como sistema gerenciador de banco de dados foi utilizado o PostgreSQL, um SGBD objeto-relacional de código aberto que trabalha com alta confiabilidade e integridade dos dados (POSTGRESQL, 2013). O PostgreSQL possui uma extensão denominada PostGIS que permite o tratamento de dados espaciais com maior eficiência (REFRACTIONS RESEARCH, 2013). Além disso, esse SGBD é compatível com a maioria dos sistemas operacionais e possui interface de programação nativa para diversas linguagens, incluindo Java (POSTGRESQL, 2013).

4.3.5.2 Arquitetura

Na interface *Web* da ferramenta FADE são configuradas as informações a respeito da tabela alvo, na qual o algoritmo CHSMST^{+} será executado, bem como são definidos os dados

espaciais e não espaciais que serão utilizados como critérios para geração dos agrupamentos. A partir disso, o algoritmo é executado, e os agrupamentos obtidos são armazenados na base de dados espaciais. Para facilitar a análise dos resultados obtidos, os objetos são projetados em formatos de pontos e os agrupamentos em formato de polígonos, por meio do *Geoserver*. As projeções obtidas são incorporadas na ferramenta junto com um mapa obtido do Google *Maps* por meio da biblioteca *OpenLayers*. A arquitetura da ferramenta de apoio ao agrupamento de dados espaciais é exibida na Figura 19.

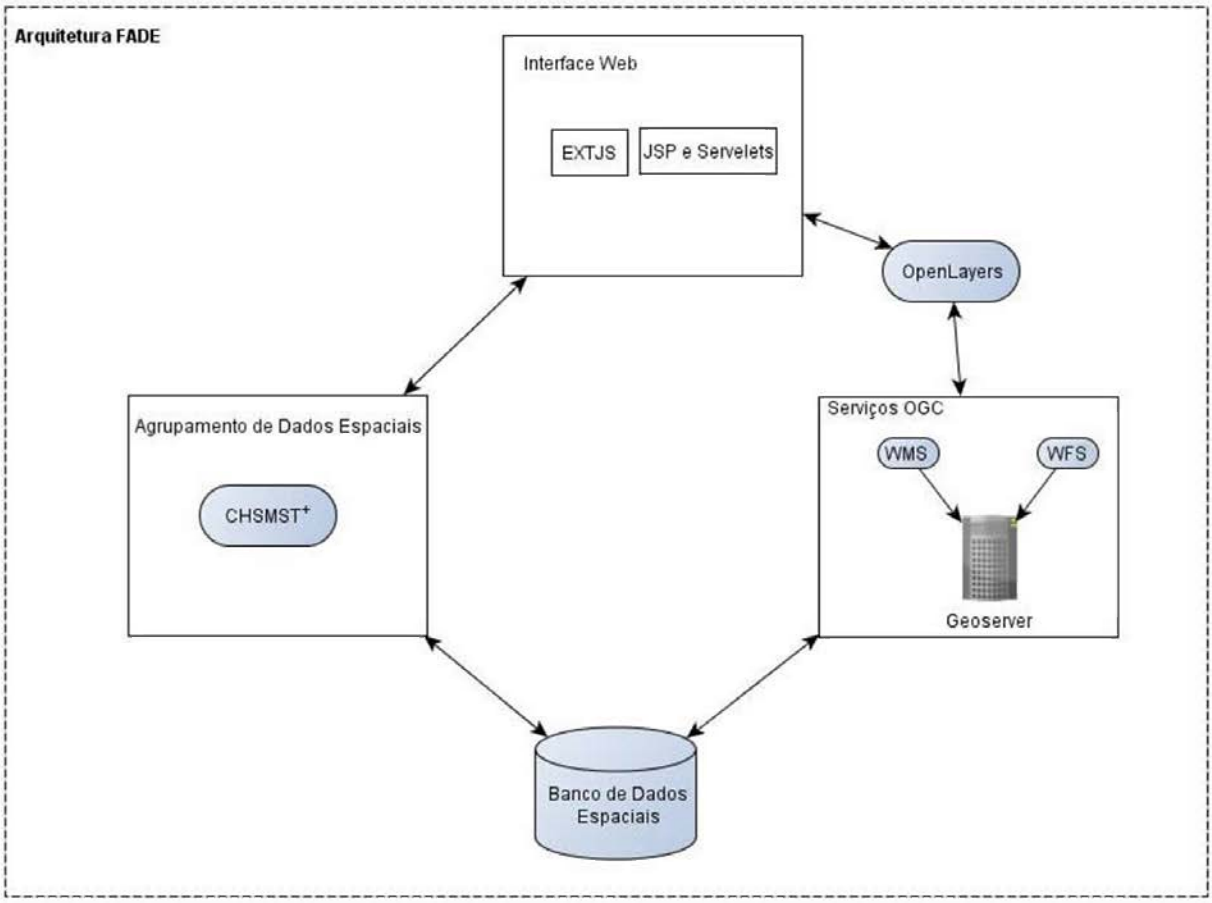

Figura 19: Arquitetura da Ferramenta.

4.3.5.3 Ferramenta

A ferramenta FADE possui interfaces para configuração da tabela alvo para execução do algoritmo CHSMST^{+}, bem como para visualização dos resultados obtidos. Na Figura 20 é apresentada a interface de conexão com a base de dados e na Figura 21 é mostrada a interface para configuração da tabela alvo com a seleção dos critérios para execução do algoritmo.

Figura 20: Conexão com a base de dados.

Figura 21: Configuração da tabela alvo.

Os agrupamentos de dados são realizados por meio da definição dos critérios de agrupamento. Dessa forma, o algoritmo procura agrupar pontos que sejam semelhantes segundo esses critérios e próximos entre si. Na Figura 22 é apresentada a interface para visualização dos agrupamentos. No exemplo, foi selecionado um dos agrupamentos gerados pelo CHSMST+ a partir de uma base que contempla registros de acidentes de trabalho.

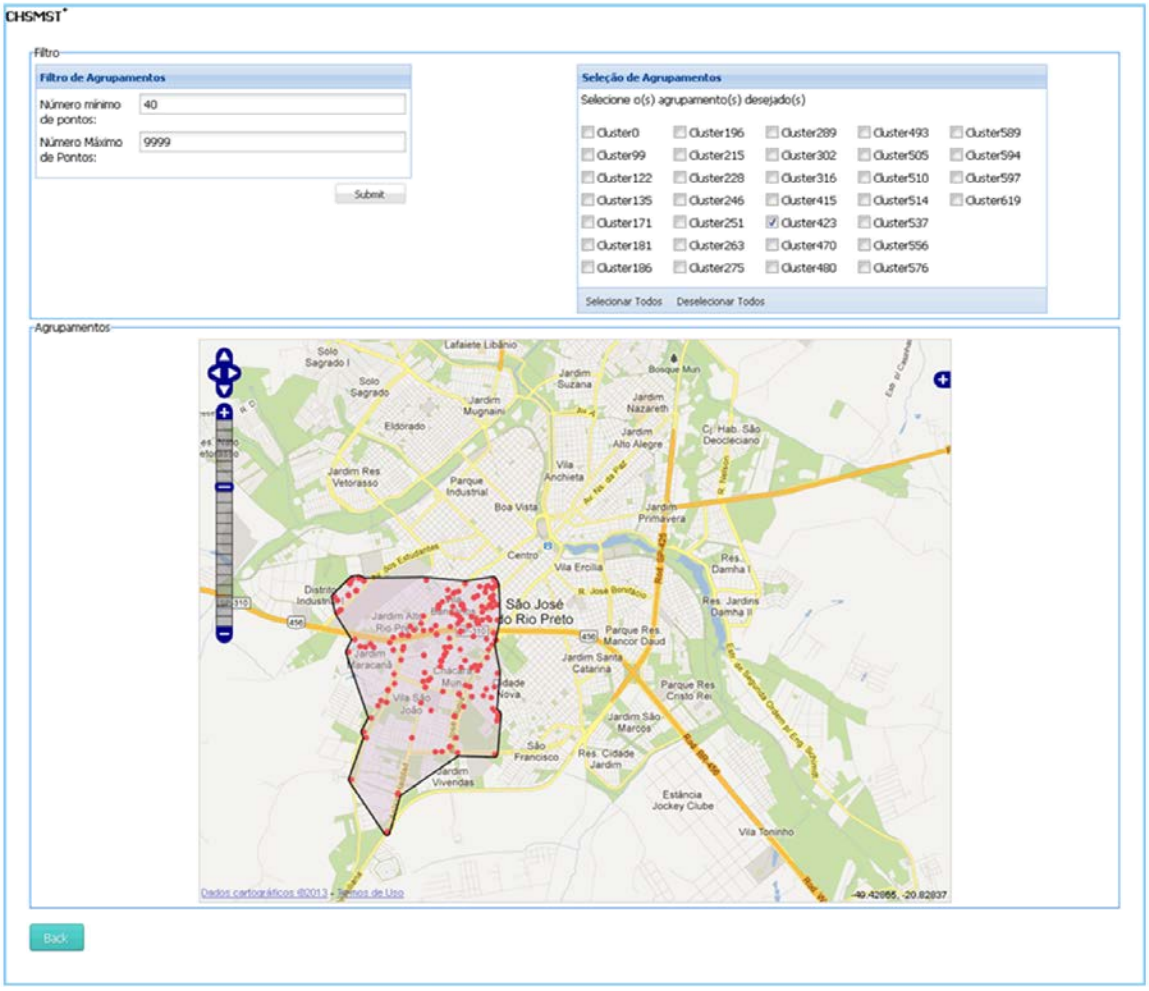

Figura 22: Interface para visualização dos agrupamentos.

Quando o usuário visualiza um polígono, não consegue entender qual a similaridade dos pontos que constituem esse polígono em relação aos dados não espaciais, selecionados para geração dos agrupamentos. Para suprir essa necessidade, são disponibilizados relatórios que apresentam estatísticas sobre cada agrupamento gerado, tais como número de objetos no agrupamento, qualidade em relação à distância e em relação à similaridade. Na Figura 23 é apresentado um exemplo de relatório gerado pela ferramenta, no qual é possível verificar que o agrupamento apresentado na Figura 22 contempla102 acidentes de trabalho, cujo coeficiente de silhueta em relação à distância é 1 e em relação à similaridade é 0.960. Verifica-se também que a maioria desses acidentes gerou afastamento do trabalho e foram típicos.

Figura 23: Relatório gerado pela ferramenta.

4.4 Considerações Finais

Neste capítulo foi apresentado o algoritmo desenvolvido CHSMST^{+}, sendo descritas as modificações realizadas com o propósito de aprimorar os resultados e o desempenho do algoritmo original.

Para visualização e análise dos resultados foi construída a ferramenta FADE que permite identificar cada agrupamento junto ao mapa da região em que o mesmo é localizado, bem como analisar as características desses agrupamentos.

CAPÍTULO 5
Experimentos e Resultados

5.1 Considerações Iniciais

Neste capítulo são apresentados os experimentos executados com o algoritmo desenvolvido, CHSMST$^+$, a fim de demonstrar a aplicação do mesmo em contextos reais, validar a estratégia para otimização dos resultados, compará-lo com um algoritmo correlato e validar a estratégia para melhoria de desempenho.

5.2 Aplicação do Algoritmo CHSMST$^+$

Nesta seção são apresentados os experimentos realizados com o objetivo de demonstrar a aplicação do algoritmo CHSMST$^+$. Para isso, foi utilizada uma base de dados que contempla registros sobre acidentes de trabalho em uma região do interior de São Paulo com mais de 100 municípios. A alimentação desses dados é realizada por meio do Sistema de Informação e Vigilância de Acidentes de Trabalho - SIVAT, sendo que a referida base de dados possui mais de 20 mil notificações georreferenciadas – o que possibilita a aplicação do algoritmo desenvolvido e análise das informações resultantes. O SIVAT é resultado do convênio entre o Grupo de Banco de Dados - GBD e o Centro de Referência de Saúde do Trabalhador - CEREST de São José do Rio Preto e de Ilha Solteira, São Paulo (CEREST, 2014).

Com o propósito de verificar a aplicabilidade em outro conjunto de dados, o algoritmo CHSMST$^+$ também foi aplicado em uma base de dados que contempla dados relacionados a pontos de captação de recursos hídricos e respectivos usuários. A referida base é alimentada pelo Sistema de Gerenciamento sobre Recursos Hídricos e Ambientais – GRH, que vem sendo desenvolvido com recursos do Fundo Estadual de Recursos Hídricos - FEHIDRO em duas áreas da bacia hidrográfica Turvo/Grande: na região do município de Monte Azul Paulista – SP, que soma 133 km^2, localizada na sub-bacia do Avanhandava (Deliberação CBH-TG, 2008) e na região do município de Votuporanga com aproximadamente 130 km^2, localizada na sub-bacia Ribeirão do Marinheiro (Deliberação CBH-TG, 2010). Atualmente, a base de

dados contempla 699 pontos georreferenciados de captação de recursos hídricos, que foram obtidos por meio de trabalho *in loco*.

5.2.1 Considerações iniciais

Para a análise dos resultados obtidos, em todos os experimentos realizados, é importante observar que:

- Os mapas apresentados estão centralizados na região das respectivas ocorrências;
- Os pontos representam locais de uma ou mais ocorrências;
- Os polígonos exibidos em destaque representam os agrupamentos de dados que possuem similaridade em relação aos critérios definidos.

A extração de conhecimento por meio de prospecção de dados espaciais requer a existência de pelo menos um atributo espacial para que a localização seja um dos critérios analisados durante a geração dos padrões. No caso do SIVAT, tem-se a característica espacial em relação ao local de ocorrência do acidente, local de atendimento e domicílio do acidentado. Tais dados são estratégicos aos CERESTs, que por meio deles tem a capacidade de realizar ações para minimizar as ocorrências, melhorar o atendimento nas unidades de saúde, bem como apontar melhores estratégias de alocação do acidentado à unidade de saúde mais próxima de sua residência. Já no caso GRH, o dado espacial corresponde à localização geográfica do ponto de captação de recursos hídricos, que pode ser utilizada pelos órgãos públicos para suportar ações de preservação dos recursos ambientais.

Nos experimentos realizados na base de dados do SIVAT, foram selecionados os atributos não espaciais considerados importantes pelos CERESTs para a análise de dados. Já no caso do experimento realizado na base de dados do GRH, foi selecionado um atributo não espacial que confirma a situação de irregularidade dos pontos de captação perante aos órgãos públicos e um atributo não espacial que mostra o perfil do usuário que realiza a captação dos recursos hídricos.

Além disso, é importante destacar que o cálculo da similaridade em todos os experimentos apresentados é realizado por meio da Equação (3), na qual são verificadas as coincidências dos atributos não espaciais dos objetos analisados. Portanto, a referida equação pode ser aplicada a qualquer tipo de dado, conforme foi discutido na Seção 4.3.1. Como exemplo, considere os atributos categóricos afastamento e tipo de acidente, cujos valores são apresentados na Tabela 3.

Tabela 3: Exemplo de cálculo da similaridade.

ID	Afastamento	Tipo de Acidente
1	Não	Típico
2	Sim	Trajeto
3	Sim	Trajeto

Ao aplicar a Equação (3) para cada par de objetos obtêm-se:

$$s(1,2) = 1 - \frac{2-0}{2} = 0$$

$$s(1,3) = 1 - \frac{2-0}{2} = 0$$

$$s(2,3) = 1 - \frac{2-2}{2} = 1$$

Ao considerar apenas a similaridade no exemplo apresentado, os resultados indicam que os registros com id = 2 e id = 3 devem ser agrupados, uma vez que estes apresentam alta similaridade.

5.2.2 Base de dados SIVAT

Os CERESTs de São José do Rio Preto e Ilha Solteira analisam os dados do ano anterior, para elaborar as estratégias do ano atual para prevenção e fiscalização dos acidentes de trabalho, além de realizar estudos comparativos em relação às ocorrências de acidentes de trabalho ao longo do tempo. No ano de 2012 em São José do Rio Preto até a confecção deste trabalho haviam sido registrados 11.248 acidentes, sendo que 4.512 foram georreferenciados quanto ao local do acidente, ou seja, em 40,11% dos acidentes foi possível identificar a coordenada geográfica referente ao local onde o acidente ocorreu. Neste contexto, os experimentos de análise sobre os agrupamentos foram realizados sobre as notificações de acidentes georreferenciados referentes ao ano de 2012 para ilustrar a aplicação do algoritmo CHSMST^{+} em um contexto real.

Com o propósito de comparar a distribuição das características das notificações de acidentes de trabalho com e sem georreferenciamento, foi realizada uma consulta na base de dados, considerando os atributos não espaciais utilizados nos experimentos apresentados nesta seção: afastamento, tipo de acidente, ramo de atividade, período do acidente e faixa etária. Os resultados são apresentados nas Figuras 24 a 28, nas quais é possível observar que a distribuição das características ocorre de forma semelhante para as notificações georreferenciadas e não georreferenciadas, exceto para o atributo tipo de acidente, que devido à dificuldade de obtenção do endereço exato do local do acidente, quando este ocorre no trajeto de casa para o

trabalho ou do trabalho para a casa, existe uma menor quantidade de notificações georreferenciadas com acidentes deste tipo.

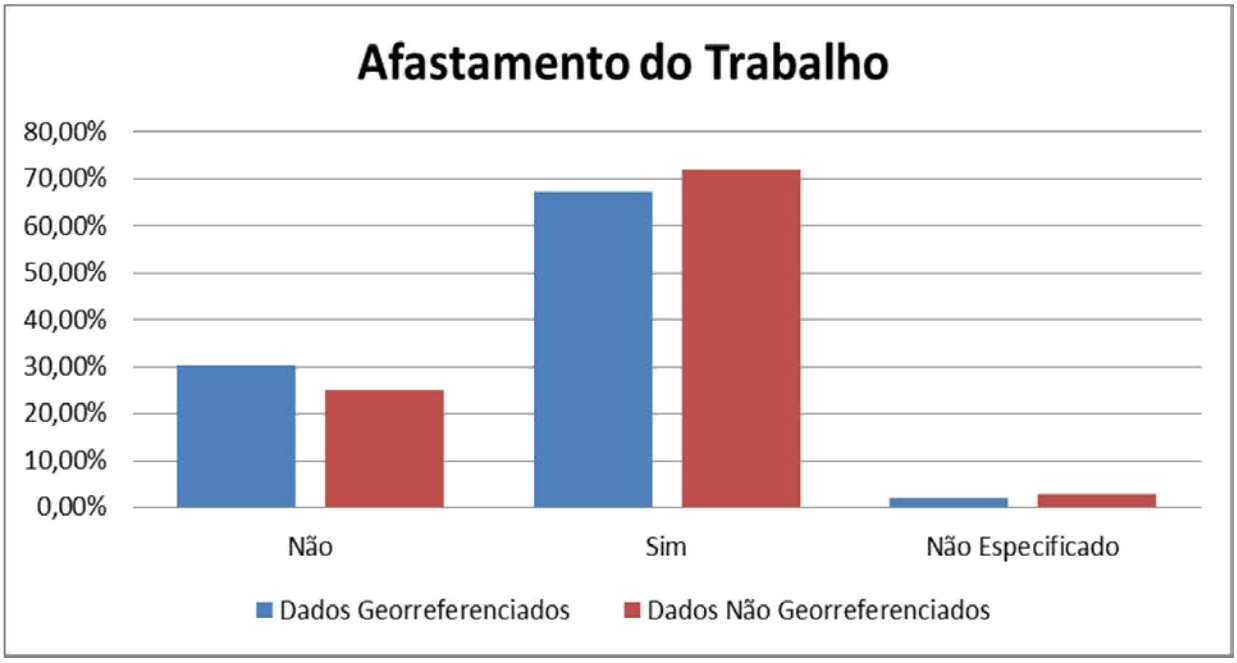

Figura 24: Distribuição das características nas notificações georreferenciadas e não georreferenciadas – Afastamento do Trabalho.

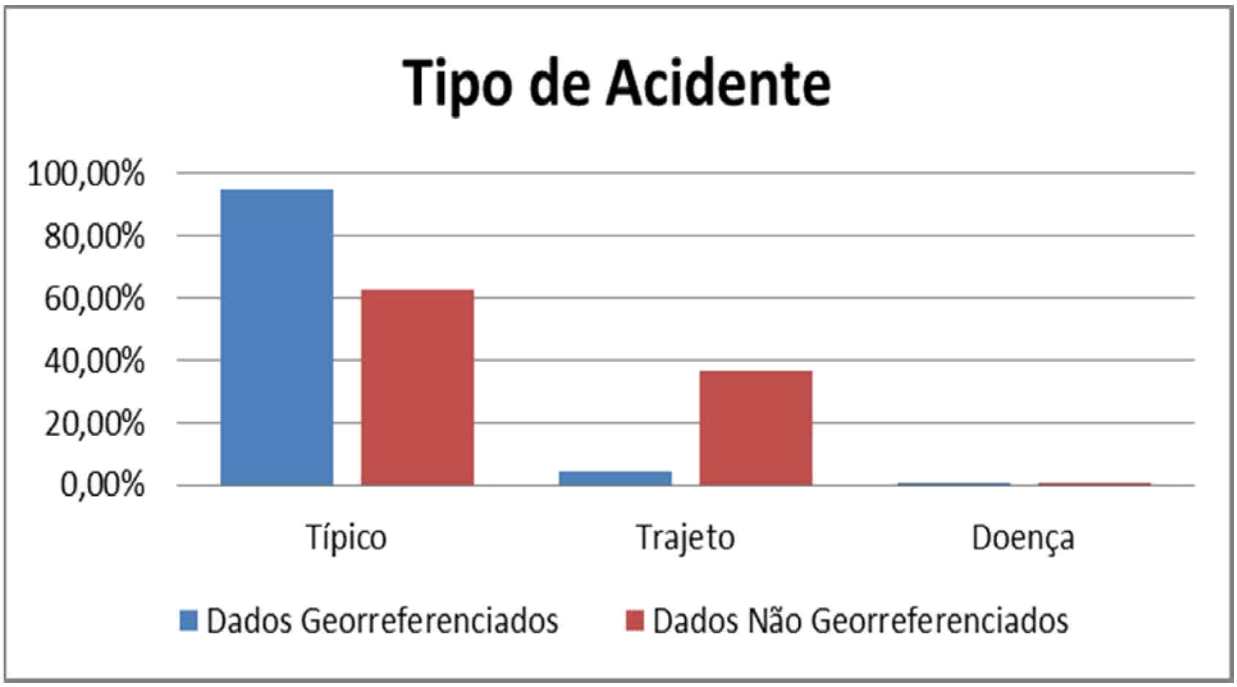

Figura 25: Distribuição das características nas notificações georreferenciadas e não georreferenciadas – Tipo de Acidente.

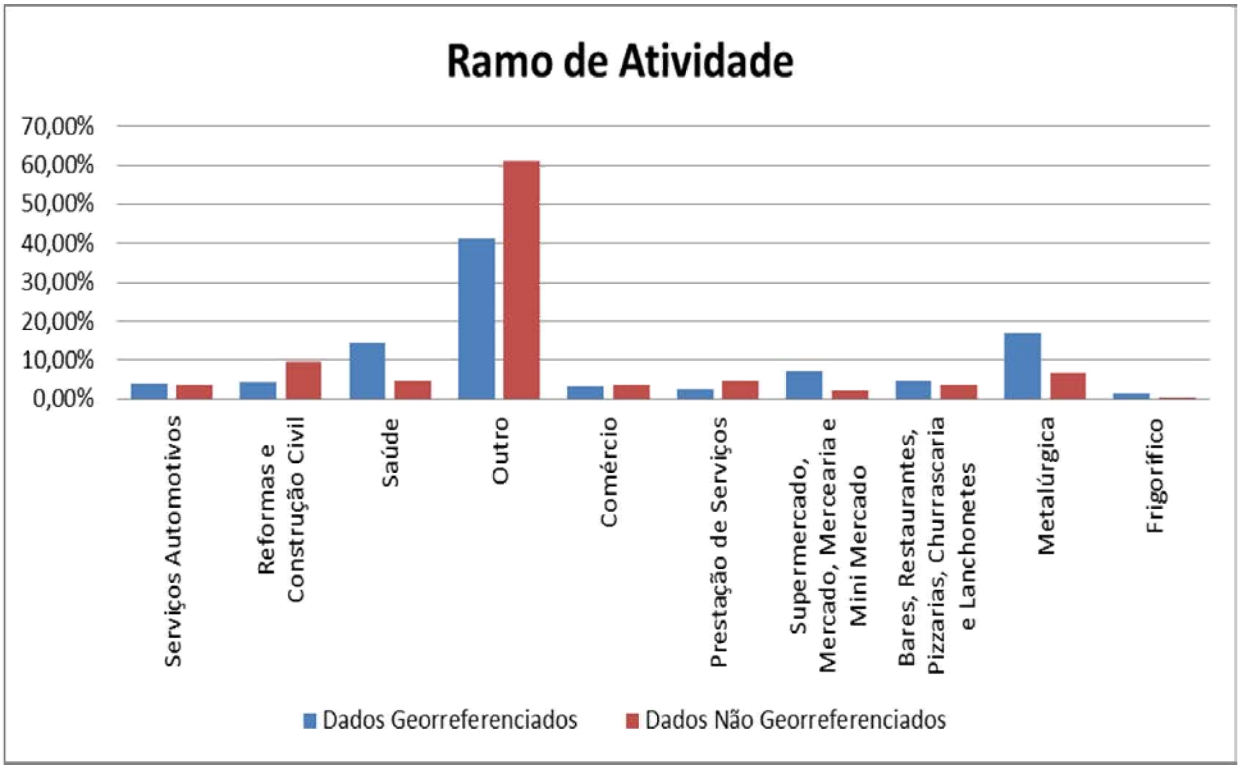

Figura 26: Distribuição das características nas notificações georreferenciadas e não georreferenciadas – Ramo de Atividade.

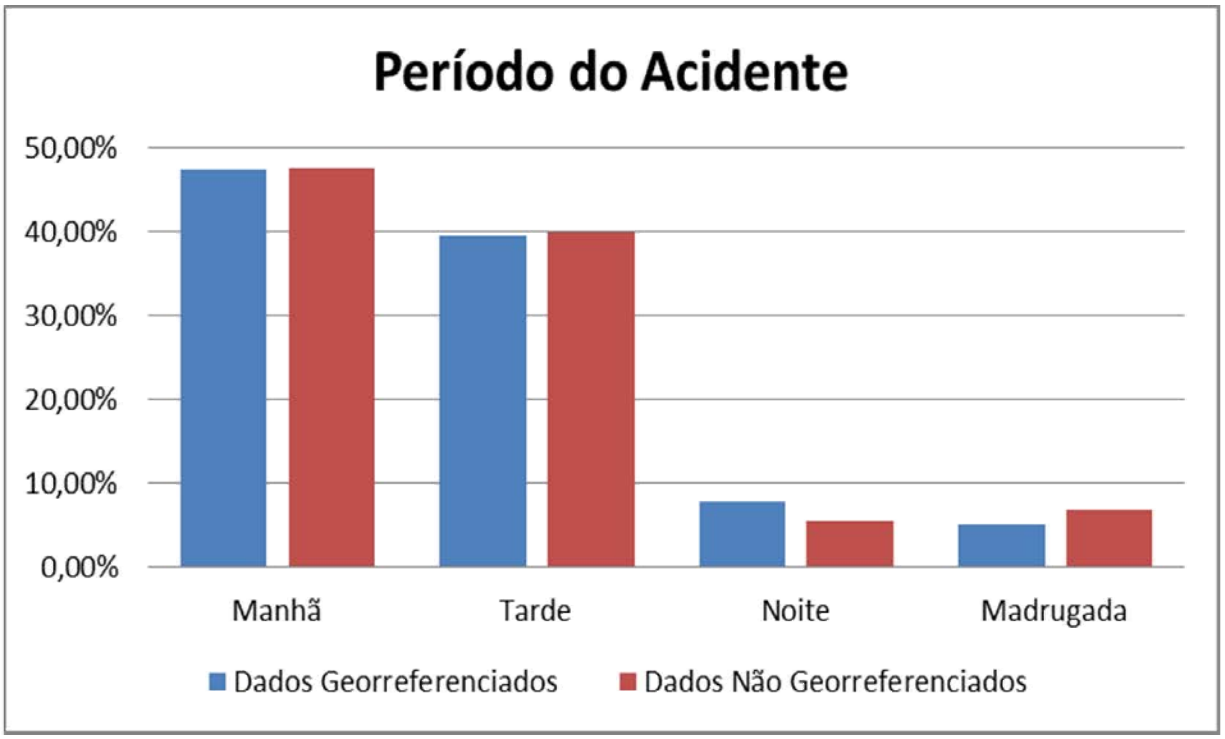

Figura 27: Distribuição das características nas notificações georreferenciadas e não georreferenciadas – Período do Acidente.

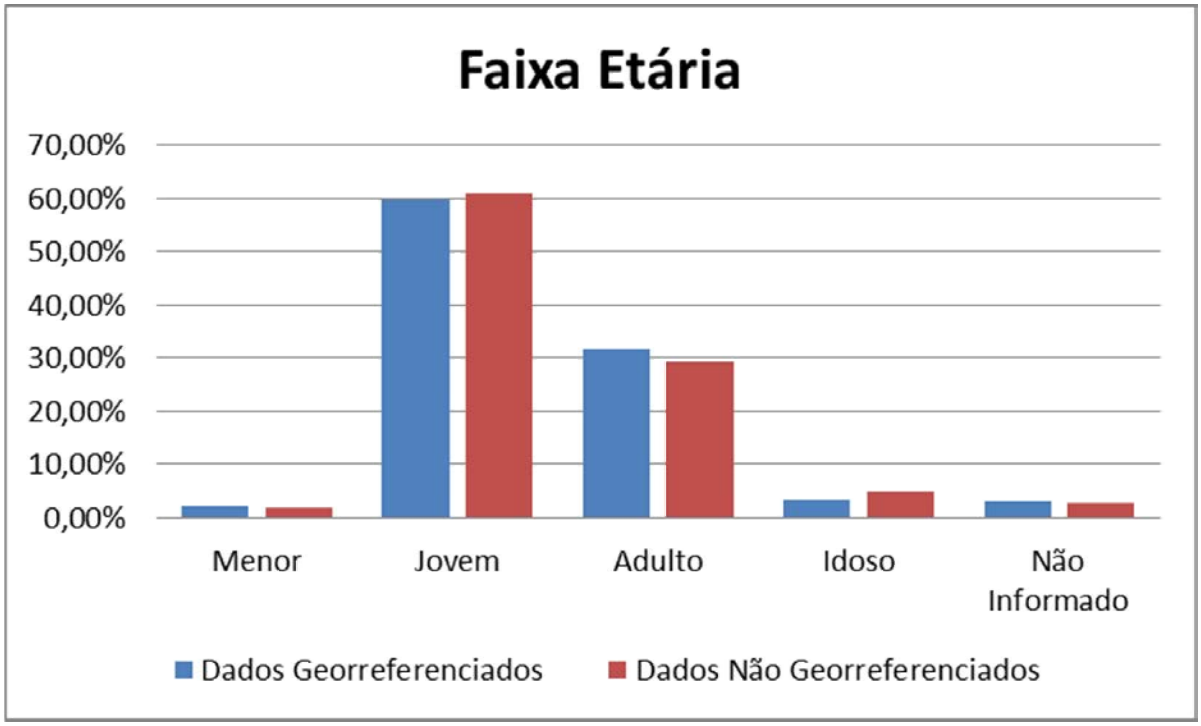

Figura 28: Distribuição das características nas notificações georreferenciadas e não georreferenciadas – Faixa Etária.

No primeiro experimento foram selecionados os atributos não espaciais afastamento e tipo de acidente, bem como o atributo espacial referente ao local do acidente. Na Figura 29 são apresentados os agrupamentos obtidos. Neste experimento foram gerados 70 agrupamentos, sendo verificados os agrupamentos que mais de 80% dos registros possuem uma determinada característica. Os resultados são apresentados na Tabela 4 e na Tabela 5.

Na Tabela 4 verifica-se que 30 agrupamentos possuem mais de 80% de ocorrência de afastamento do trabalho e 68 agrupamentos possuem mais de 80% de acidentes típicos. A partir disso, foram verificados quais agrupamentos possuem mais de 100 notificações de acidentes de trabalho, com o propósito de analisar os agrupamentos cujos resultados fossem mais representativos em relação ao conjunto de dados analisado e quais destes apresentavam maior índice de afastamento do trabalho, uma vez que esta é uma característica bastante importante ao CERESTs, já que indica que o acidente é grave. A partir disso, foi identificado um agrupamento, localizado próximo à rodovia Washington Luís e que possui alto índice de qualidade tanto em relação à similaridade quanto em relação à distância, e que, portanto foi escolhido para ser apresentado neste trabalho. Na Figura 30 é exibida a seleção do referido agrupamento, que a fim de facilitar a discussão sobre o mesmo será denominado como agrupamento 1.

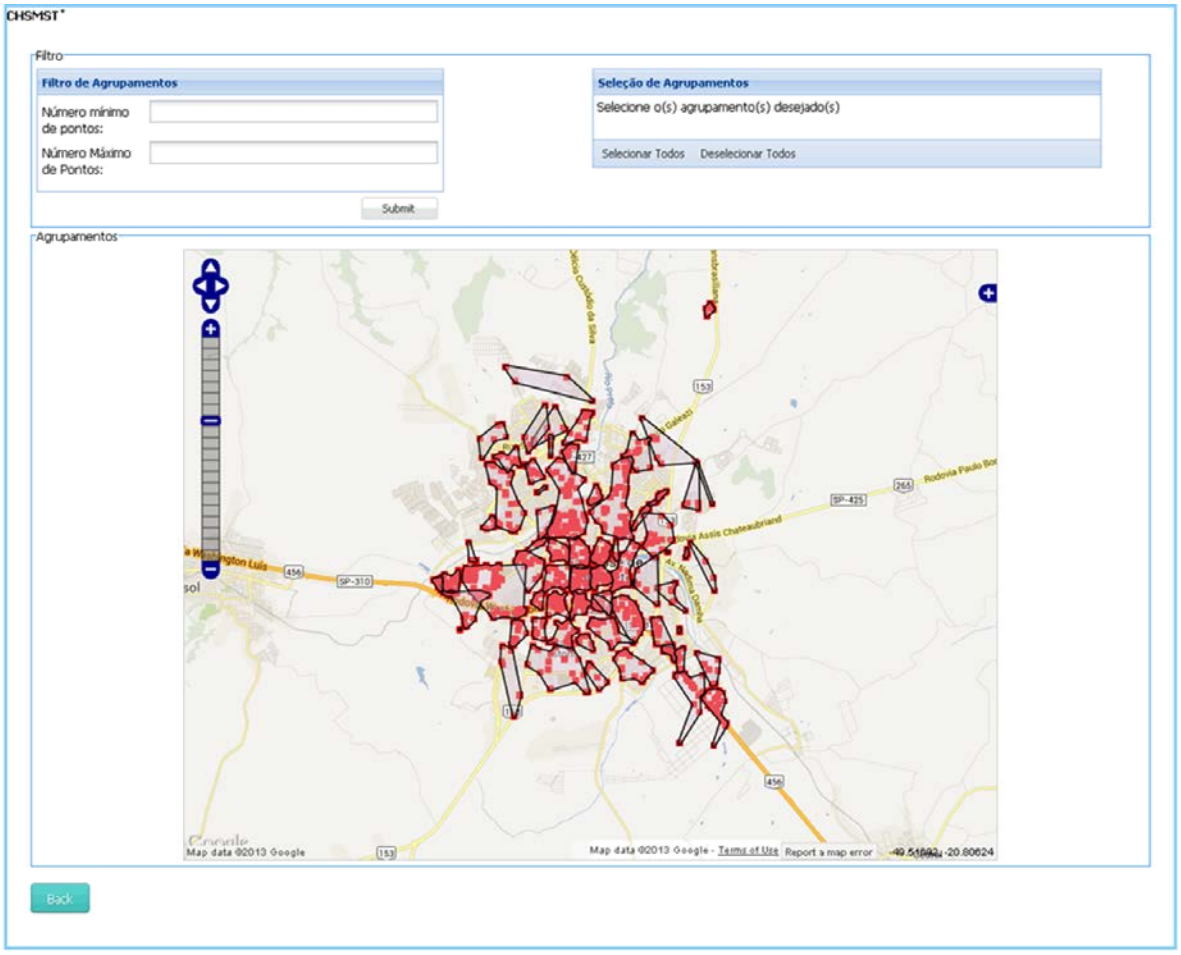

Figura 29: Agrupamentos gerados para a base de dados SIVAT - Experimento 1.

Tabela 4: Comparação dos agrupamentos em que mais de 80% dos registros apresentam determinada característica em relação a afastamento do trabalho.

	Não	Sim	Não Especificado
Número de Agrupamentos	2	30	0
Média das Porcentagens	100%	89,41%	-

Tabela 5: Comparação dos agrupamentos em que mais de 80% dos registros apresentam determinada característica em relação ao tipo de acidente.

	Típico	Trajeto	Doença do Trabalho
Número de Agrupamentos	68	0	0
Média das Porcentagens	97,26%	-	-

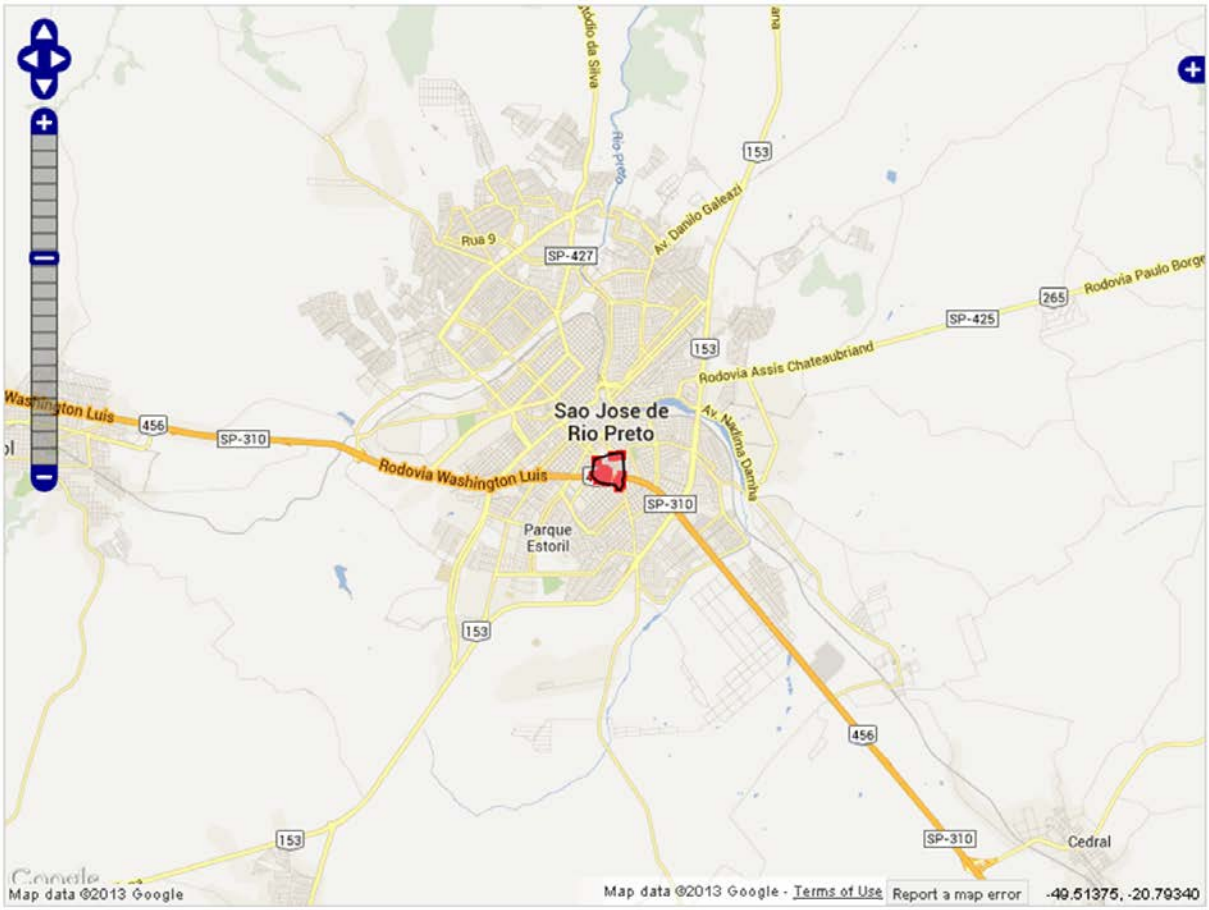

Figura 30: Agrupamento 1 gerado a partir dos atributos afastamento, tipo de acidente e localização espacial da ocorrência.

O agrupamento 1 é composto por 102 ocorrências de acidentes de trabalho, e com o propósito de permitir uma análise sobre o seu perfil, na Figura 31 e na Figura 32 são apresentados os gráficos referentes aos atributos não espaciais utilizados como critério pelo algoritmo CHSMST^{+}.

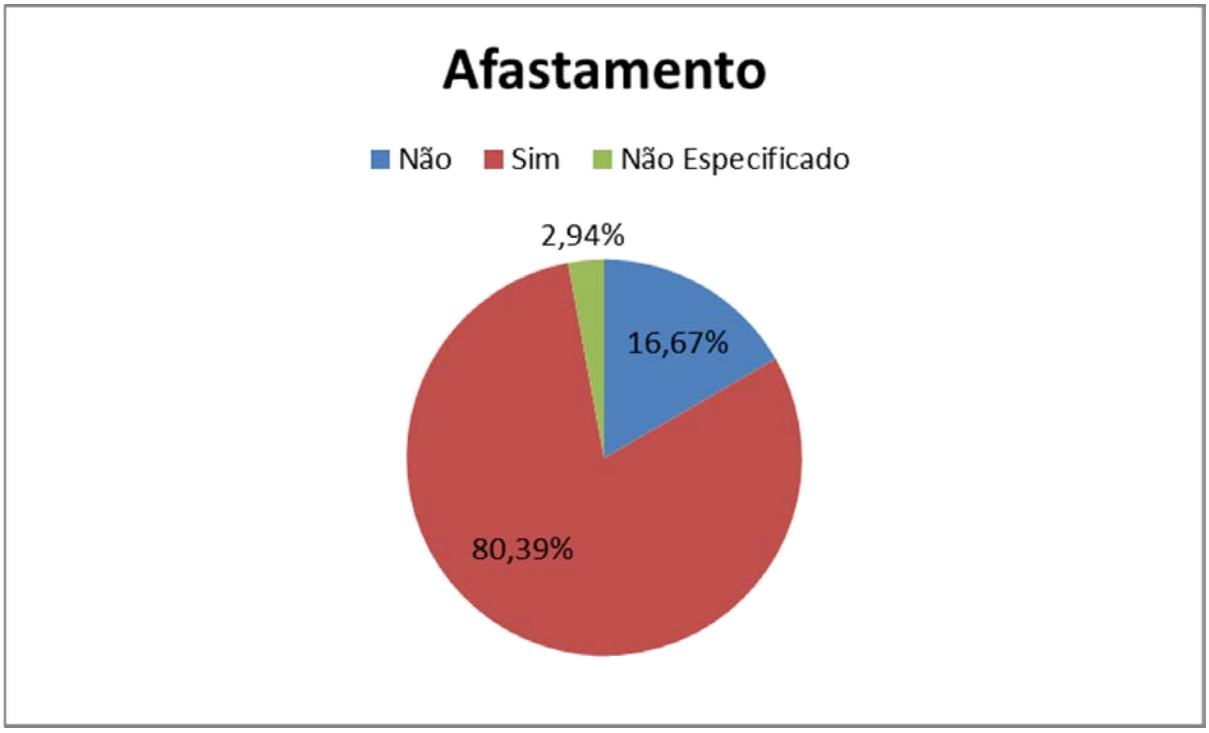

Figura 31: Gráfico referente ao agrupamento 1, considerando atributo afastamento.

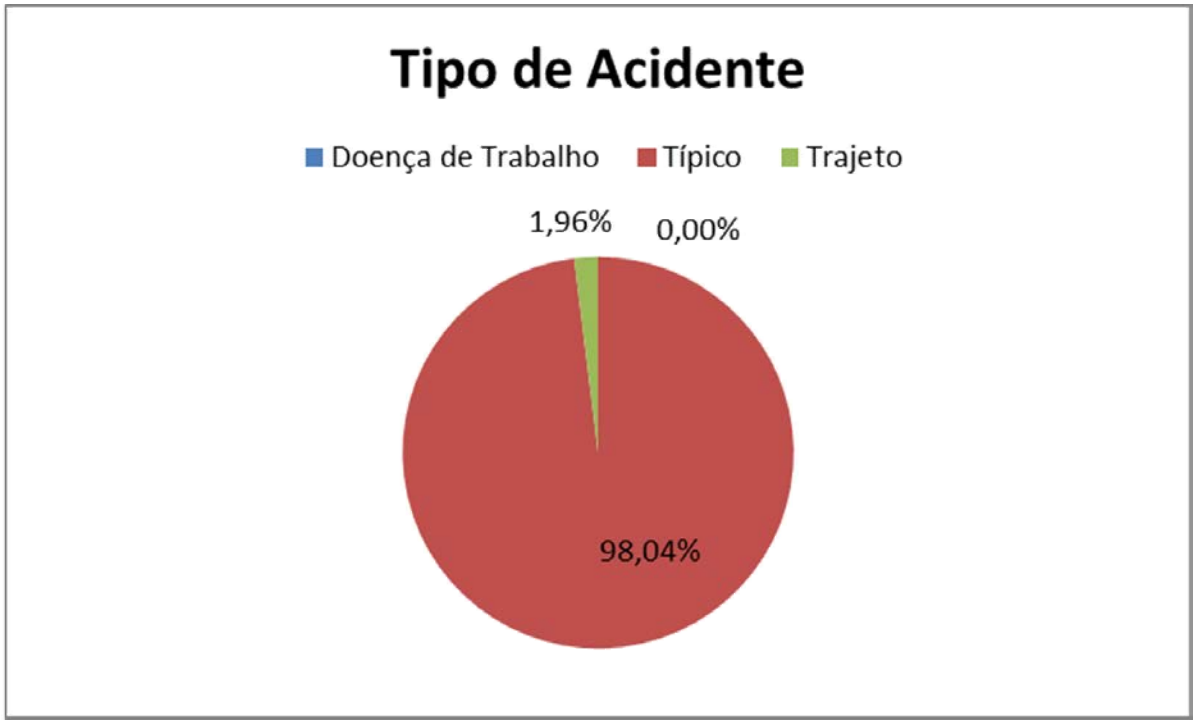

Figura 32: Gráfico referente ao agrupamento 1, considerando o atributo tipo de acidente.

Na Figura 31 verifica-se que dos acidentes de trabalho referentes ao agrupamento 1, 80,39% ocorreram afastamento de trabalho, ou seja, os acidentados tiveram que se ausentar do trabalho devido aos ferimentos ocasionados pelos acidentes, sendo assim caracterizados como acidentes mais graves. Na Figura 32, verifica-se que 98,04% dos acidentes foram típi-

cos, ou seja, ocorreram durante o expediente, e não foram ocasionados por doença do trabalho e também não ocorreram durante o trajeto de ida ou volta ao local de trabalho.

Ao encontrar uma região na qual foram registrados vários casos de afastamento do trabalho, o algoritmo possibilita que o CEREST de São José do Rio Preto direcione esforços para minimizar tais ocorrências, que sem a aplicação de um algoritmo de agrupamento de dados espaciais, não seriam identificadas com a mesma facilidade uma vez que o município de São José do Rio Preto possui uma área de 431,96 km² com mais de 400 mil habitantes, além disso, o CEREST de São José do Rio Preto é responsável por mais de 100 municípios, o que dificulta ainda mais a vigilância dos acidentes de trabalho.

Com o propósito de medir a qualidade dos agrupamentos gerados no experimento 1 foi aplicado o algoritmo para o cálculo da qualidade do agrupamento, baseado no coeficiente da silhueta, tanto em relação à distância quanto à similaridade. Como resultado para o agrupamento 1 obteve-se 0,96 para similaridade e 1,00 para distância, o que configura-o como sendo bastante adequado em relação a compactação dos objetos dentro do agrupamento e dissimilaridade em relação aos demais agrupamentos. Já a qualidade média para todos os agrupamentos foi de 0,86 para similaridade e 0,97 para distância.

No segundo experimento foram selecionados os atributos não espaciais: ramo de atividade, período do acidente e faixa etária, bem como o atributo espacial referente ao local do acidente. Os agrupamentos gerados são exibidos na Figura 33.

No experimento 2 foram gerados 41 agrupamentos, sendo verificados os agrupamentos em que mais de 80% dos registros possuem uma determinada característica. Os resultados são apresentados nas Tabelas 6, 7 e 8.

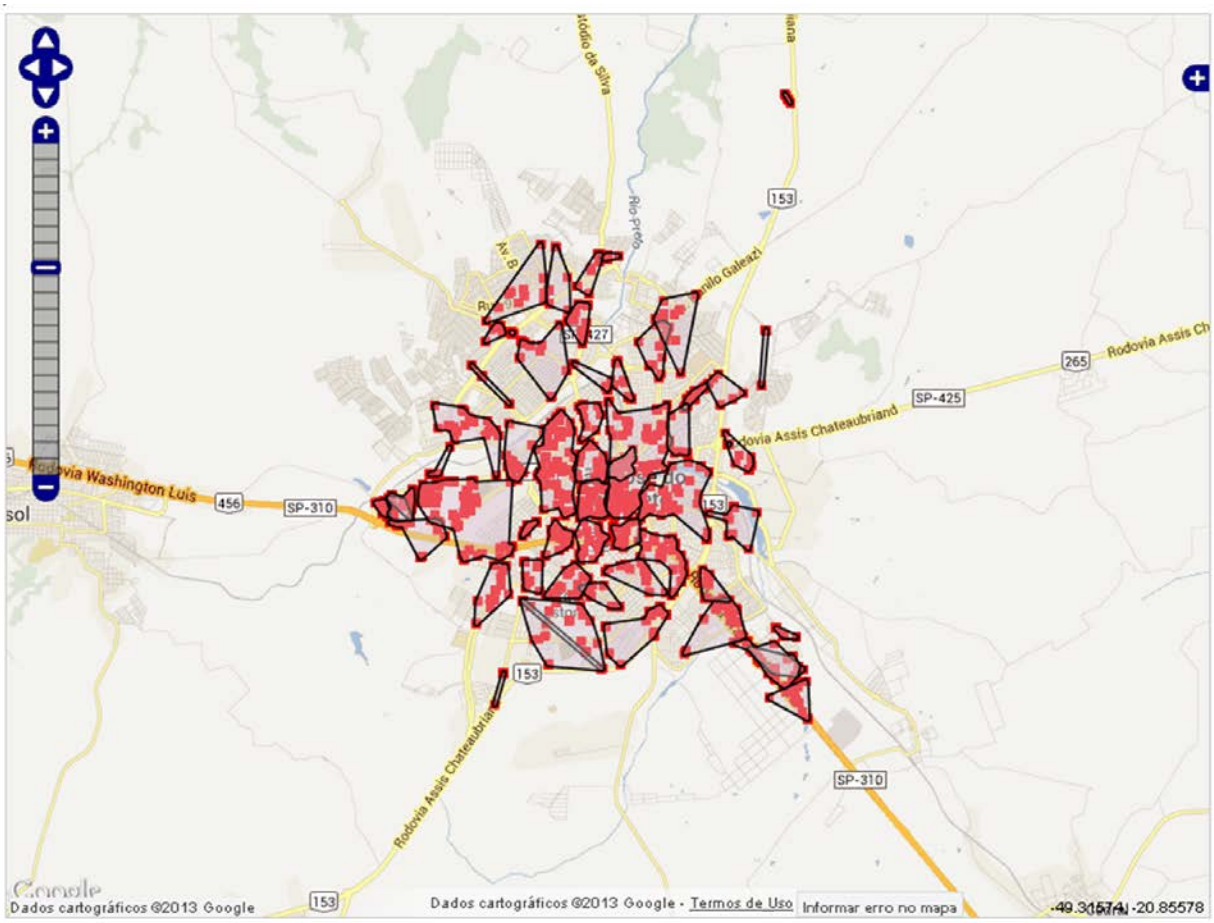

Figura 33: Agrupamentos gerados para a base de dados SIVAT - Experimento 2.

Tabela 6: Comparação dos agrupamentos em que mais de 80% dos registros apresentam determinada característica em relação ao ramo de atividade.

	Metalúrgica	Supermercado, mercado, mercearia e mini mercado	Reformas e Construção Civil	Saúde	Outro
Número de Agrupamentos	3	1	0	3	5
Média das Porcentagens	84,74%	83,82%	-	93,44%	96%

Tabela 7: Comparação dos agrupamentos em que mais de 80% dos registros apresentam determinada característica em relação ao período do acidente.

	Manhã	Tarde	Noite	Madrugada
Número de Agrupamen-	2	2	0	0
Média das Porcentagens	95%	100%	-	-

Tabela 8: Comparação dos agrupamentos em que mais de 80% dos registros apresentam determinada característica em relação à faixa etária.

	Menor	Jovem	Adulto	Idoso
Número de Agrupamentos	0	9	0	0
Média das Porcentagens	-	89,71%	-	-

Nas Tabelas 6, 7 e 8 verifica-se que 12 agrupamentos possuem mais de 80% em relação ao ramo de atividade, 4 agrupamentos possuem mais de 80% de acidentes que ocorreram no período da manhã ou tarde e 9 agrupamentos possuem mais de 80% de acidentes com jovens. A partir disso, foram verificados quais destes agrupamentos possuem mais de 100 notificações de acidentes de trabalho, com o propósito de analisar os agrupamentos cujos resultados fossem mais representativos em relação ao conjunto de dados. Foram identificados 5 agrupamentos, sendo que o mais interessante possui um percentual significativo em relação a área da saúde e bom índice de qualidade tanto em relação à similaridade quanto em relação à distância, e que, portanto foi escolhido para ser apresentado neste trabalho. Na Figura 34 é apresentado o referido agrupamento, localizado na região sul do município de São José do Rio Preto, que será denominado como agrupamento 2. Nas Figuras 35, 36 e 37 são apresentados os gráficos referentes aos atributos não espaciais utilizados como critério pelo algoritmo CHSMST^{+} neste experimento.

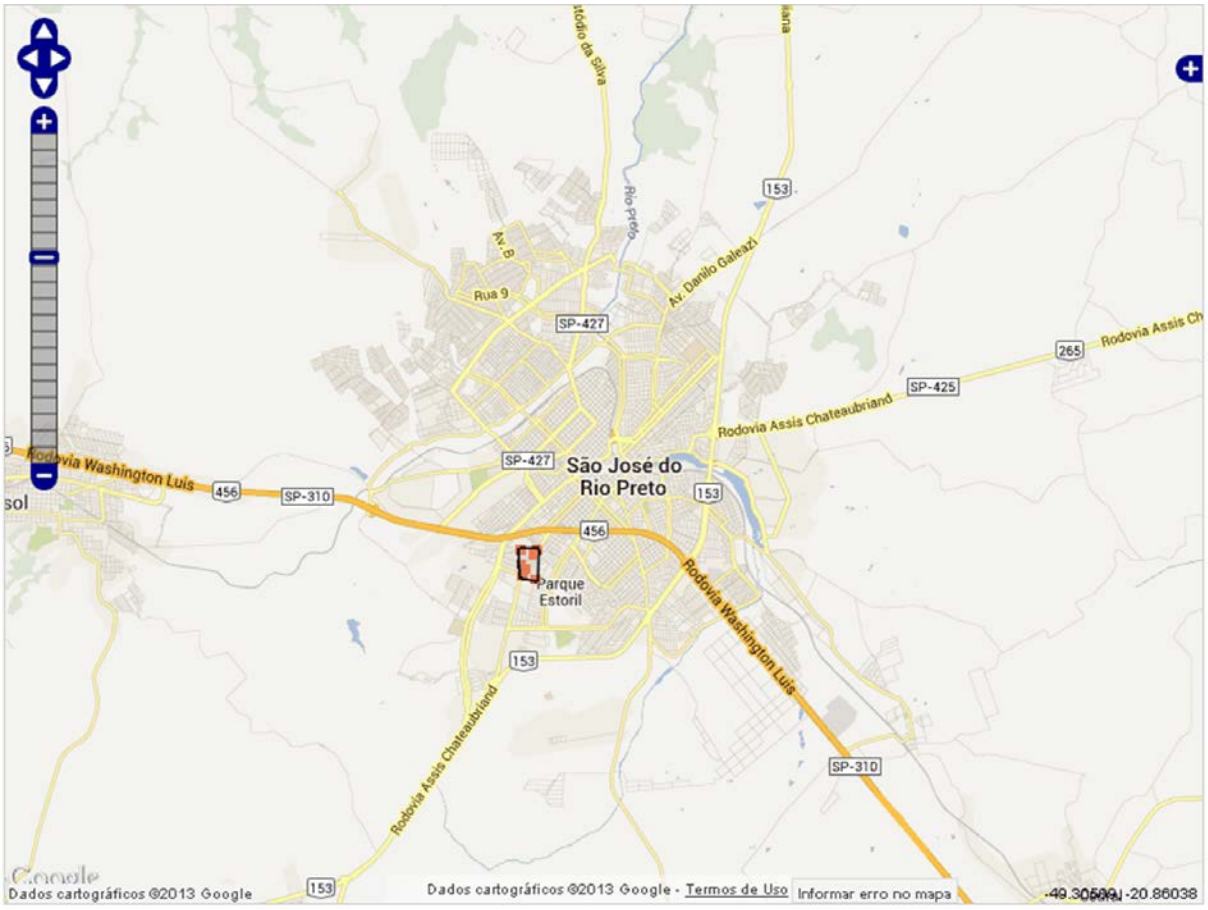

Figura 34: Agrupamento 2 gerado a partir dos atributos ramo de atividade, período do acidente, faixa etária e localização espacial da ocorrência.

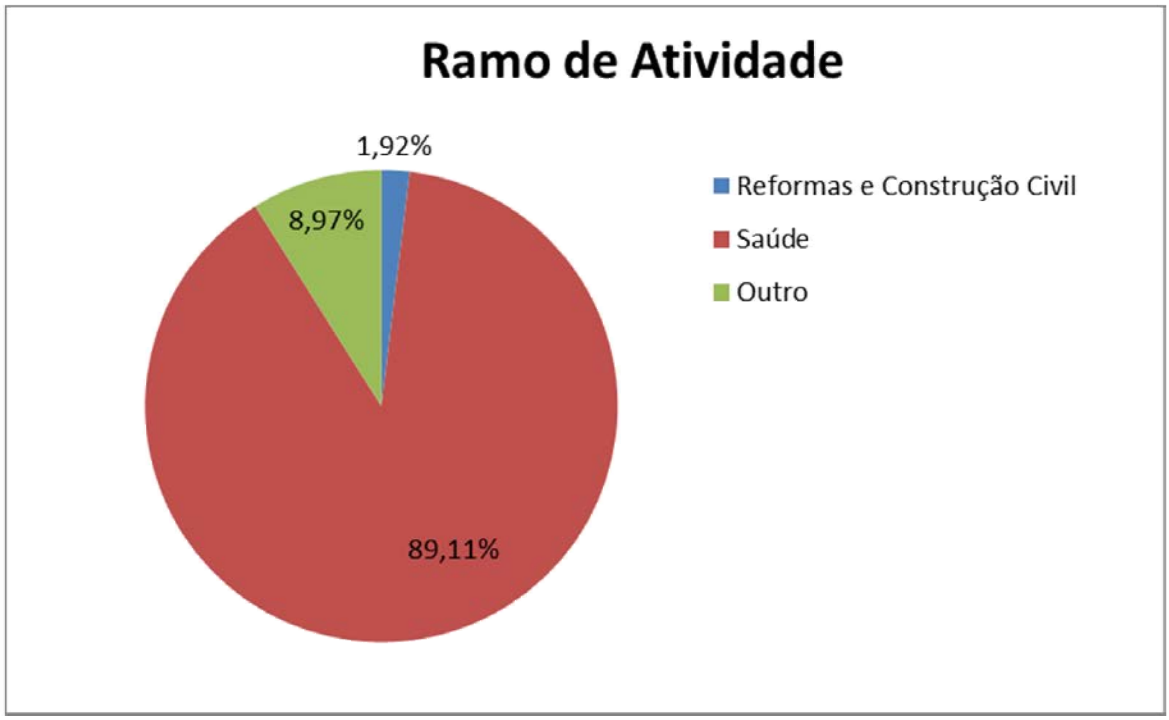

Figura 35: Gráfico referente ao agrupamento 2, considerando o atributo ramo de atividade.

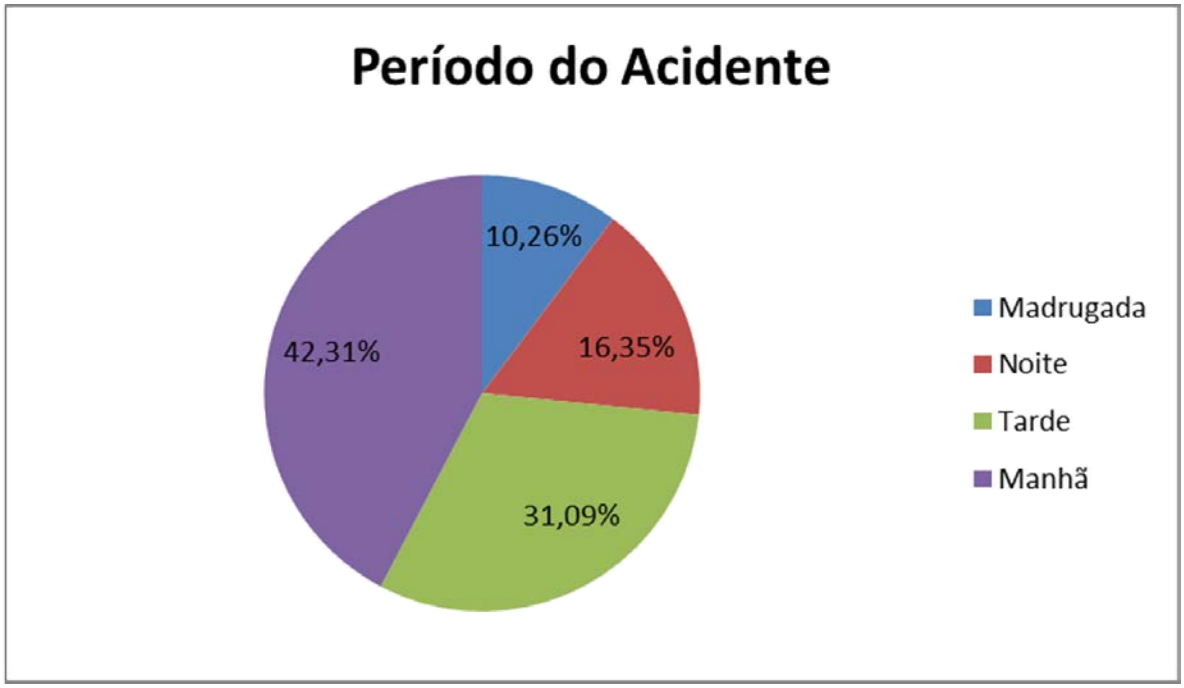

Figura 36: Gráfico referente ao agrupamento 2, considerando o atributo período do acidente.

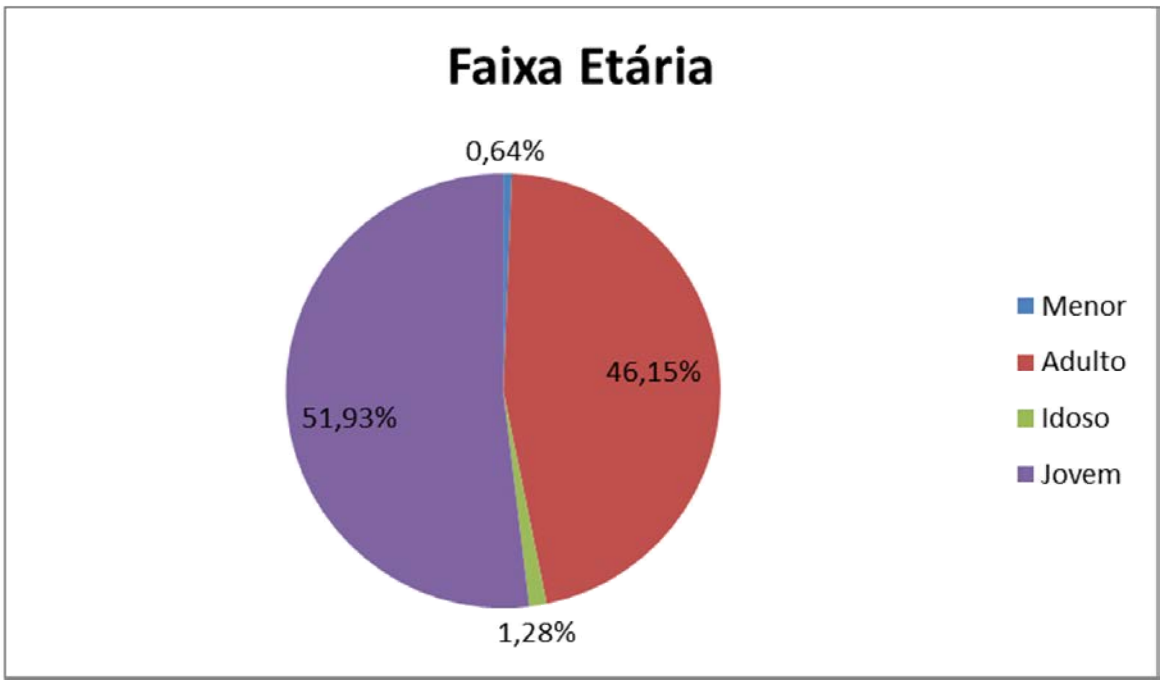

Figura 37: Gráfico referente ao agrupamento 2, considerando o atributo faixa etária.

Ao analisar o gráfico apresentado na Figura 35 verifica-se que 89,11% dos acidentes pertencentes ao agrupamento 2 são da área da saúde, o que equivale a 156 notificações. Este número é representativo, pois, no conjunto de dados analisado o ramo de atividade é bastante diversificado, sendo que o ramo "Outros" corresponde a 41,47% de todos os acidentes de trabalho analisados, enquanto que a área da saúde corresponde a apenas 14,61%. Assim, o número de acidentes da área da saúde do agrupamento 2 se torna ainda mais relevante ao verifi-

car que estes correspondem a 23,67% de todos os acidentes de trabalho do ramo da saúde presentes no conjunto de dados analisado. Portanto, o experimento 2 permitiu detectar uma região responsável por uma quantidade significativa dos acidentes na área da saúde no ano de 2012, o que pode contribuir para uma ação dos órgãos públicos direcionada para esta região a fim de reduzir tais acidentes de trabalho.

Na Figura 36 é possível verificar que a maioria dos acidentes ocorreu no período da manhã, porém 26,61% ocorreram durante a noite ou madrugada. Na Figura 37 verifica-se que a maioria dos acidentes do agrupamento 2 foram com jovens.

A aplicação do algoritmo para o cálculo da qualidade do agrupamento 2, baseado no coeficiente da silhueta, resultou em 1,00 para distância e 0,85 para similaridade, o que configura o agrupamento como sendo bastante eficiente em relação a compactação dos objetos dentro do agrupamento e dissimilaridade em relação aos demais agrupamentos, sendo que a qualidade média para todos os agrupamentos foi de 0,90 para distância e 0,74 para similaridade.

O experimento 1 foi realizado considerando dois atributos não espaciais, enquanto que o experimento 2 considerou três atributos. Assim é possível verificar que o algoritmo CHSMST$^+$ além considerar o atributo espacial no agrupamento dos dados, também permite correlacionar diferentes números de atributos não espaciais, o que possibilita uma descoberta de conhecimento mais flexível e significativa do que a maioria dos algoritmos estudados.

5.2.3 Base de dados GRH

Para a aplicação do algoritmo CHSMST$^+$ na base de dados GRH foram considerados 699 pontos de captação de recursos hídricos georreferenciados, sendo selecionados os atributos não espaciais situação administrativa e perfil do usuário, bem como o atributo espacial referente ao local de captação dos recursos hídricos. Na Figura 38 são apresentados os agrupamentos gerados.

Neste experimento foram gerados 5 agrupamentos, sendo verificados aqueles que mais de 80% dos registros possuem uma determinada característica. Os resultados são apresentados na Tabela 9 e na Tabela 10.

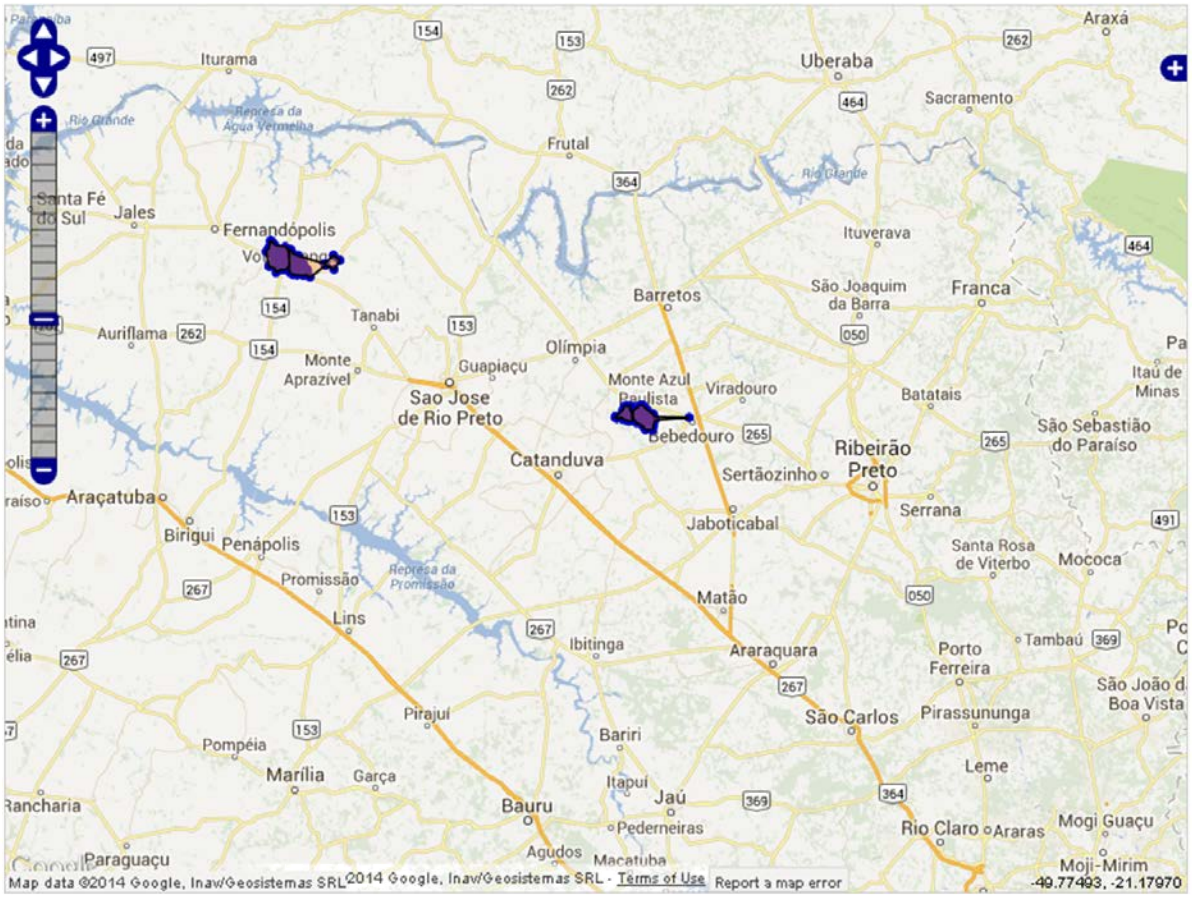

Figura 38: Agrupamentos gerados para a base de dados GRH.

Tabela 9: Comparação dos agrupamentos em que mais de 80% dos registros apresentam determinada característica em relação à situação administrativa.

	Cadastrado	Licença de Operação	Implantação Autorizada	Outorga de Direito e Uso	Sem Autorização	Não Informado
Número de Agrupamentos	1	1	0	1	1	1
Média da Porcentagem	81,25%	100%	-	100%	91,38%	100%

Tabela 10: Comparação dos agrupamentos em que mais de 80% dos registros apresentam determinada característica em relação ao perfil do usuário.

	Particular Consumo Próprio	Comércio	Chácaras de Recreio	Indústria	Abastecimento Público	Não Informado
Número de Agrupamentos	5	0	0	0	0	0
Média da Porcentagem	97,96%	-	-	-	-	-

A partir disso, foi selecionado o agrupamento que apresenta um elevado índice de pontos de captação de recursos hídricos sem autorização dos órgãos públicos, uma vez que esta situação administrativa é mais interessante para análise que as demais. Na Figura 39 é apresentada a seleção do referido agrupamento, denominado como agrupamento 3, localizado próximo ao município de Monte Azul Paulista.

O agrupamento 3 contempla 58 pontos de captação de recursos hídricos, sendo que na Figura 40 e na Figura 41 são apresentados os gráficos referentes aos atributos não espaciais utilizados como critério pelo algoritmo CHSMST^{+}. Em relação à medida da eficiência do agrupamento, no cálculo do coeficiente da silhueta, obteve-se 0,97 em relação à similaridade e 0,56 em relação à distância, o que significa que o agrupamento possui ótima qualidade em relação à similaridade e boa qualidade em relação à distância, sendo que a qualidade média para todos os agrupamentos em relação à similaridade e a distância foi de 0,92 e 0,50 respectivamente.

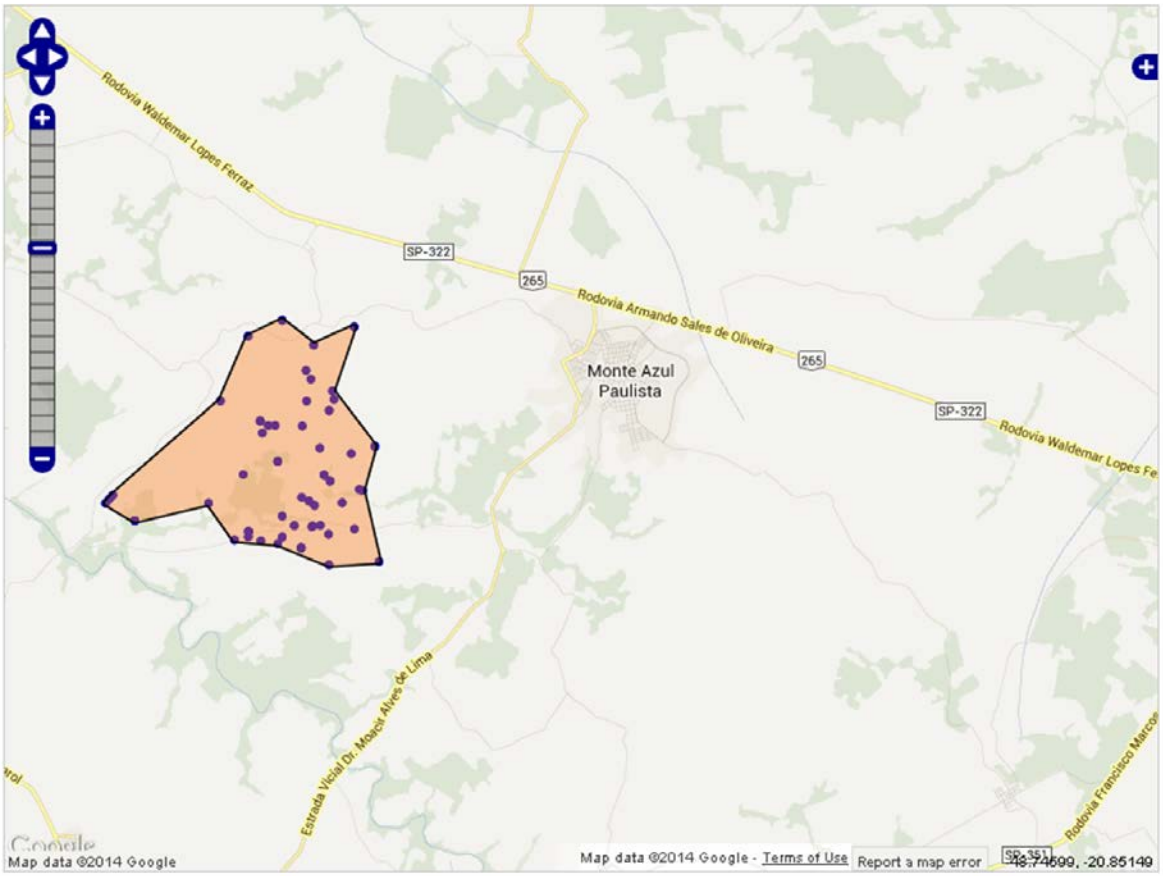

Figura 39: Agrupamento 3 gerado próximo ao município de Monte Azul Paulista.

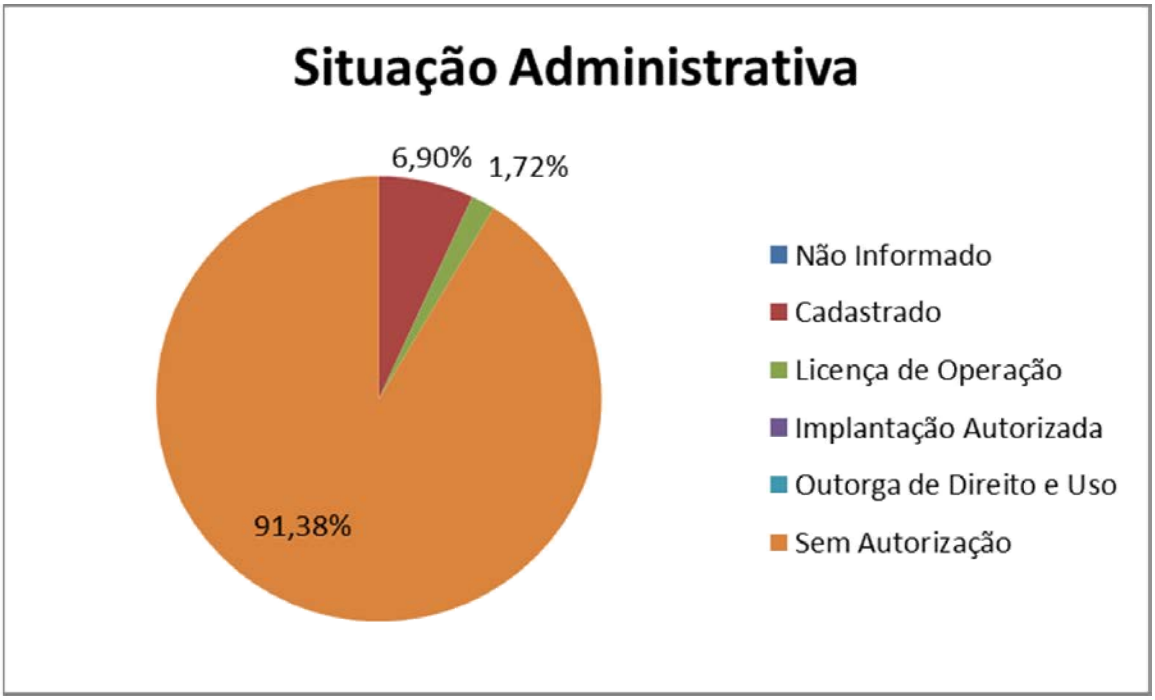

Figura 40: Gráfico referente ao agrupamento 3, considerando atributo situação administrativa.

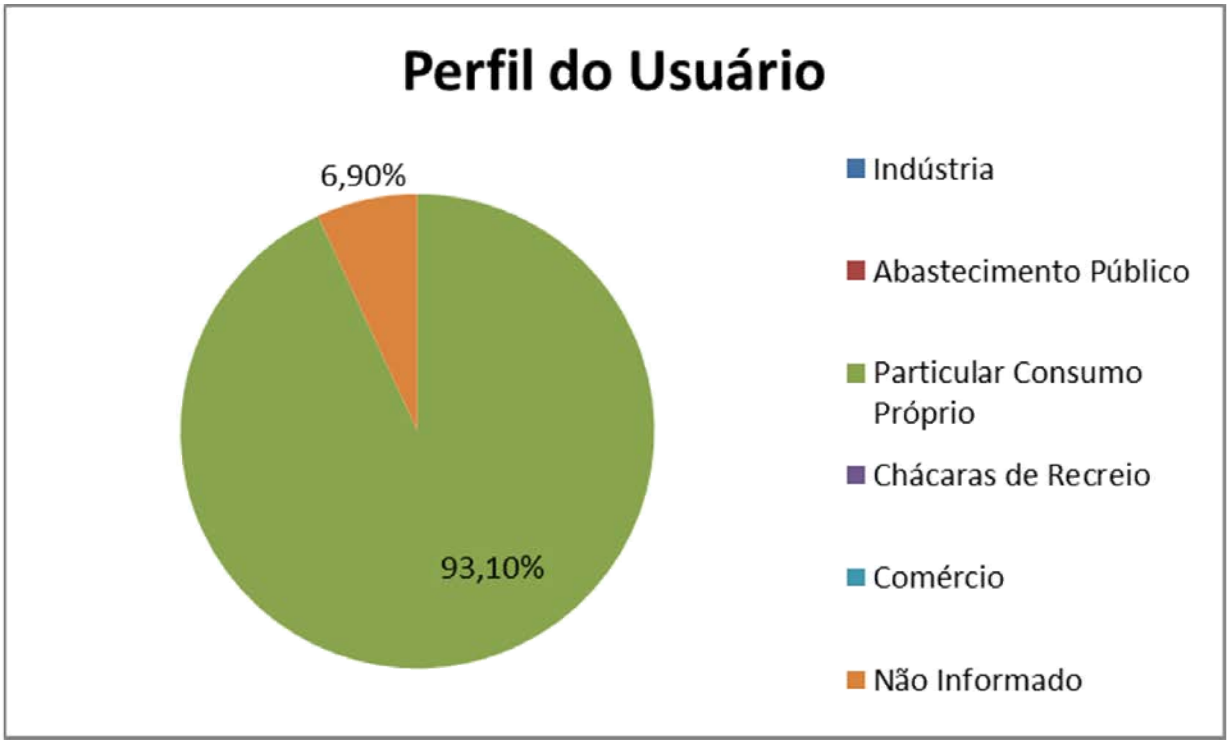

Figura 41 Gráfico referente ao agrupamento 3, considerando atributo perfil do usuário.

Na Figura 40 é possível verificar que 91,38% dos pontos de captação pertencentes ao agrupamento 3 estão sem autorização dos órgãos públicos e na Figura 41 verifica-se que 93,10% dos recursos hídricos são utilizados para consumo próprio. A situação identificada neste agrupamento reflete a dificuldade dos órgãos públicos em controlar o número de pontos de captação e o volume de recursos hídricos coletados, o que pode resultar em um colapso no

abastecimento. Assim, a aplicação de algoritmos de agrupamentos de dados espaciais, tal como o CHSMST^{+} pode contribuir para uma gestão mais eficiente com o objetivo de preservar os recursos ambientais, uma vez que permite identificar o perfil de uma região de acordo com as características selecionadas.

Os experimentos realizados nas bases de dados SIVAT e GRH mostraram a capacidade de aplicação do algoritmo CHSMST^{+} tanto na área da saúde quanto na área ambiental, ressaltando ainda mais a importância do trabalho desenvolvido.

5.3 Experimento Comparativo da Estratégia de Similaridade

Nesta seção é realizada uma comparação entre os resultados obtidos com e sem a estratégia de similaridade no algoritmo CHSMST$^{+,}$ uma vez que esta é uma das principais alterações realizadas para aprimorar os resultados obtidos com o algoritmo desenvolvido. Nestes experimentos o conjunto de dados contempla 4.512 registros de acidentes de trabalho ocorridos em São José do Rio Preto no ano de 2012 e o atributo espacial analisado é referente ao local do acidente de trabalho.

No primeiro experimento foi aplicado o algoritmo CHSMST^{+} considerando a similaridade em relação a ramo de atividade, sendo que os agrupamentos gerados são apresentados na Figura 42.

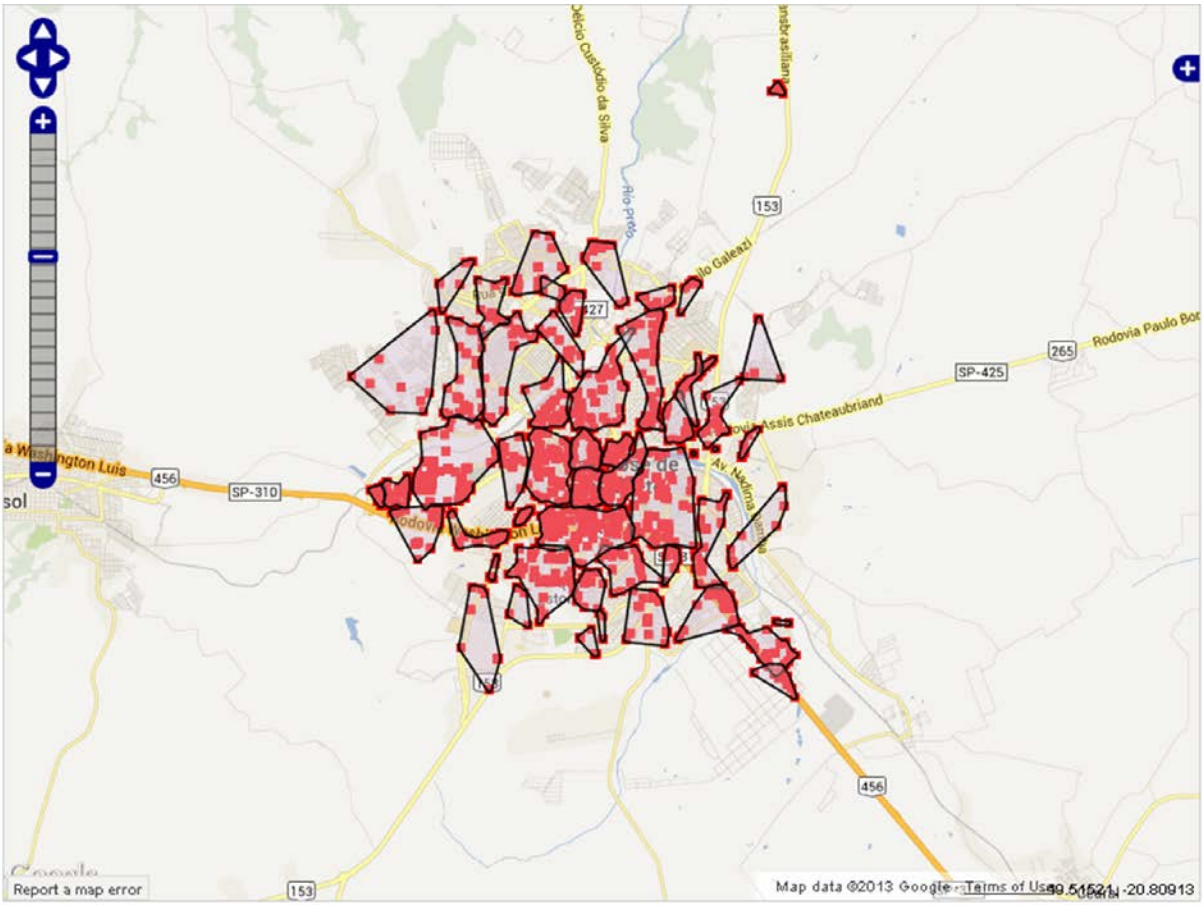

Figura 42: Agrupamentos considerando similaridade em relação ao ramo de atividade.

A análise sobre os agrupamentos gerados é enriquecida com a utilização dos recursos disponibilizados na ferramenta FADE, tais como filtro do agrupamento por número de pontos, seleção de um ou mais agrupamentos, geração de gráfico para visualização da distribuição da característica não espacial, bem como para obtenção do número de pontos no agrupamento e qualidade do agrupamento em relação à distância e similaridade. Na Figura 43 é apresentado um exemplo da utilização da ferramenta FADE.

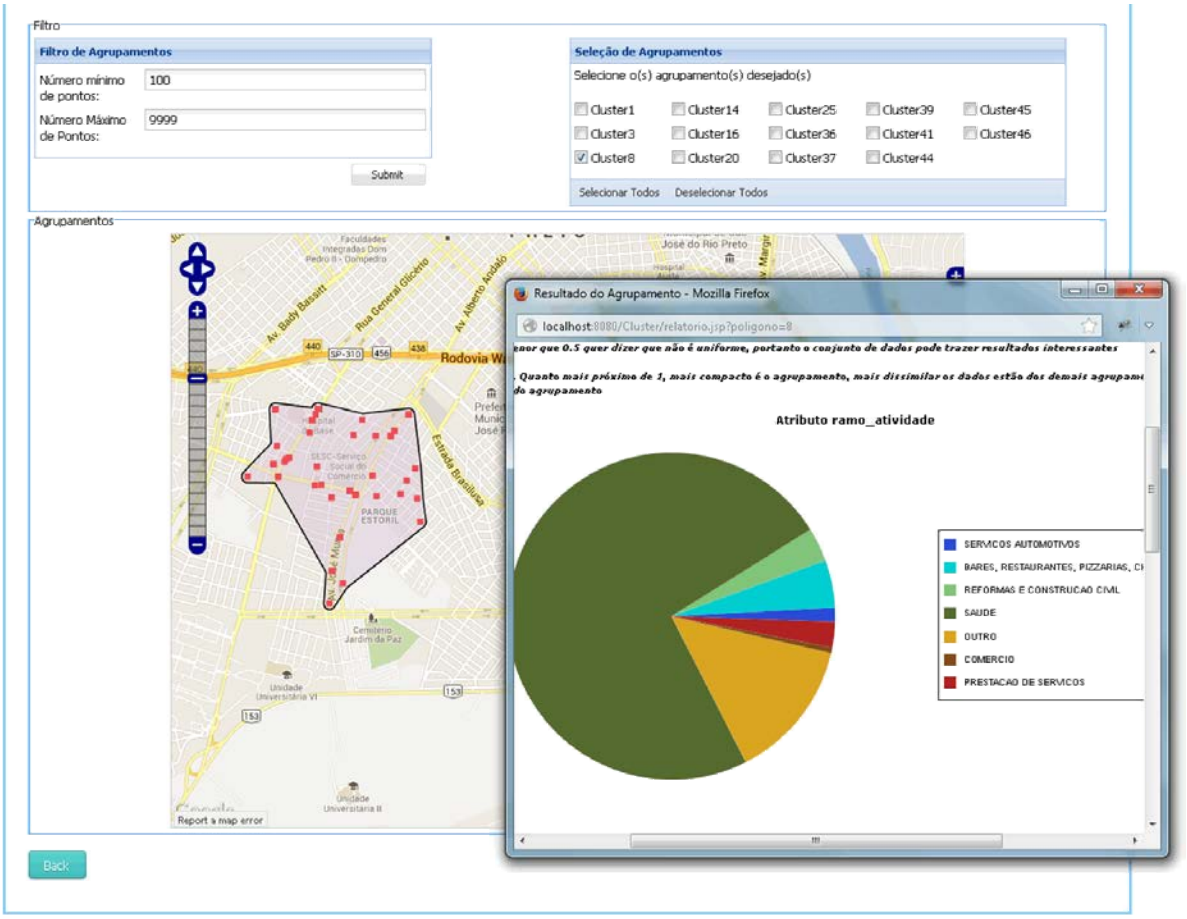

Figura 43: Aplicação dos recursos para análise dos agrupamentos gerados.

No segundo experimento, a similaridade não foi considerada para geração dos agrupamentos, sendo assim nenhum atributo não espacial foi selecionado. Portanto, os agrupamentos gerados nesta seção são baseados apenas no atributo espacial, que neste caso corresponde ao local do acidente de trabalho. Na Figura 44 são apresentados os agrupamentos gerados.

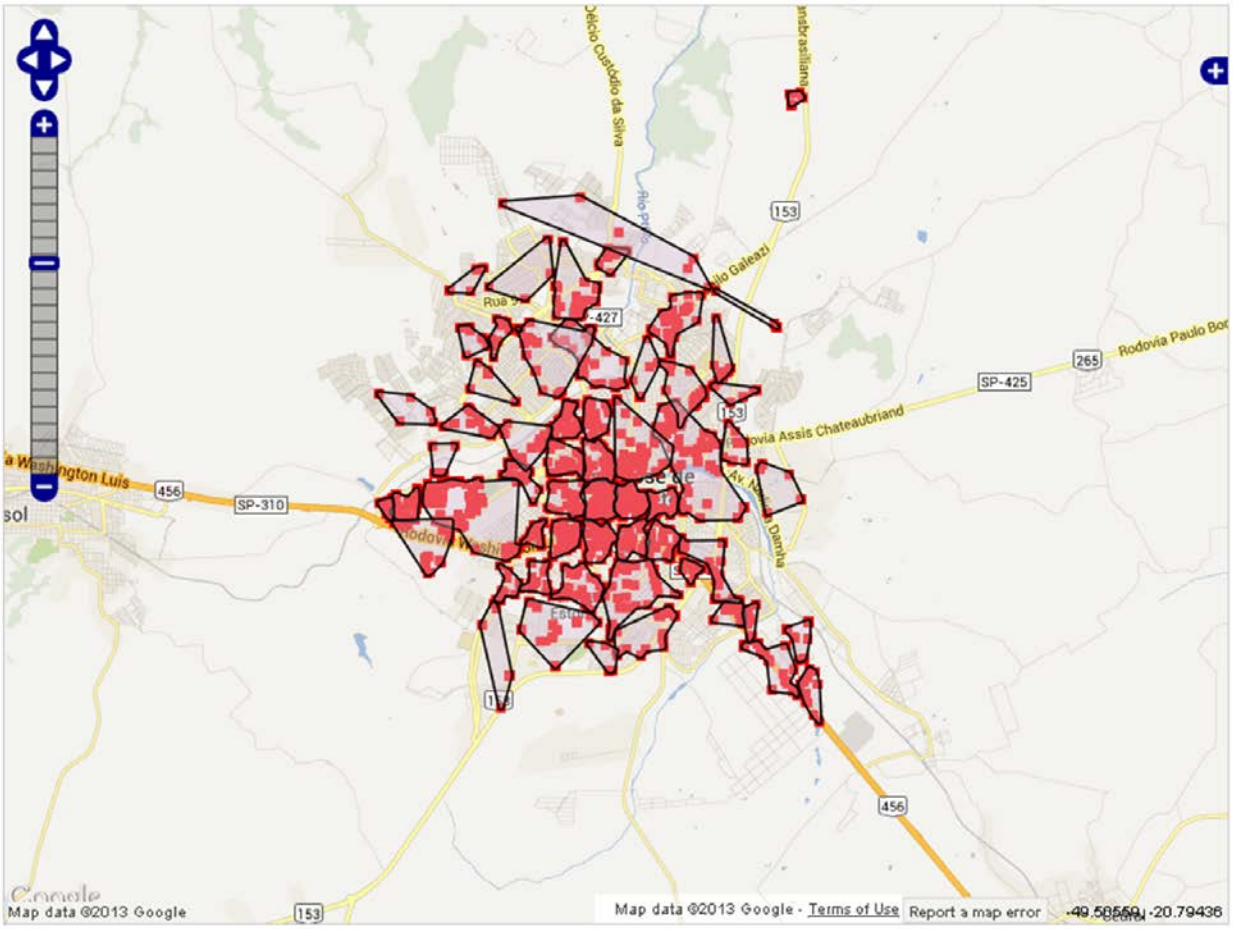

Figura 44: Agrupamentos gerados sem considerar a similaridade.

Como o experimento 2 não considerou a similaridade, a ferramenta FADE não gera os gráficos para análise desses agrupamentos. Para visualizar a distribuição das características não espaciais seria necessário realizar a busca no banco de dados. Para exemplificar isso, foi realizada uma consulta no banco de dados para analisar a característica ramo de atividade, uma vez que esta foi utilizada com critério de similaridade para geração dos agrupamentos no experimento 1. Ao realizar a análise dos resultados de forma manual é necessário verificar todas as opções para cada característica não espacial utilizada, que no caso de ramo de atividade são: Saúde; Bares, Restaurantes, Pizzarias, Churrascarias e Lanchonetes; Supermercado, Mercado, Mercearia e Mini Mercado; Metalúrgica; Prestação de Serviços; Serviços Automotivos; Reformas e Construção; Indústria e Comercio de Móveis; Frigorífico; Comércio; Outros. A consulta é apresentada na Figura 45 e os resultados obtidos são exibidos na Figura 46.

```
select distinct cluster,count(*) as qtd,
    (round(sum(
                CASE WHEN ramo_atividade='SAUDE'
                     then 1 else 0 end)*100.0 / sum(1),2)
    ) as ramo_atividade_saude,
    (round(sum(
                CASE WHEN ramo_atividade='BARES, RESTAURANTES, PIZZARIAS, CHURRASCARIA E LANCHONETES'
                     then 1 else 0 end)*100.0 / sum(1),2)
    ) as ramo_atividade_bares,
    (round(sum(
                CASE WHEN ramo_atividade='SUPERMERCADO, MERCADO, MERCEARIA E MINI MERCADO'
                     then 1 else 0 end)*100.0 / sum(1),2)
    ) as ramo_atividade_super,
    (round(sum(
                CASE WHEN ramo_atividade='METALURGICA'
                then 1 else 0 end)*100.0 / sum(1),2)
    ) as ramo_atividade_metalurgica,
    (round(sum(
                CASE WHEN ramo_atividade='PRESTACAO DE SERVICOS'
                then 1 else 0 end)*100.0 / sum(1),2)
    ) as ramo_atividade_servicos,
    (round(sum(
                CASE WHEN ramo_atividade='SERVICOS AUTOMOTIVOS'
                then 1 else 0 end)*100.0 / sum(1),2)
    ) as ramo_atividade_automotivos,
    (round(sum(
                CASE WHEN ramo_atividade='REFORMAS E CONSTRUCAO CIVIL'
                then 1 else 0 end)*100.0 / sum(1),2)
    ) as ramo_atividade_construcao,
    (round(sum(
                CASE WHEN ramo_atividade='INDUSTRIA E COMERCIO DE MOVEIS'
                then 1 else 0 end)*100.0 / sum(1),2)
    ) as ramo_atividade_moveis,
    (round(sum(
                CASE WHEN ramo_atividade='FRIGORIFICO' then 1 else 0 end)*100.0 / sum(1),2)
    ) as ramo_atividade_frigorifico,
    (round(sum(
                CASE WHEN ramo_atividade='COMERCIO' then 1 else 0 end)*100.0 / sum(1),2)
    ) as ramo_atividade_comercio,
    (round(sum(
                CASE WHEN ramo_atividade='OUTRO' then 1 else 0 end)*100.0 / sum(1),2)
    ) as ramo_atividade_outros
from resultado
JOIN dm_rp2012 USING (id)
group by cluster
order by cluster;
```

Figura 45: Consulta na base de dados do SIVAT.

	cluster integer	qtd bigint	saude numeric	bares numeric	super numeric	metalurgica numeric	servicos numeric	automotivos numeric	construcao numeric	moveis numeric	frigorifico numeric	comercio numeric	outros numeric
1	0	51	0.00	31.37	37.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	31.37
2	1	46	2.17	0.00	13.04	13.04	0.00	4.35	0.00	0.00	0.00	2.17	65.22
3	2	11	0.00	9.09	0.00	0.00	0.00	9.09	9.09	0.00	0.00	0.00	72.73
4	3	129	0.78	0.00	0.00	58.91	0.00	4.65	4.65	0.00	0.00	4.65	26.36
5	4	74	0.00	20.27	6.76	0.00	5.41	1.35	12.16	0.00	0.00	8.11	45.95
6	5	28	0.00	0.00	3.57	0.00	0.00	28.57	0.00	0.00	0.00	0.00	67.86
7	6	65	3.08	3.08	20.00	4.62	1.54	6.15	10.77	0.00	0.00	0.00	50.77
8	7	61	1.64	19.67	1.64	0.00	22.95	6.56	4.92	0.00	0.00	1.64	40.98
9	8	24	4.17	12.50	0.00	0.00	8.33	4.17	20.83	0.00	0.00	0.00	50.00
10	9	44	2.27	0.00	0.00	27.27	2.27	13.64	0.00	0.00	0.00	11.36	43.18
11	10	11	9.09	27.27	0.00	0.00	0.00	18.18	0.00	0.00	0.00	0.00	45.45
12	11	341	0.00	2.93	0.00	43.70	0.88	0.00	1.76	0.00	0.00	2.64	48.09
13	12	36	0.00	8.33	5.56	11.11	0.00	22.22	8.33	0.00	0.00	5.56	38.89
14	13	70	52.86	0.00	0.00	2.86	0.00	0.00	5.71	0.00	0.00	1.43	37.14
15	14	78	17.95	11.54	0.00	1.28	0.00	1.28	1.28	0.00	0.00	2.56	64.10
16	15	60	0.00	5.00	3.33	75.00	0.00	3.33	0.00	0.00	0.00	1.67	11.67
17	16	166	0.00	0.60	0.00	15.66	0.60	0.00	26.51	0.00	0.00	9.04	47.59
18	17	53	0.00	3.77	0.00	1.89	1.89	3.77	28.30	0.00	0.00	1.89	58.49
19	18	20	0.00	5.00	0.00	0.00	20.00	20.00	0.00	0.00	0.00	5.00	50.00
20	19	70	37.14	20.00	0.00	0.00	4.29	2.86	2.86	0.00	0.00	4.29	28.57
21	20	43	0.00	0.00	0.00	30.23	0.00	0.00	0.00	0.00	0.00	0.00	69.77
22	21	125	16.80	14.40	0.00	2.40	8.00	1.60	4.00	0.00	0.00	0.80	52.00
23	22	131	10.69	21.37	6.87	0.00	0.00	1.53	5.34	0.00	0.00	2.29	51.91
24	23	48	4.17	10.42	35.42	0.00	10.42	2.08	2.08	0.00	0.00	2.08	33.33
25	24	110	2.73	1.82	58.18	4.55	0.91	0.91	2.73	0.00	0.00	0.00	28.18
26	25	37	0.00	0.00	0.00	70.27	0.00	0.00	0.00	0.00	0.00	0.00	29.73
27	26	75	0.00	0.00	0.00	49.33	0.00	0.00	0.00	0.00	0.00	5.33	45.33
28	27	166	24.70	4.22	20.48	1.81	4.82	3.61	0.60	0.00	0.00	5.42	34.34
29	28	34	0.00	2.94	58.82	0.00	0.00	0.00	11.76	0.00	0.00	2.94	23.53
30	29	66	0.00	0.00	0.00	3.03	0.00	0.00	1.52	0.00	0.00	0.00	95.45
31	30	59	0.00	0.00	0.00	49.15	0.00	0.00	0.00	0.00	0.00	0.00	50.85
32	31	25	4.00	4.00	0.00	12.00	0.00	0.00	4.00	0.00	0.00	4.00	72.00
33	32	43	2.33	2.33	0.00	4.65	4.65	2.33	11.63	0.00	0.00	0.00	72.09
34	33	28	7.14	3.57	28.57	0.00	0.00	7.14	0.00	0.00	0.00	0.00	53.57
35	34	52	0.00	9.62	17.31	0.00	1.92	3.85	0.00	0.00	0.00	1.92	65.38
36	35	121	19.01	7.44	2.48	1.65	14.05	1.65	0.83	0.00	3.31	9.09	40.50
37	36	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
38	37	10	0.00	0.00	0.00	0.00	0.00	30.00	0.00	0.00	0.00	0.00	70.00
39	38	37	0.00	0.00	0.00	32.43	0.00	8.11	0.00	0.00	0.00	2.70	56.76
40	39	49	20.41	6.12	0.00	6.12	2.04	8.16	6.12	0.00	2.04	4.08	44.90
41	40	89	1.12	0.00	14.61	2.25	2.25	1.12	1.12	0.00	55.06	0.00	22.47
42	41	16	0.00	6.25	0.00	31.25	0.00	6.25	12.50	0.00	0.00	0.00	43.75
43	42	114	44.74	0.00	0.00	13.16	4.39	6.14	3.51	0.00	0.00	1.75	26.32
44	43	15	0.00	6.67	0.00	0.00	0.00	66.67	0.00	0.00	0.00	0.00	26.67
45	44	66	0.00	12.12	37.88	1.52	0.00	16.67	0.00	0.00	0.00	0.00	31.82
46	45	9	0.00	0.00	0.00	11.11	0.00	11.11	11.11	0.00	0.00	11.11	55.56
47	46	29	3.45	0.00	0.00	13.79	0.00	20.69	0.00	0.00	0.00	0.00	62.07
48	47	35	0.00	0.00	0.00	20.00	0.00	0.00	0.00	0.00	0.00	0.00	80.00
49	48	67	5.97	0.00	1.49	10.45	0.00	1.49	4.48	0.00	16.42	0.00	59.70
50	49	22	0.00	0.00	0.00	0.00	0.00	4.55	0.00	0.00	0.00	0.00	95.45
51	50	7	14.29	0.00	0.00	14.29	0.00	0.00	28.57	0.00	0.00	0.00	42.86
52	51	13	0.00	0.00	0.00	15.38	15.38	0.00	0.00	0.00	0.00	0.00	69.23

Figura 46: Resultado da consulta para verificar o resultado sem a estratégia de similaridade.

Ao analisar os dados verifica-se que apenas três agrupamentos obtiveram similaridade superior a 90%. Com o propósito de comparar tais resultados com os resultados obtidos ao aplicar a similaridade, foi realizada a mesma consulta sobre os resultados do experimento 1, sendo que os resultados são apresentados na Figura 47.

	cluster integer	qtd bigint	saude numeric	bares numeric	super numeric	metalurgica numeric	servicos numeric	automotivos numeric	construcao numeric	moveis numeric	frigorifico numeric	comercio numeric	outros numeric
1	0	22	0.00	0.00	0.00	0.00	0.00	9.09	0.00	0.00	0.00	0.00	90.91
2	1	171	0.00	0.00	2.92	47.95	0.00	8.19	4.09	0.00	0.00	5.85	30.99
3	2	52	0.00	30.77	36.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	32.69
4	3	115	2.61	0.00	0.00	39.13	0.00	2.61	0.00	0.00	0.00	4.35	51.30
5	4	15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
6	5	44	0.00	0.00	13.64	9.09	0.00	0.00	0.00	0.00	0.00	0.00	77.27
7	6	13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
8	7	49	0.00	30.61	10.20	0.00	0.00	0.00	4.08	0.00	0.00	12.24	42.86
9	8	388	73.71	4.64	0.00	0.00	2.32	1.55	3.35	0.00	0.00	0.52	13.92
10	9	24	8.33	0.00	0.00	0.00	45.83	0.00	20.83	0.00	0.00	0.00	25.00
11	10	40	0.00	0.00	32.50	0.00	0.00	7.50	5.00	0.00	0.00	0.00	55.00
12	11	38	0.00	0.00	0.00	18.42	0.00	15.79	0.00	0.00	0.00	26.32	39.47
13	12	10	0.00	40.00	0.00	0.00	0.00	10.00	0.00	0.00	0.00	0.00	50.00
14	13	11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
15	14	283	18.37	4.95	7.42	7.07	1.06	2.83	5.30	0.00	0.00	6.71	46.29
16	15	49	2.04	0.00	16.33	0.00	4.08	4.08	4.08	0.00	0.00	2.04	67.35
17	16	149	12.08	2.01	8.72	4.70	1.34	2.68	1.34	0.00	32.89	0.00	34.23
18	17	58	15.52	6.90	0.00	3.45	0.00	8.62	10.34	0.00	0.00	3.45	51.72
19	18	55	0.00	0.00	0.00	12.73	0.00	5.45	3.64	0.00	0.00	0.00	78.18
20	19	11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
21	20	128	3.91	1.56	46.88	0.00	1.56	4.69	0.00	0.00	0.00	0.00	41.41
22	21	7	0.00	0.00	0.00	28.57	0.00	0.00	28.57	0.00	0.00	0.00	42.86
23	22	40	5.00	0.00	40.00	25.00	0.00	0.00	0.00	0.00	0.00	0.00	30.00
24	23	26	15.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	84.62
25	24	26	0.00	0.00	50.00	0.00	0.00	0.00	3.85	0.00	0.00	0.00	46.15
26	25	203	29.56	6.90	8.37	0.00	5.42	1.97	9.85	0.00	0.00	3.94	33.99
27	26	68	0.00	0.00	19.12	26.47	0.00	14.71	1.47	0.00	0.00	2.94	35.29
28	27	11	0.00	0.00	0.00	63.64	0.00	0.00	0.00	0.00	0.00	0.00	36.36
29	28	35	0.00	22.86	5.71	0.00	8.57	8.57	0.00	0.00	0.00	0.00	54.29
30	29	16	0.00	6.25	0.00	31.25	0.00	0.00	12.50	0.00	0.00	0.00	50.00
31	30	35	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
32	31	59	0.00	0.00	42.37	0.00	0.00	35.59	0.00	0.00	0.00	0.00	22.03
33	32	68	0.00	4.41	0.00	2.94	0.00	0.00	2.94	0.00	0.00	0.00	89.71
34	33	28	0.00	0.00	71.43	0.00	0.00	0.00	10.71	0.00	0.00	0.00	17.86
35	34	20	10.00	0.00	0.00	20.00	0.00	0.00	0.00	0.00	0.00	0.00	70.00
36	35	7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
37	36	394	0.00	0.00	0.00	52.54	1.02	1.27	1.52	0.00	0.00	2.03	41.62
38	37	123	61.79	3.25	7.32	0.00	0.00	1.63	0.00	0.00	0.00	0.00	26.02
39	38	22	0.00	0.00	0.00	9.09	13.64	0.00	0.00	0.00	0.00	9.09	68.18
40	39	135	29.63	11.11	0.00	2.96	5.93	1.48	5.19	0.00	0.00	3.70	40.00
41	40	72	4.17	11.11	4.17	2.78	20.83	0.00	0.00	0.00	5.56	13.89	37.50
42	41	114	16.67	16.67	0.00	2.63	7.02	1.75	4.39	0.00	0.00	0.00	50.88
43	42	39	0.00	25.64	0.00	0.00	0.00	5.13	0.00	0.00	0.00	0.00	69.23
44	43	61	0.00	3.28	3.28	77.05	0.00	1.64	0.00	0.00	0.00	0.00	14.75
45	44	156	0.00	0.00	0.00	14.74	0.00	0.00	28.21	0.00	0.00	10.26	46.79
46	45	139	0.00	0.00	0.00	79.86	0.00	0.00	0.00	0.00	0.00	0.00	20.14
47	46	114	10.53	21.93	7.89	0.00	0.00	2.63	5.26	0.00	0.00	2.63	49.12
48	47	9	0.00	0.00	11.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	88.89
49	48	17	17.65	0.00	0.00	23.53	0.00	0.00	11.76	0.00	0.00	0.00	47.06
50	49	7	0.00	0.00	42.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	57.14
51	50	53	0.00	0.00	0.00	33.96	0.00	0.00	0.00	0.00	0.00	7.55	58.49
52	51	18	0.00	0.00	0.00	44.44	0.00	22.22	11.11	0.00	0.00	0.00	22.22
53	52	26	0.00	0.00	0.00	26.92	0.00	0.00	0.00	0.00	0.00	0.00	73.08
54	53	32	0.00	0.00	0.00	12.50	0.00	81.25	6.25	0.00	0.00	0.00	0.00

Figura 47: Resultado da consulta para verificar o resultado com a estratégia de similaridade.

Na Figura 47 verifica-se que com a aplicação da similaridade foi possível obter sete agrupamentos com similaridade superior a 90%, sendo que um dos agrupamentos obteve 100% da similaridade na área da saúde, o que é bastante interessante ao passo que a área da saúde corresponde a apenas 14,61% de todo o conjunto de dados analisado, conforme pode ser observado na Figura 48.

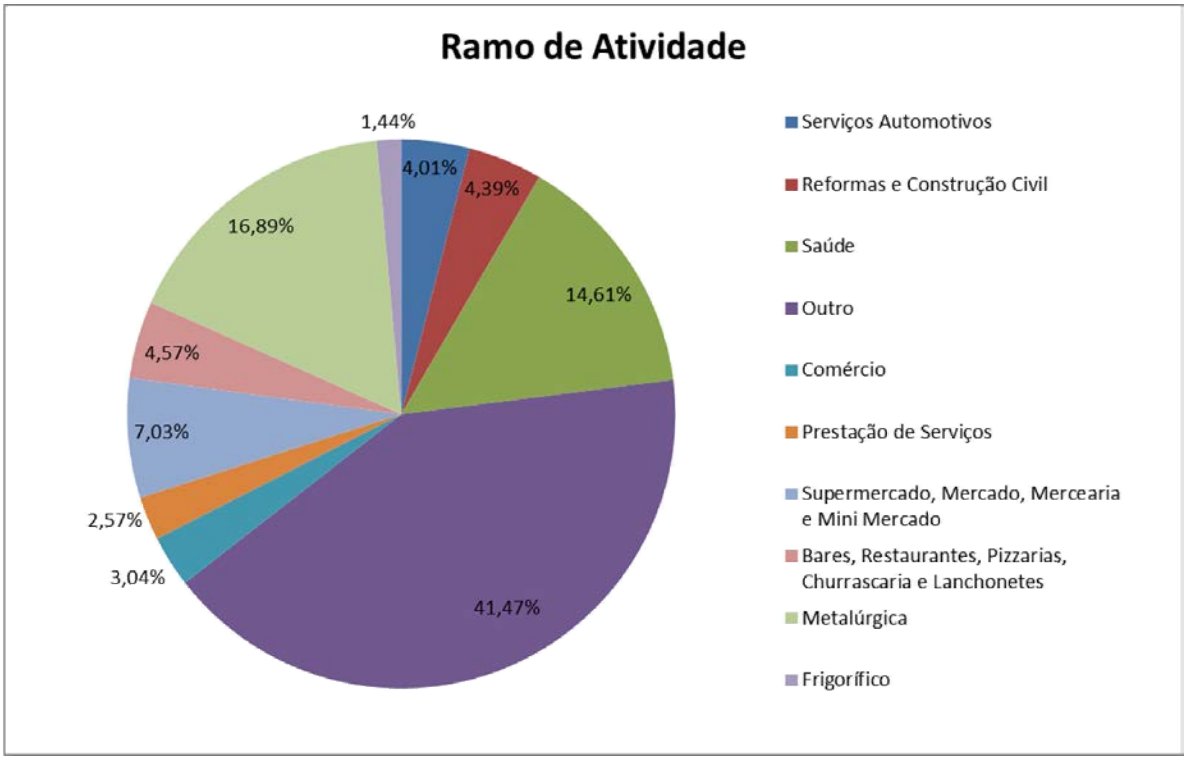

Figura 48: Distribuição do ramo de atividade no conjunto de dados.

Os resultados apresentados nos experimentos 1 e 2 mostram que embora seja possível analisar os agrupamentos de dados espaciais sem considerar a similaridade, na maioria dos casos é difícil obter algum conhecimento útil com esta metodologia, uma vez que este processo de interação usuário e base de dados é muito mais penoso e na maioria das vezes não obtém o mesmo nível de similaridade, ao passo que ao aplicar o algoritmo sem considerar a similaridade as características dos dados agrupados tendem a ser mais diversificadas do que quando a similaridade é considerada. Assim, para a maioria das aplicações a utilização da similaridade contribui para extração de conhecimento útil e facilita a correlação dos atributos espaciais e não espaciais.

5.4 Comparativo do CHSMST^{+} com VDBSCAN

Nesta seção são apresentados alguns dos experimentos realizados com o propósito de comparar os resultados alcançados pelo CHSMST^{+} com o algoritmo VSBSCAN, uma vez que no estudo comparativo apresentado na Seção 3.7, o VDBSCAN foi verificado com um dos melhores algoritmos de agrupamento de dados espaciais.

Foram selecionados 4.512 registros georreferenciados de acidentes que ocorreram durante o ano de 2012 no município de São José do Rio Preto. Para efeito de comparação dos algoritmos, foi utilizada a mesma equação de similaridade no processo de geração dos agru-

pamentos para o algoritmo CHSMST+ e o VDBSCAN, sendo utilizado como critério para geração dos agrupamentos o atributo espacial referente ao local do acidente com a variação de um, dois e três atributos não espaciais. O algoritmo VDBSCAN requer como parâmetro de entrada a variável k (número mínimo de pontos no agrupamento), que neste caso foi atribuído o valor 3, uma vez que foram realizados experimentos que comprovaram que para k=3, o algoritmo é capaz de gerar agrupamentos satisfatórios (LIU; ZHOU; WU, 2007). A partir disso, foi calculada a média da qualidade em relação à distância, a média da qualidade em relação à similaridade e a qualidade média dos agrupamentos. Os resultados obtidos com o algoritmo CHSMST+ são apresentados na Tabela 11 e os resultados referentes à aplicação do VDBSCAN são exibidos na Tabela 12.

Ao comparar os resultados apresentados pelos algoritmos verifica-se que para todos os casos, a qualidade média dos agrupamentos gerados pelo CHSMST^{+} foi superior que os agrupamentos obtidos com VDBSCAN, sendo que o algoritmo desenvolvido obteve em média 0,91 no cálculo do coeficiente da silhueta, enquanto que o segundo algoritmo obteve 0,72.

A seguir são discutidos os agrupamentos gerados pelos algoritmos, ao considerar como critério para geração dos agrupamentos, os atributos não espaciais: afastamento e tipo de acidente. Os resultados obtidos com o VDBSCAN foram projetados na ferramenta desenvolvida, conforme é ilustrado na Figura 49. Na Figura 50 são apresentados os resultados obtidos com o CHSMST^{+}.

Tabela 11: Algoritmo CHSMST^{+}.

Atributos Não Espaciais	Qualidade Similaridade	Qualidade Distância	Qualidade Média
Afastamento	0,98	0,93	0,96
Afastamento e Tipo de Acidente	0,86	0,97	0,92
Afastamento e Faixa Etária	0,88	0,90	0,89
Local do Acidente e Afastamento	0,88	0,88	0,88
Tipo de Acidente, Afastamento e Escolaridade	0,86	0,98	0,92
Média	0,89	0,93	0,91
Valor Ideal	1,00	1,00	1,00

Tabela 12: Algoritmo VDBSCAN.

Atributos Não Espaciais	Qualidade Similaridade	Qualidade Distância	Qualidade Média
Afastamento	0,53	0,88	0,71
Afastamento e Tipo de Acidente	0,88	0,54	0,71
Afastamento e Faixa Etária	0,92	0,54	0,73
Local do Acidente e Afastamento	0,47	0,94	0,71
Tipo de Acidente, Afastamento e Escolaridade	0,95	0,54	0,75
Média	0,75	0,69	0,72
Valor Ideal	1,00	1,00	1,00

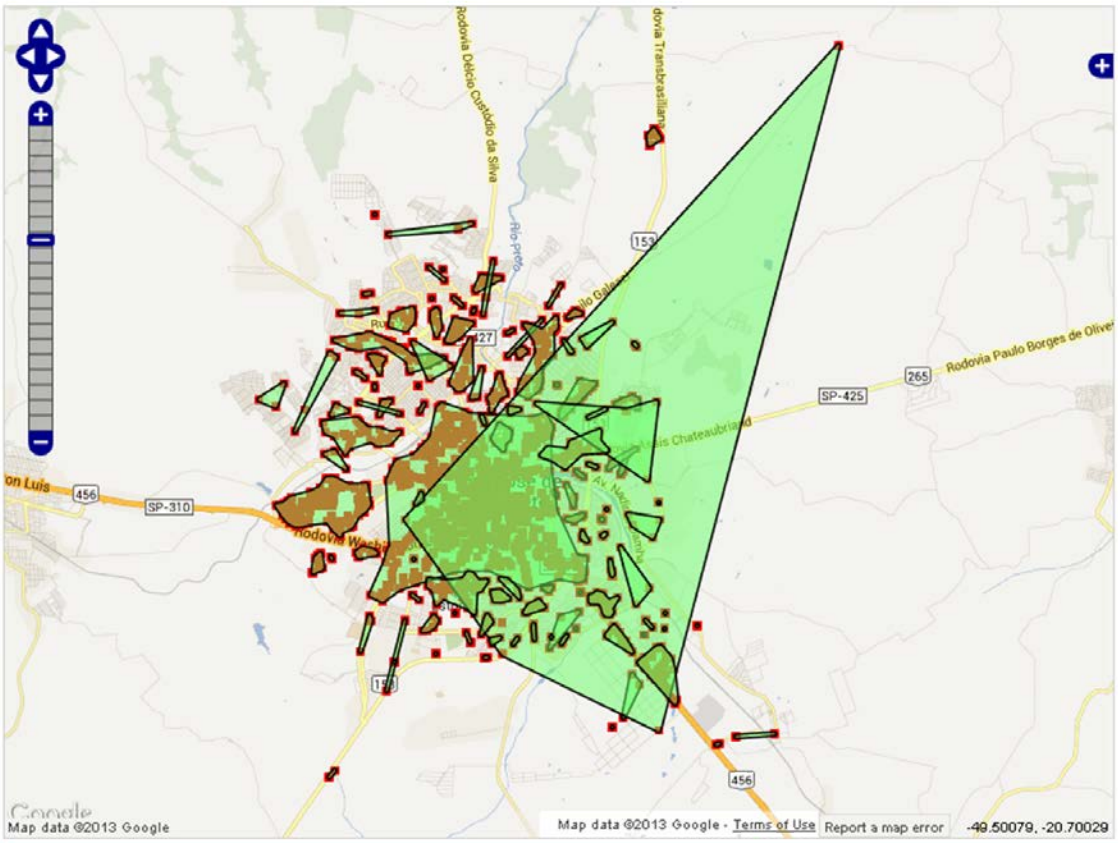

Figura 49: Aplicação do algoritmo VDBSCAN.

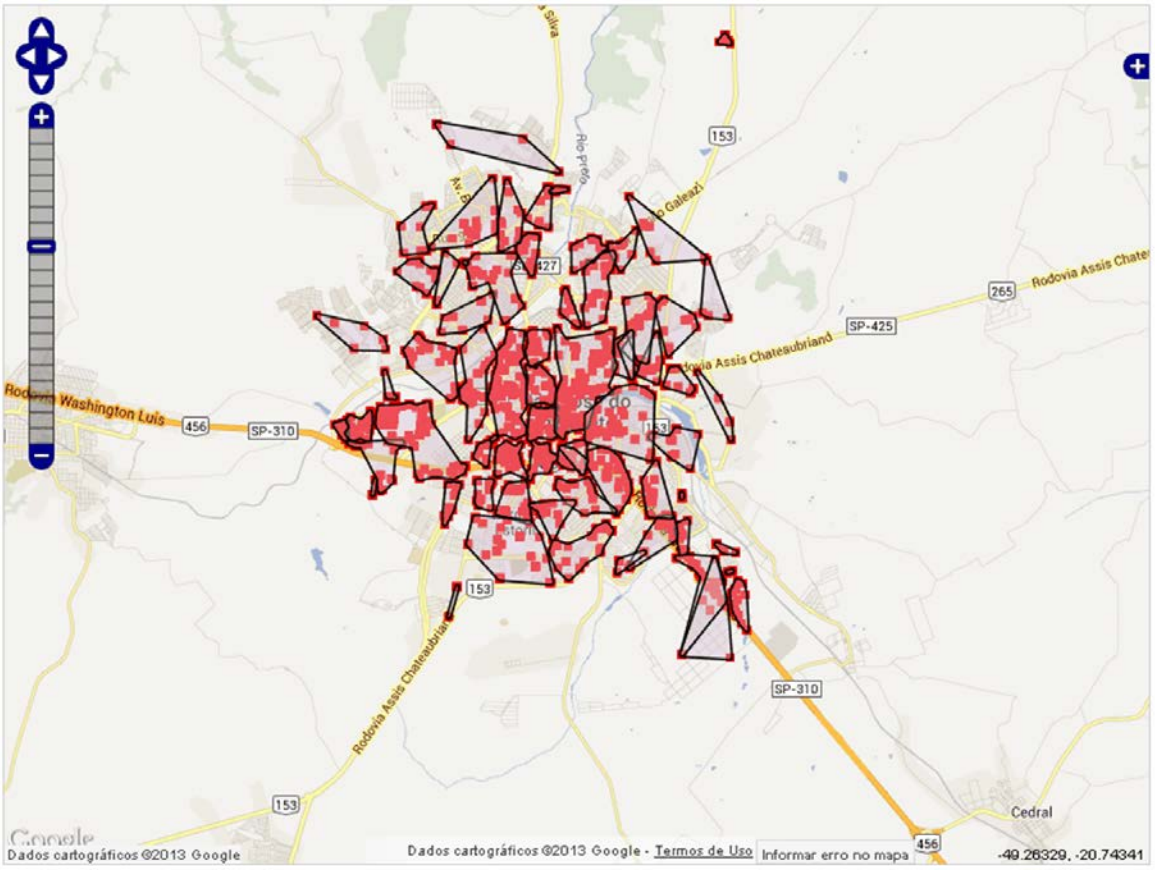

Figura 50: Aplicação do algoritmo CHSMST^{+}.

O VDBSCAN requer a definição do parâmetro k, enquanto que no CHSMST^{+} nenhum parâmetro é necessário. O algoritmo VDBSCAN gerou dois agrupamentos com mais de 200 pontos. Na Figura 51 são apresentados os agrupamentos gerados pelo VDBSCAN denominados como 4 e 5.

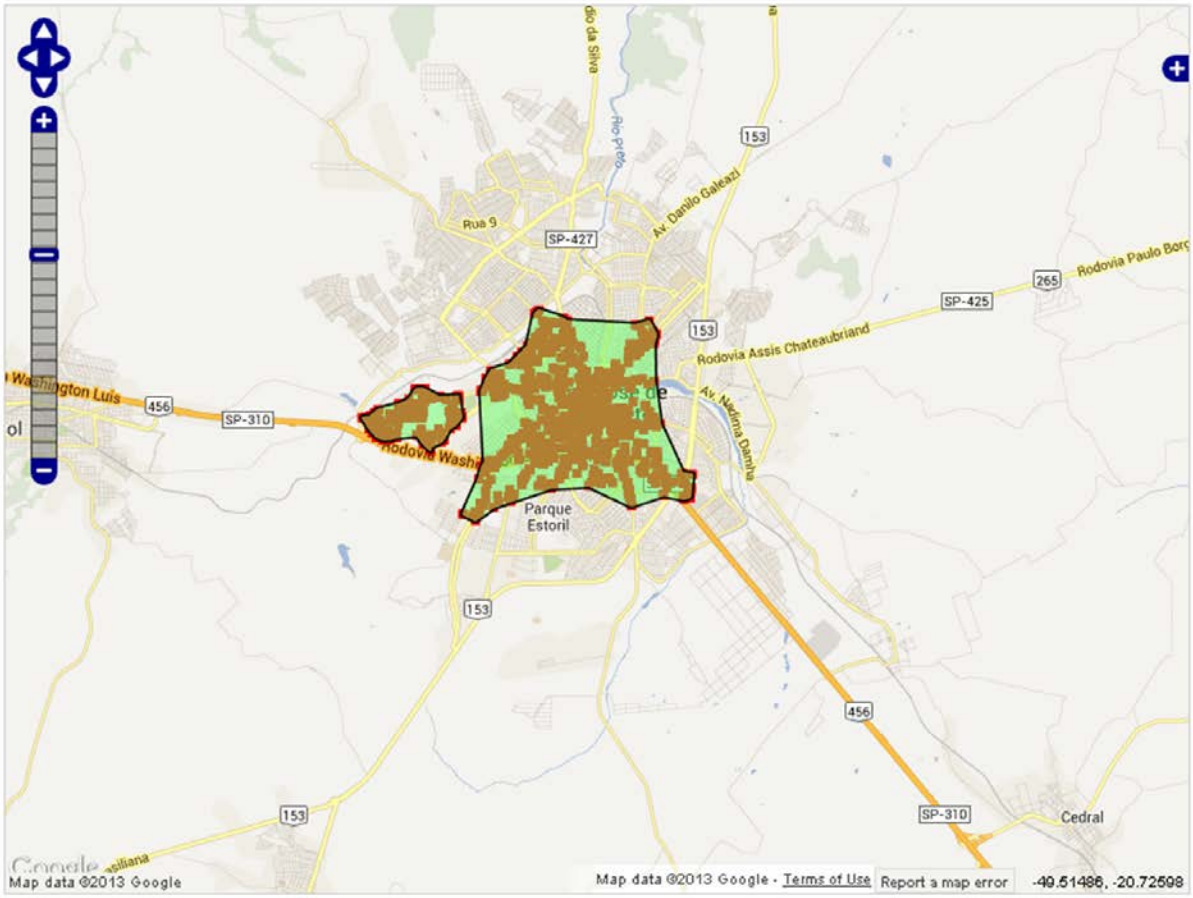

Figura 51: Agrupamentos com mais de 200 pontos gerados pelo VDBSCAN.

O agrupamento 4 contempla 1.961 ocorrências de acidentes de trabalho, que corresponde a 43,46% de todo o conjunto de dados analisado. Tal resultado é indesejado, pois, da mesma forma que é difícil extrair conhecimento de agrupamentos com poucos dados, também é difícil extrair conhecimento de um agrupamento com muitos dados como este, sendo que tal dificuldade é potencializada por este agrupamento corresponder a uma área bastante significativa em relação ao município de São José do Rio Preto. Na Figura 52 e na Figura 53 são apresentados os gráficos referentes aos atributos não espaciais utilizados como critério pelo algoritmo VDBSCAN.

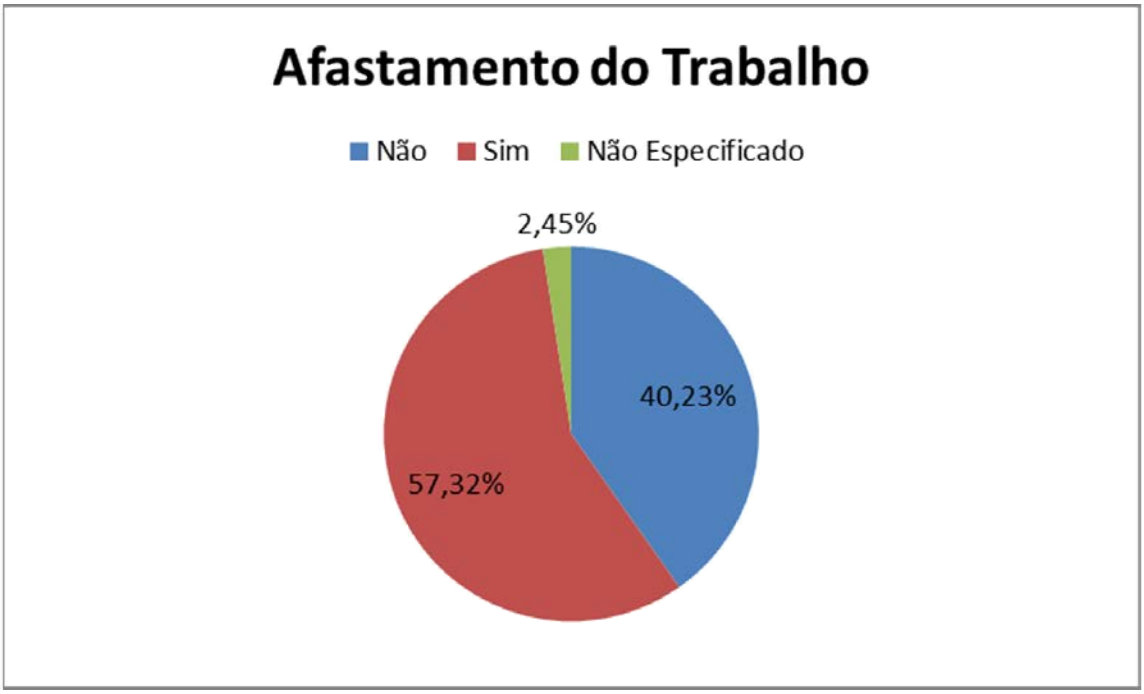

Figura 52: Gráfico referente ao agrupamento 4, considerando o atributo afastamento do trabalho.

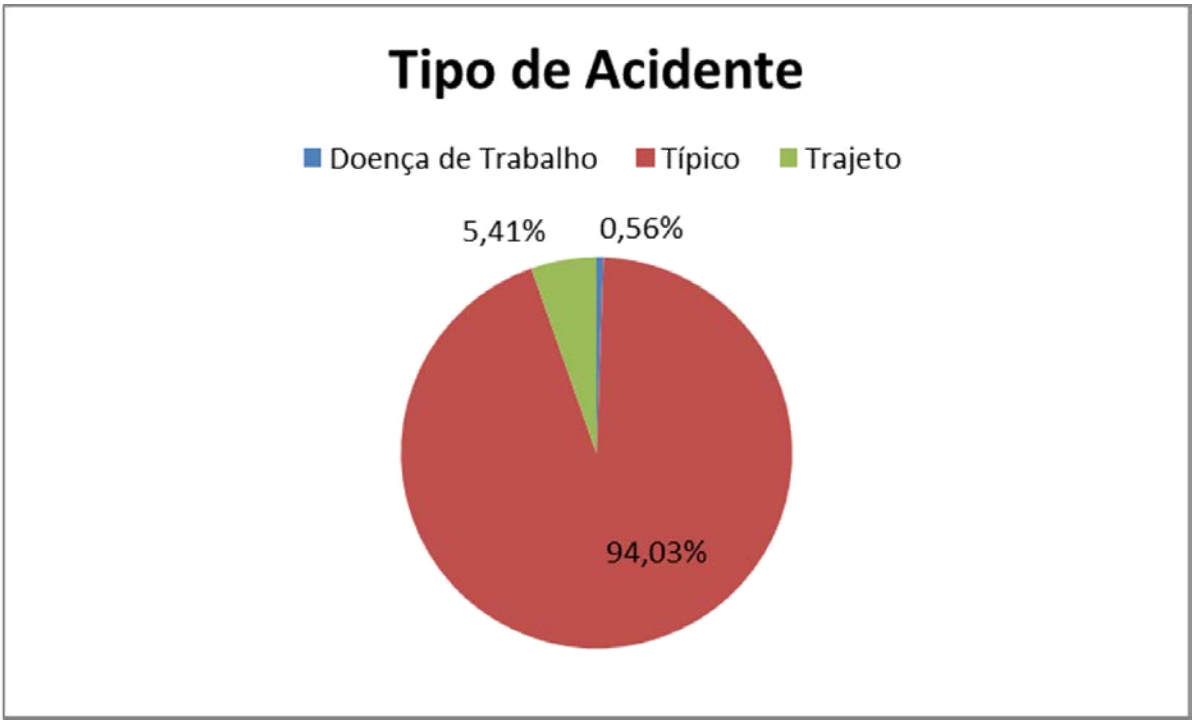

Figura 53: Gráfico referente ao agrupamento 4, considerando o atributo tipo de acidente.

Na Figura 52 verifica-se que 57,32% das ocorrências referentes ao agrupamento 4 ocasionaram afastamento do trabalho e na Figura 53 é observado que 94,03% dos acidentes foram típicos. Como as ocorrências pertencentes a este agrupamento correspondem a mais de 43% da base de dados e contemplam uma grande área referente ao munícipio em que o algoritmo foi aplicado, esses dados não representam uma descoberta interessante para a análise

dos órgãos públicos no sentido de elaborar ações de prevenção ou fiscalização dos acidentes de trabalho. Em relação à eficiência do agrupamento gerado pelo VDBSCAN, a qualidade do agrupamento também não apresentou um resultado desejável ao passo que a qualidade em relação à similaridade foi de 0,77 e em relação à distância foi de 0,09, enquanto que o ideal é que a qualidade seja próxima de 1,00, de acordo com a medida de eficiência utilizada para este cálculo.

O agrupamento 5 é mais interessante, uma vez que contempla 750 ocorrências em uma área menor que do agrupamento 4. Neste agrupamento 73,39% dos acidentados sofreram ausência do trabalho e 97,37% dos acidentes foram típicos, conforme é verificado na Figura 54 e na Figura 55, respectivamente. No entanto, em relação à medida de eficiência o algoritmo não apresentou um bom resultado em relação à distância que foi de 0,66, por outro lado em relação à similaridade a qualidade foi de 0,91.

O experimento realizado com o CHSMST^{+} utilizando os mesmos atributos não espaciais, isto é afastamento e tipo de acidente possibilitou encontrar três agrupamentos com mais de 200 pontos. Na Figura 56 tais agrupamentos são apresentados, denominados como 6, 7 e 8.

O agrupamento 6 contempla 298 ocorrências de acidentes de trabalho, que corresponde a 6,60% de todo o conjunto de dados analisado. Na Figura 57 e na Figura 58 são apresentados os gráficos referentes aos atributos não espaciais utilizados como critério pelo algoritmo CHSMST^{+}.

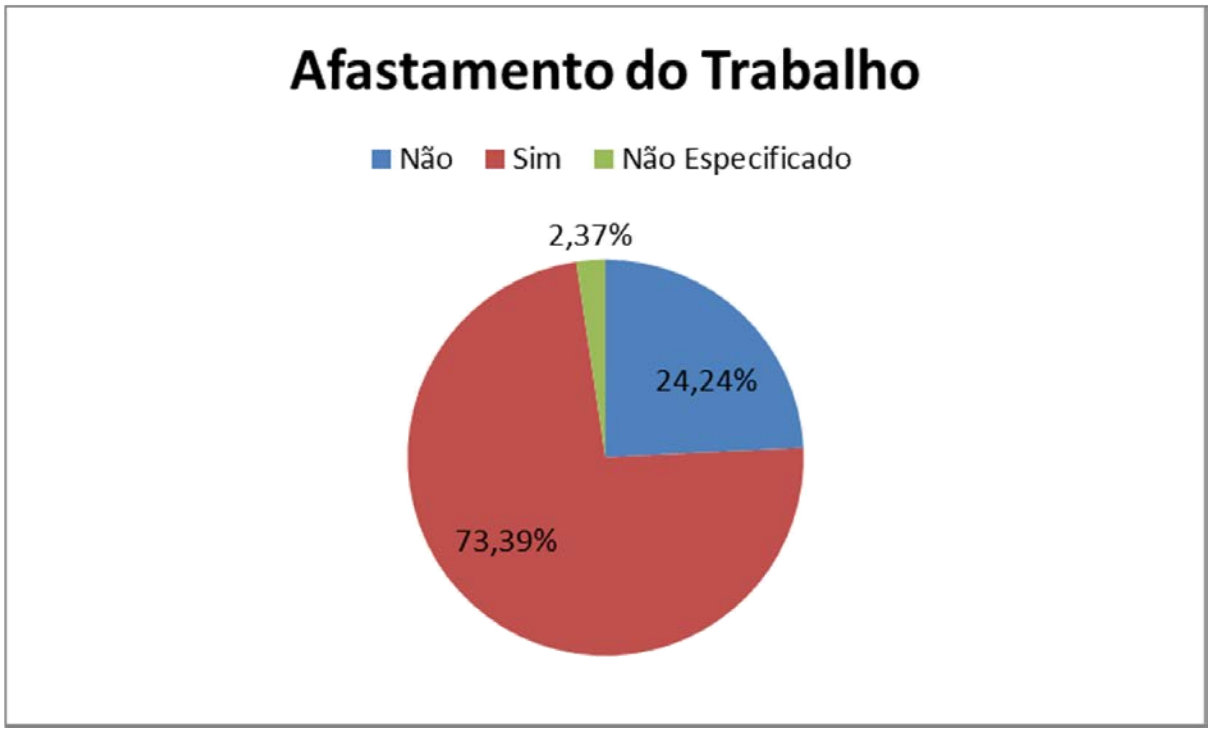

Figura 54: Gráfico referente ao agrupamento 5, considerando o atributo afastamento do trabalho.

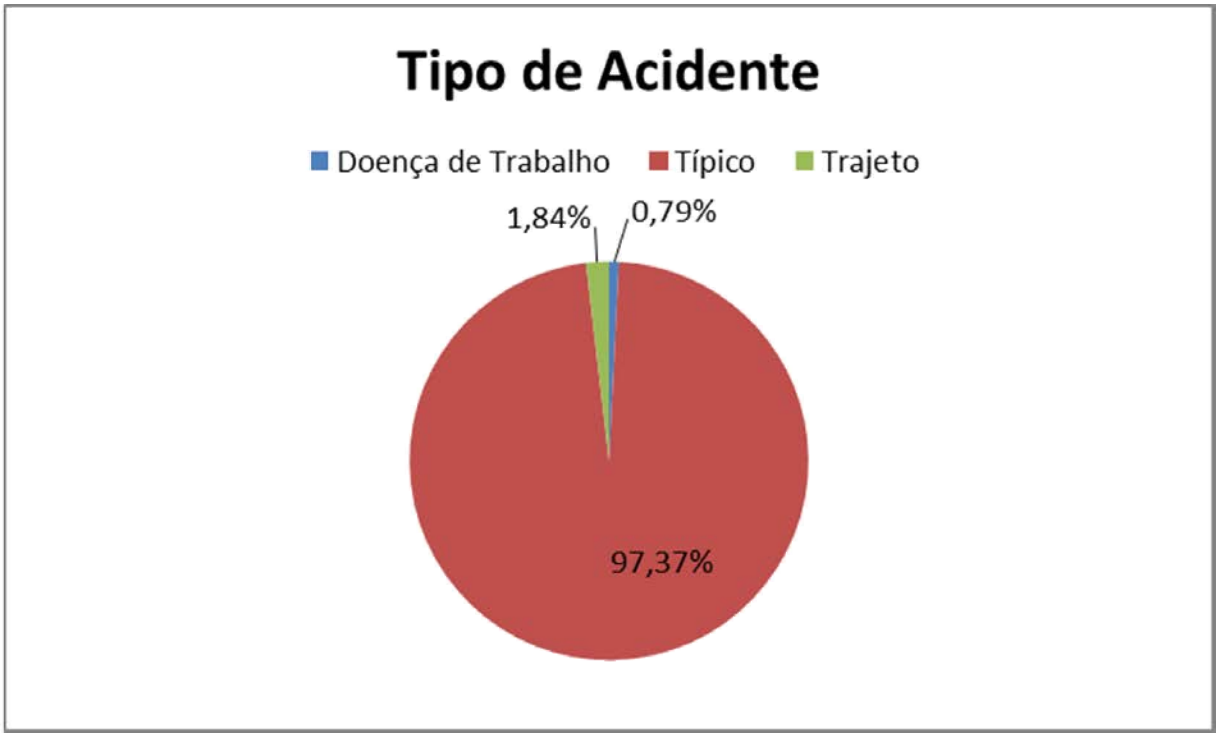

Figura 55: Gráfico referente ao agrupamento 5, considerando o atributo tipo de acidente.

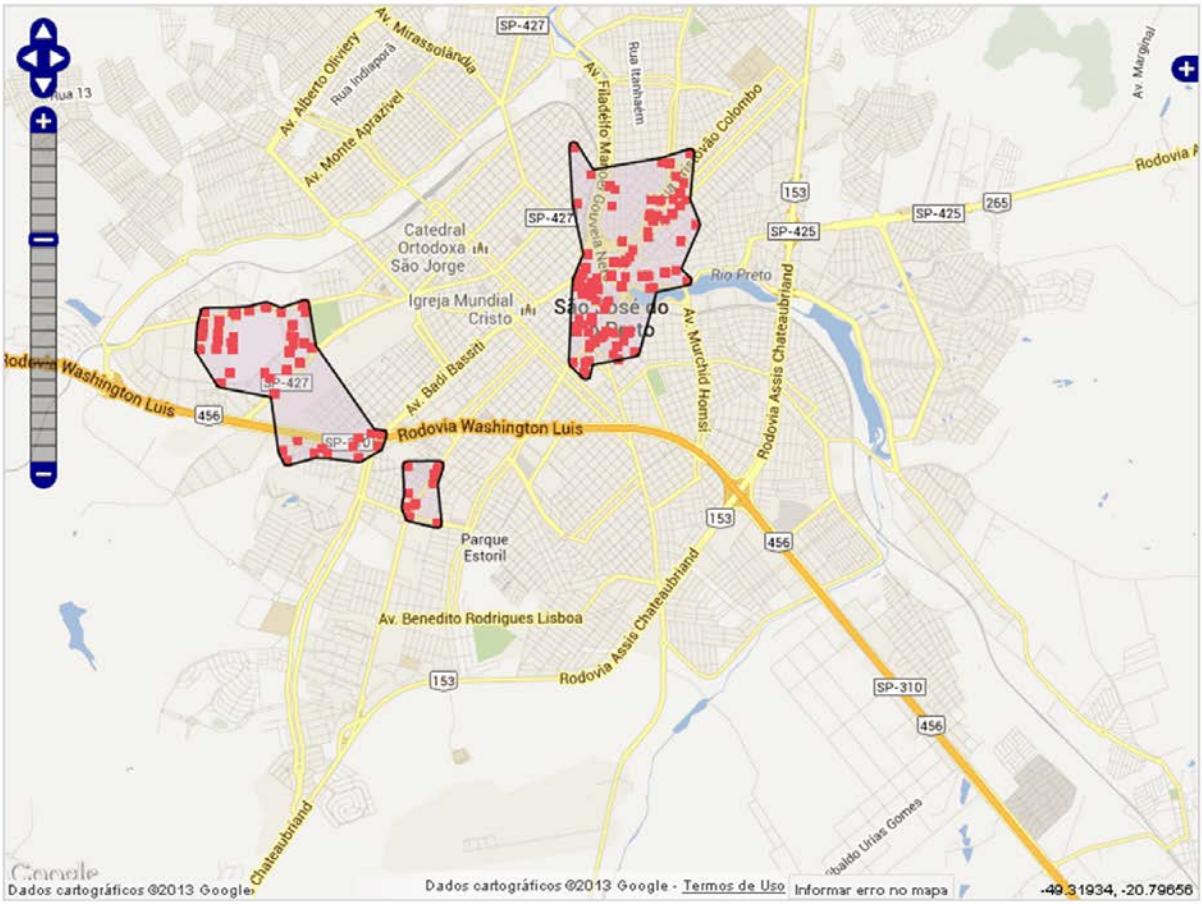

Figura 56: Agrupamentos com mais de 200 pontos gerados pelo CHSMST^{+}.

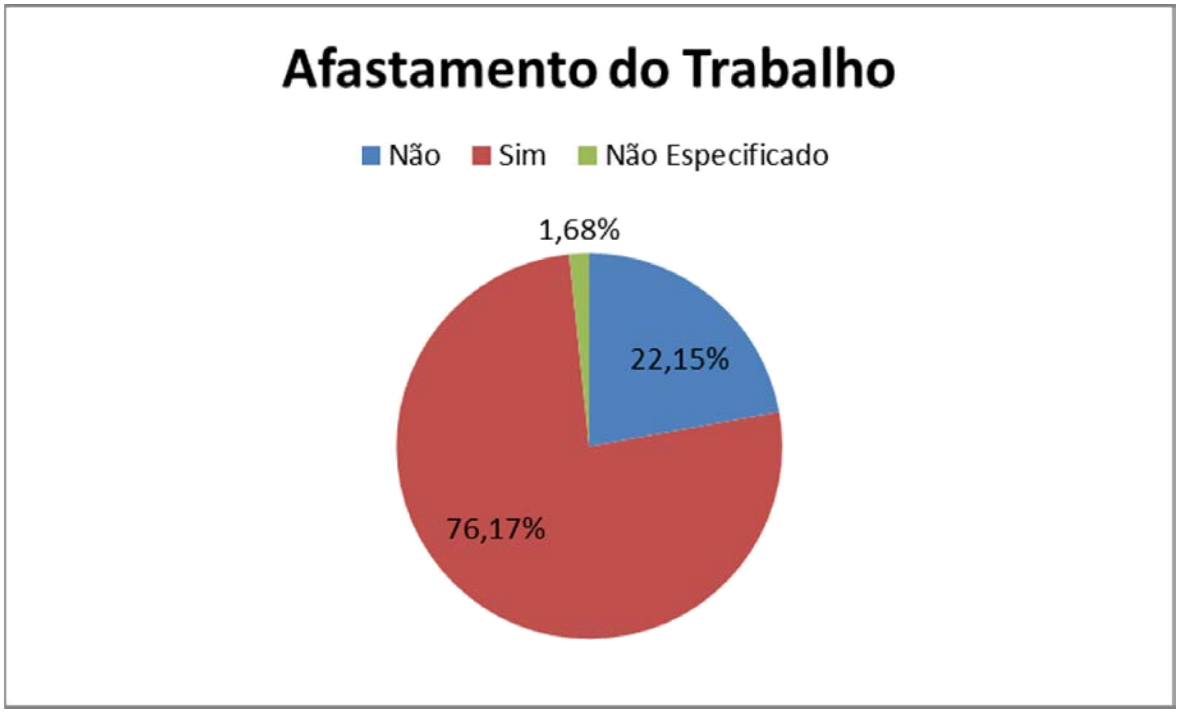

Figura 57: Gráfico referente ao agrupamento 6, considerando o atributo afastamento do trabalho.

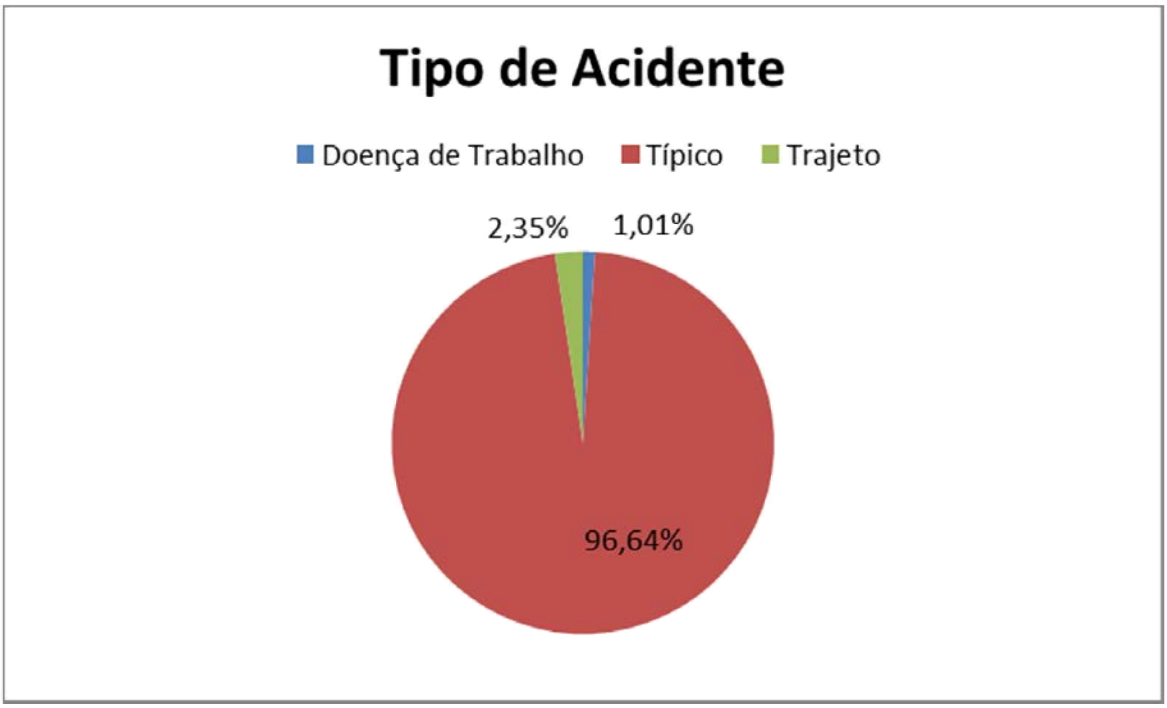

Figura 58: Gráfico referente ao agrupamento 6, considerando o atributo tipo de acidente.

No agrupamento 6, verifica-se que 76,17% dos acidentados sofreram ausência do trabalho e 96,64% dos acidentes foram típicos. O algoritmo CHSMST^{+} obteve como uma boa medida de eficiência em relação à distância que foi de 0,99, por outro lado em relação à similaridade a qualidade foi de 0,64.

O agrupamento 7 contempla 257 ocorrências de acidentes de trabalho, que corresponde a 5,70% de todo o conjunto de dados analisado. Na Figura 59 e na Figura 60 são apresentados os gráficos referentes aos atributos não espaciais utilizados como critério pelo algoritmo CHSMST+.

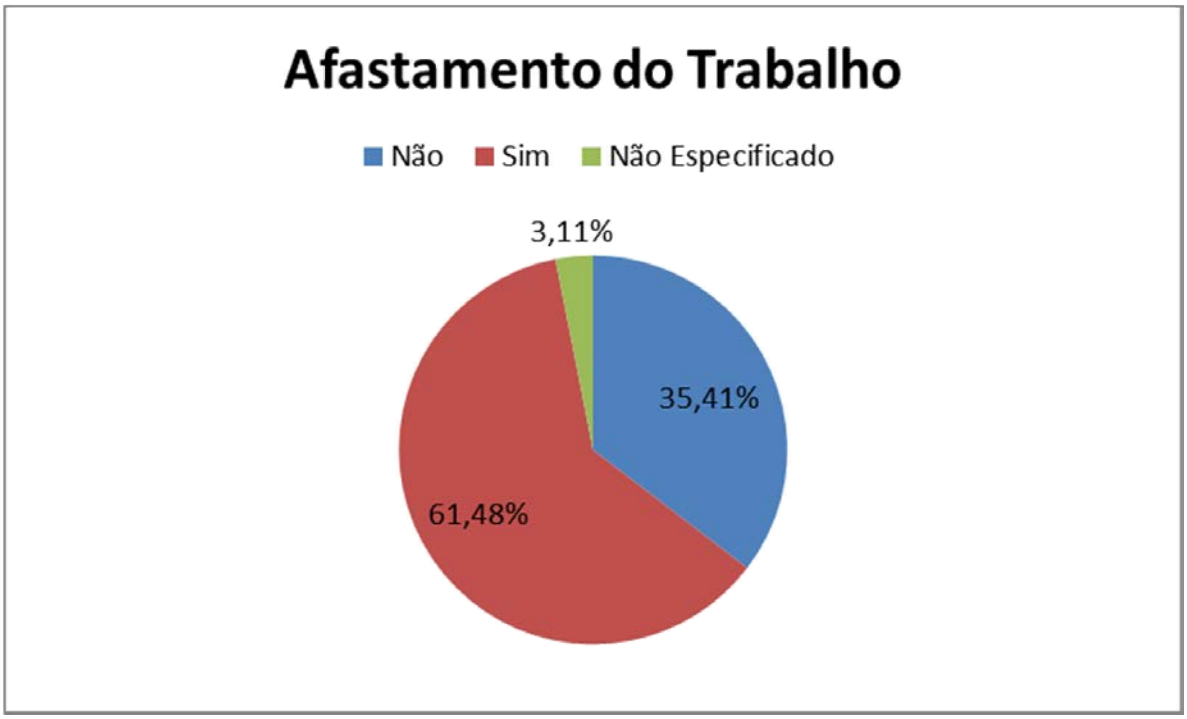

Figura 59: Gráfico referente ao agrupamento 7, considerando o atributo afastamento do trabalho.

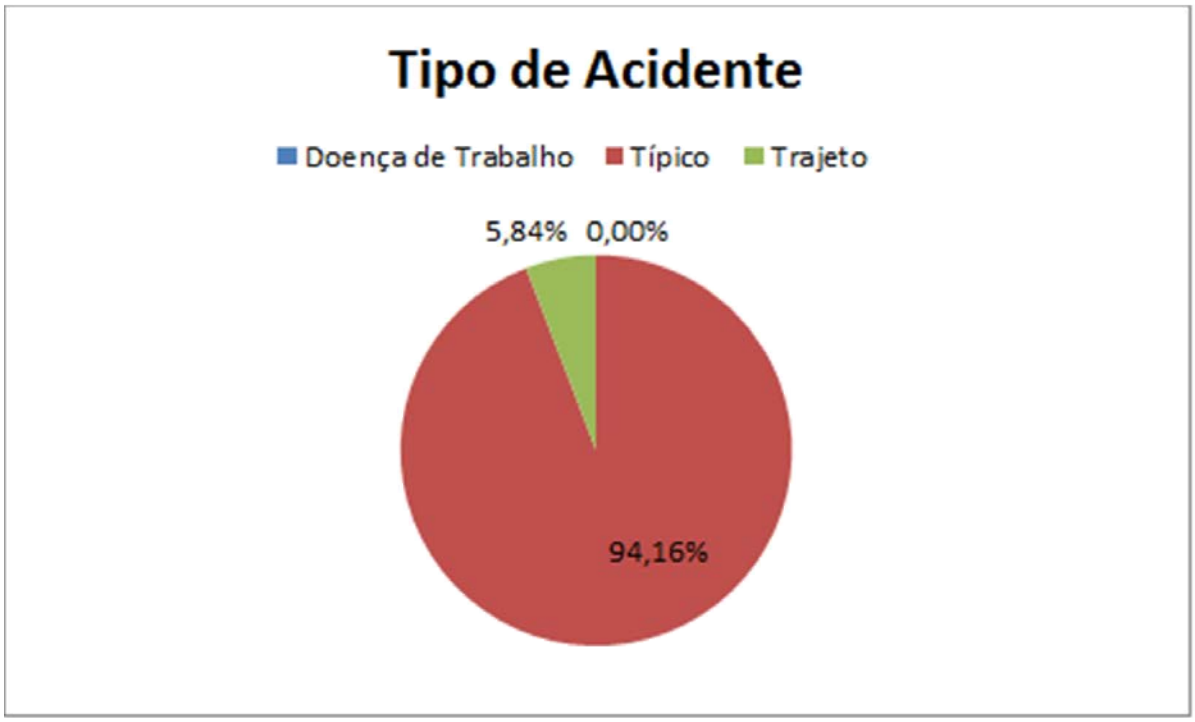

Figura 60: Gráfico referente ao agrupamento 7, considerando o atributo tipo de acidente.

No agrupamento 7, verifica-se que 61,48% dos acidentados sofreram ausência do trabalho e 94,16% dos acidentes foram típicos. O algoritmo CHSMST^{+} obteve como uma boa medida de eficiência em relação à distância que foi de 0,82 e em relação à similaridade foi de 0,75.

O agrupamento 8 contempla 323 ocorrências de acidentes de trabalho, que corresponde a 7,16% de todo o conjunto de dados analisado. Nas Figuras 61 e 62 são apresentados os gráficos referentes aos atributos não espaciais utilizados como critério pelo algoritmo CHSMST^{+}.

No agrupamento 8 verifica-se que 67,18% dos acidentados não sofreram ausência do trabalho e 98,14% dos acidentes foram típicos. O algoritmo CHSMST^{+} obteve como uma boa medida de eficiência em relação à distância que foi de 1,00 e em relação à similaridade que foi de 0,92.

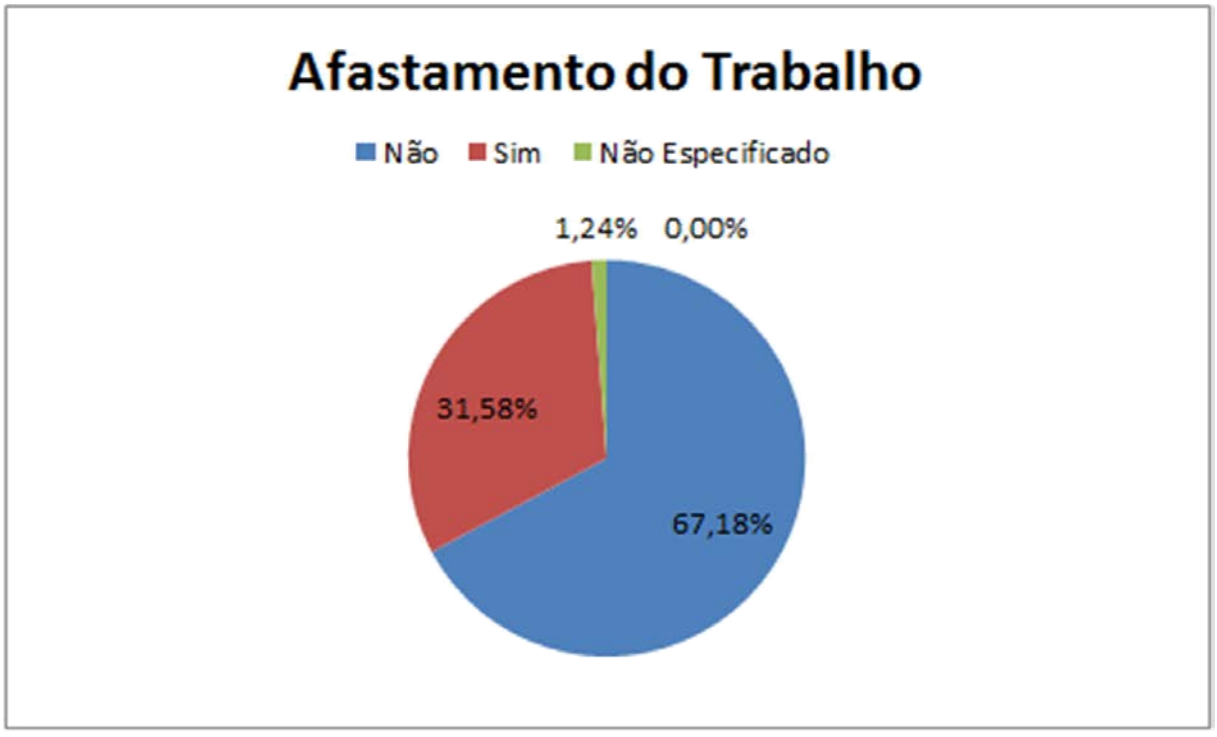

Figura 61: Gráfico referente ao agrupamento 8,considerando o atributo afastamento do trabalho.

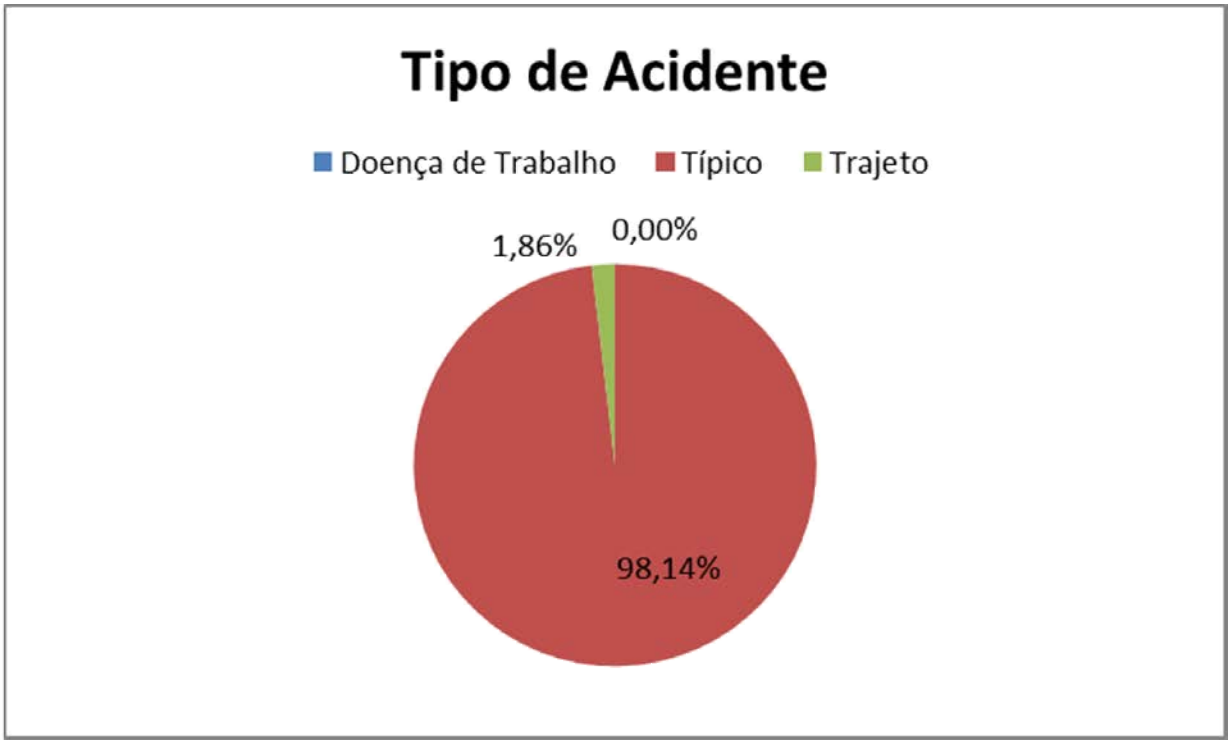

Figura 62: Gráfico referente ao agrupamento 8, considerando o atributo tipo de acidente.

5.5 Análise do Desempenho do Algoritmo CHSMST^{+}

O processamento do algoritmo foi otimizado por meio da incorporação da técnica de *multithreading* no algoritmo CHSMST^{+}. Com o propósito de demostrar a eficiência da estratégia adotada foram realizados experimentos para comparação do tempo de execução do algoritmo com e sem a otimização por *multithreading*, sendo que para realizar os testes em um volume de dados significativo, foi utilizada a base de dados espaciais disponibilizada pelo *Academy of Natural Sciences*, que contempla registros de espécies de animais georreferenciados (Academy of Natural Sciences, 2014). Os testes foram realizados em um sistema computacional com a seguinte configuração: um processador Intel *Core* i3-2350m (2.3 GHz) e 4 *threads*, memória RAM de 4GB e sistema operacional *Windows* 7 *Home Edition*.

Para aferição dos tempos de execução, foram realizados três conjuntos de testes e, posteriormente, foi calculada a média desses tempos em segundos. Cada conjunto de testes considerou a aplicação do algoritmo utilizando 1, 2 e 3 atributos não espaciais além do atributo espacial, uma vez que o objetivo deste algoritmo é considerar a similaridade dos objetos além da proximidade espacial. Também foi verificado o comportamento do algoritmo para diferentes volumes de dados, a saber: 25.000, 50.000 e 100.000 registros.

Na primeira etapa dos testes, o agrupamento de dados espaciais foi realizado considerando o atributo família e o atributo espacial referente à localização do animal. Na Tabela 13 é apresentado o tempo médio de execução para cada caso e na Figura 63 o respectivo gráfico.

Tabela 13: Tempo médio de execução utilizando um atributo não espacial.

Experimento	Número de Registros	Tempo sem *Multithreading* (em segundos)	Tempo com *Multithreading* (em segundos)
1	25.000	595,00	389,67
2	50.000	1.856,33	1.113,33
3	100.000	7.820,33	5.373,00

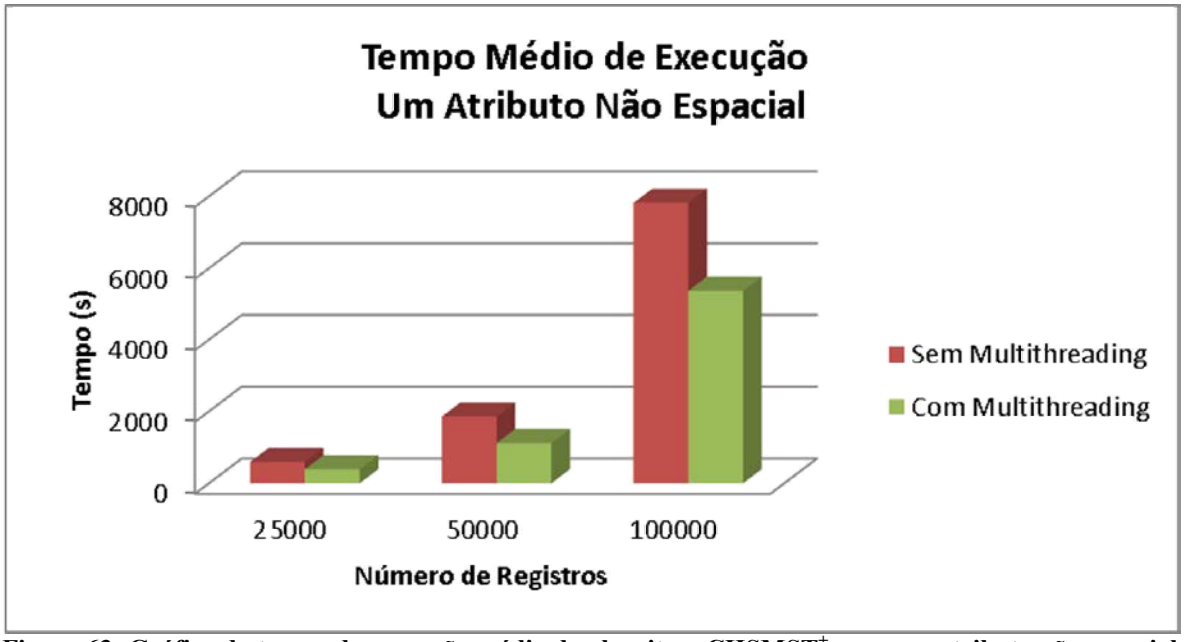

Figura 63: Gráfico do tempo de execução médio do algoritmo CHSMST$^+$ para um atributo não espacial.

No gráfico apresentado na Figura 63 verifica-se que a versão do algoritmo CHSMST$^+$ com *multithreading* apresentou um tempo inferior para todas as quantidades de registros, sendo que a redução média foi de 35,28%.

Para a segunda etapa foram selecionados os atributos família e gênero como critérios de agrupamento além do atributo espacial referente à localização espacial do animal. Na Tabela 14 é apresentado o tempo de execução para cada caso e na Figura 64 o respectivo gráfico, no qual verifica-se que a versão do algoritmo CHSMST$^+$ com *multithreading* apresentou um tempo inferior para todas as quantidades de registros analisadas, sendo a redução média foi de 47,23%.

Tabela 14: Tempo médio de execução utilizando dois atributos não espaciais.

Experimento	Número de Registros	Tempo sem *Multithreading* (em segundos)	Tempo com *Multithreading* (em segundos)
1	25.000	962,00	307,67
2	50.000	2.524,67	1.357,67
3	100.000	8.278,67	6.007,00

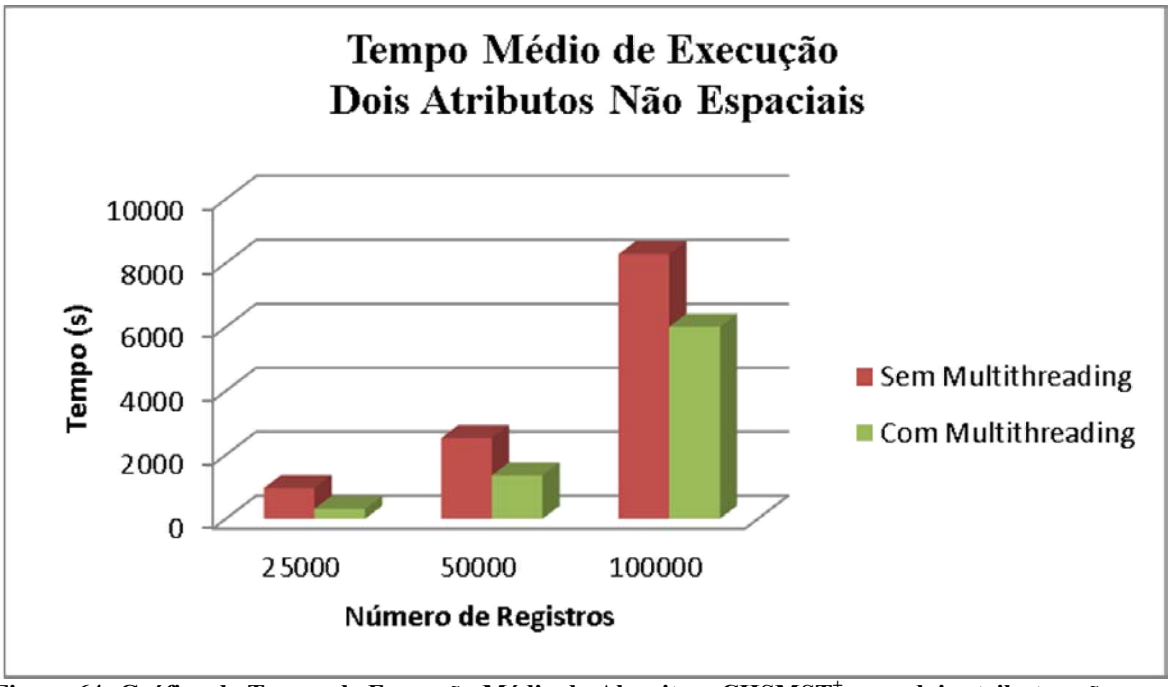

Figura 64: Gráfico do Tempo de Execução Médio do Algoritmo CHSMST^{+} para dois atributos não espaciais.

Para a terceira etapa foram selecionados os atributos família, gênero e região como critérios de agrupamento além do atributo espacial. Na Tabela 15 é apresentado o tempo de execução para cada caso e na Figura 65 o respectivo gráfico.

Tabela 15: Tempo médio de execução utilizando três atributos não espaciais.

Experimento	Número de Registros	Tempo sem *Multithreading* (em segundos)	Tempo com *Multithreading* (em segundos)
1	25.000	706,67	421,00
2	50.000	3.129,00	1.886,67
3	100.000	9.812,33	7.945,33

Figura 65: Gráfico do tempo de execução médio do algoritmo CHSMST$^+$ para três atributos não espaciais.

No gráfico apresentado na Figura 65 verifica-se que a versão do algoritmo CHSMST$^+$ com *multithreading* apresentou um tempo inferior para todas as quantidades de registros, sendo que a redução média foi de 33,05%.

Conforme pode ser observado nos experimentos realizados, a estratégia de otimização do algoritmo CHSMST$^+$ por meio da técnica de *multithreading* se mostrou satisfatória ao passo que houve uma redução média no tempo de execução de 38,52%. A redução foi significativa para diferentes quantidades registros e número de atributos não espaciais, utilizados para a formação do agrupamento. Para comprovar isto, foi realizada a média dos tempos de execução ao aumentar o número de registros e o resultado é exibido na Figura 66. Também foi calculada a média do tempo de execução ao se aumentar o número de atributos e os resultados são apresentados na Figura 67.

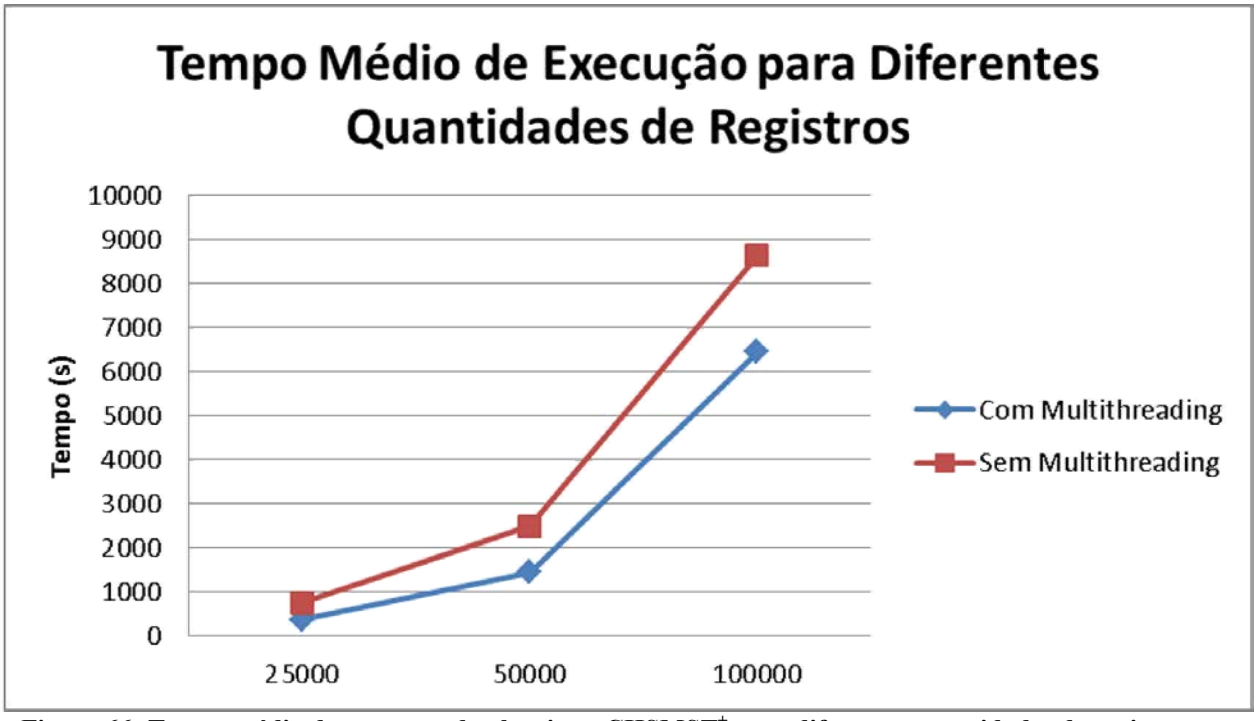

Figura 66: Tempo médio de execução do algoritmo CHSMST^{+} para diferentes quantidades de registros.

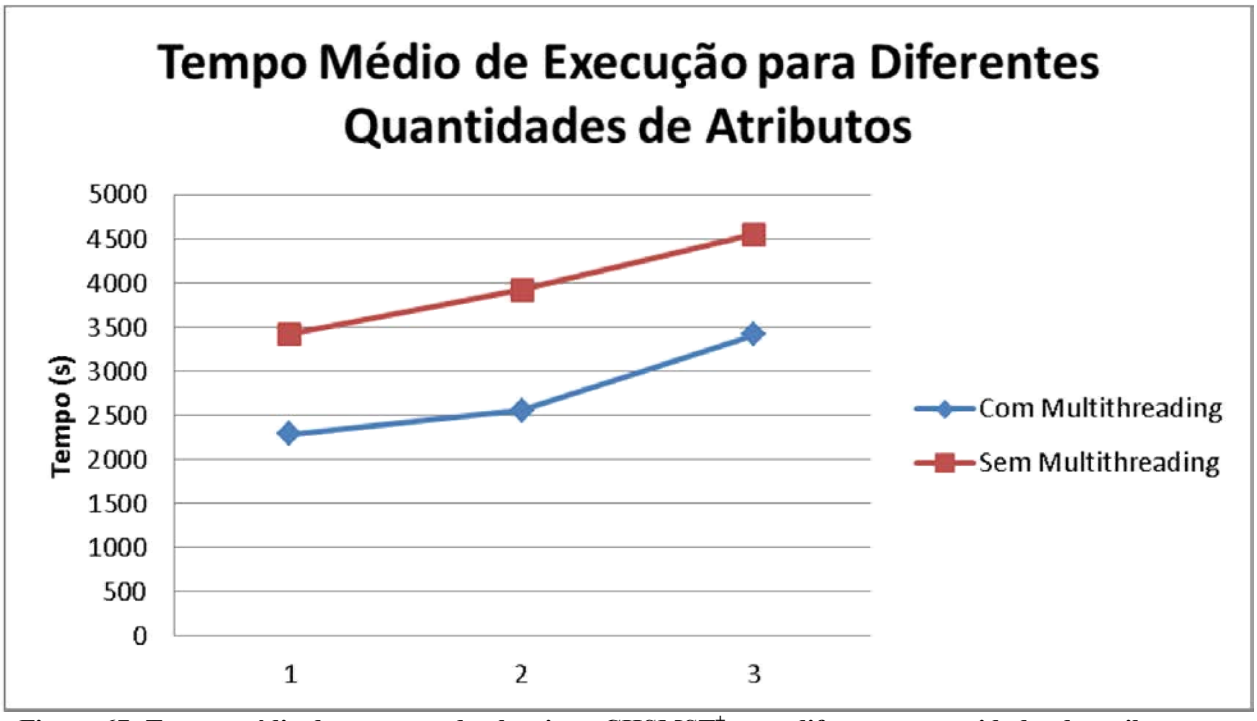

Figura 67: Tempo médio de execução do algoritmo CHSMST^{+} para diferentes quantidades de atributos.

5.6 Considerações Finais

Nesta seção foram apresentados os experimentos realizados com o algoritmo CHSMST^{+}, em que foi possível verificar a aplicabilidade do algoritmo em contextos reais da área da saúde e meio ambiente, a importância da estratégia de similaridade, comparação o

algoritmo desenvolvido com o VDBSCAN, bem como foram realizados testes para verificar a melhora de desempenho ao incluir a estratégia de *multithreading*.

CAPÍTULO 6
Conclusões

6.1 Considerações

O agrupamento de dados espaciais vem se mostrando um método importante da tarefa de prospecção de dados espaciais, visto que pode ser empregado em diversos domínios de aplicações. No entanto, para que os algoritmos de fato tragam resultados satisfatórios muitos desafios ainda necessitam ser superados.

Nesse cenário, foi encontrada a motivação para a confecção do algoritmo CHSMST^{+}, sendo que as principais contribuições deste trabalho são: geração de agrupamentos considerando tanto a distância quanto a similaridade; sem parâmetros de entrada; sem interação com usuário e utilização de técnica de *multithreading* para reduzir o tempo de processamento.

O algoritmo desenvolvido foi aplicado em bases de dados reais, o que possibilitou verificar a sua capacidade em correlacionar um atributo espacial com diferentes quantidades de atributos não espaciais, além de demonstrar a aplicabilidade do algoritmo CHSMST^{+} e da ferramenta FADE tanto para área da saúde quanto para a área ambiental, ressaltando ainda mais a importância do trabalho desenvolvido.

Nos experimentos também foi analisada a eficácia do algoritmo desenvolvido com e sem a estratégia de similaridade, o que possibilitou verificar que a referida estratégia contribui para extração de conhecimento útil e facilita a correlação dos atributos espaciais e não espaciais, uma vez que a análise manual da similaridade, por meio de consultas no banco de dados, é um processo penoso e na maioria dos casos não obtém o mesmo nível de similaridade obtido com aplicação da referida estratégia no algoritmo CHSMST^{+}.

O algoritmo desenvolvido foi comparado com o VDBSCAN, uma vez que este foi identificado como um dos melhores por meio do estudo comparativo entre os algoritmos de agrupamento de dados espaciais. Para isso, foi utilizado o mesmo conjunto de dados e diferentes quantidades de atributos não espaciais. Como resultado obteve-se que nos experimentos realizados a qualidade dos agrupamentos gerados pelo CHSMST^{+} foi superior que os agrupamentos obtidos com VDBSCAN.

Ao final foram realizados testes para verificar o desempenho do algoritmo CHSMST$^+$ com e sem a estratégia de *multithreading*. Para isso, o algoritmo foi aplicado em uma base de dados com registros de espécies de animais georreferenciados. Foram realizados três conjuntos de testes com 25.000, 50.000 e 100.000 registros, considerando a correlação de um, dois e três atributos não espaciais. Em todos os casos foi possível executar o algoritmo mais rapidamente ao utilizar a estratégia de *multithreading*, sendo obtida uma redução média de 38,52% no tempo de execução do CHSMST$^+$.

Diante do exposto verifica-se que o algoritmo CHSMST$^+$ cumpre o objetivo de contribuir no contexto computacional ao aprimorar os resultados e desempenho do algoritmo original e que, além disso, o algoritmo tem potencial para contribuir com os avanços das pesquisas nas áreas da saúde, ambiental e em outras áreas, uma vez que o mesmo pode ser aplicado em diferentes contextos para a descoberta de conhecimento em bases de dados espaciais.

6.2 Comparativo do CHSMST$^+$ com Algoritmos Estudados

As estratégias utilizadas pelo algoritmo CHSMST$^+$ possibilitaram contemplar características não consideradas no algoritmo original, conforme é apresentado na Tabela 16.

Tabela 16: Comparativo do CHSMST$^+$ e demais algoritmos estudados (PATTABIRAMAN et al., 2009, DENG et al., 2011, XIAO-PING; ZHENG-YUAN; JAIN-HUA, 2010, XI, 2009, ANDREOPOULOS et al., 2009, PETER; ANTONYSAMY, 2010, LIU et al., 2012, SONG; JIN; SHEN, 2011, NOSOVSKIY; LIU; SOURINA, 2008, PRASAD; SARMAH, 2011).

Algoritmo	CHSMST$^+$	CHSMST	ADACLUS	VDBSCAN	WAVECLUSTER	STING	CLIQUE	DENCLUE	OPTICS	MST	DBSCAN	BIRCH	CURE	SINGLE LINK	CLARANS
Formatos Variados	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗
Densidade Variada	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Robusto a Ruído	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓	✓	✓	✗	✓
Considera a Similaridade	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Sem Parâmetros de Entrada	✓	✗	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗
Insensível à Ordem dos Dados	✓	✓	✓	✗	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✗
Suporta a Alta Dimensionalidade	✓	✓	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗
Multithreading	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

6.3 Trabalhos Futuros

Com o objetivo de prosseguir nos avanços apresentados pelo algoritmo CHSMST$^+$ pretende-se incorporar a estratégia *MapReduce* para melhorar ainda mais o desempenho do algoritmo. Com isso, em um trabalho futuro o algoritmo poderia ser disponibilizado em um ambiente de Nuvem, uma vez que alguns *frameworks* como *Hadoop* são oferecidos segundo um modelo PaaS – *Plataform-as-a-Service*.

Os algoritmos de agrupamentos podem gerar resultados equivocados devido à presença de obstáculos, tais como rios e montanhas, e de facilitadores, como pontes e rodovias, na região em que os objetos estão localizados. Neste cenário, um trabalho futuro poderia contemplar o tratamento dos obstáculos e facilitadores, para que o algoritmo reconheça a presença dos mesmos e não gere agrupamentos que os contenham, o que seria uma contribuição importante já que poucos algoritmos encontrados na literatura realizam esta tarefa. No entanto, a dificuldade deste desafio está em desenvolver recursos de processamento de imagens para identificar automaticamente os obstáculos e facilitadores em um mapa e armazenar os respectivos dados espaciais. A partir disso, seriam necessárias poucas modificações no algoritmo CHSMST$^+$ para que o mesmo realize o tratamento dos obstáculos e facilitadores.

Para ampliar ainda mais a capacidade de análise do algoritmo desenvolvido, sugere-se a incorporação de características espaços-temporais na geração dos agrupamentos, sendo que na literatura, diversos são os trabalhos que tem direcionado esforços com este propósito.

Por fim, verifica-se como uma opção para os avanços na área de prospecção de dados espaciais, investir esforços para ampliar a capacidade de análise por meio da ferramenta FADE, por meio do desenvolvimento de novos recursos gráficos e incorporação de novos algoritmos de prospecção de dados espaciais.

6.4 Produção Científica durante o Período de Mestrado

Durante o período de mestrado foi possível produzir e publicar vários trabalhos na área de prospecção de dados espaciais, conforme apresentado a seguir:

VALÊNCIO, C. R. et al. **CHSMST$^+$: an algorithm for spatial clustering**. In: INTERNATIONAL CONFERENCE ON DATABASE AND EXPERT SYSTEMS APPLICATIONS (DEXA 2014), 25., 2014. Trabalho submetido aguardando aprovação.

VALÊNCIO, C. R. et al. **Knowledge extraction in spatial databases: CHSMST^{+}** Algorithm. In: GEOINFORMATICA, 2014. Trabalho submetido aguardando aprovação.

VALÊNCIO, C. R. et al. Spatial clustering applied to health area. In: INTERNATIONAL CONFERENCEON PARALLEL AND DISTRIBUTED COMPUTING APPLICATIONS AND TECHNOLOGIES (PDCAT), 12., 2011. **Proceedings**... Gwangju, 2011. p. 427-432.

MEDEIROS, C. A. et al. Ferramenta de apoio ao spatial data mining. In: CONFERÊNCIA IADIS IBERO-AMERICANA WWW/INTERNET, 2011. **Proceedings**... Rio de Janeiro, 2011, p. 396-398.

VALÊNCIO, C. R. et al. VDBSCAN+: Performance Optimization Based on GPU Parallelism. In: INTERNATIONAL CONFERENCE ON PARALLEL AND DISTRIBUTED COMPUTING, APPLICATIONS AND TECHNOLOGIES, 14., 2013. **Proceedings**... Taipei, 2013.

VALÊNCIO, C. R. et al. 3D Geovisualisation techniques applied in spatial data mining. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND DATA MINING - MLDM, 2013,

New York. Machine Learning and Data Mining in Pattern Recognition, 9., 2013. **Proceedings**... Berlin: Springer-Verlag, 2013. v. 7988. p. 57-68.

VALÊNCIO, C. R. et al. Web geographic information system to support environmental resource management. In: INTERNATIONAL CONFERENCE IN ECOLOGICAL INFORMATICS, 8., 2012. **Proceedings**... Brasília, DF, 2012.

VALÊNCIO, C. R. et al. **Relatório de análise de dados espaciais de acidentes de trabalho do ano de 2011**. São José do Rio Preto, 2013.

VALÊNCIO, C. R. et al. **Relatório de análise de dados espaciais de acidentes de trabalho do ano de 2012**. São José do Rio Preto, 2013.

VALÊNCIO, C. R. et al. **Relatório de análise de dados espaciais de acidentes de trabalho do ano de 2011**. São José do Rio Preto, 2013.

Referências Bibliográficas

ABDUL-RAHMAN, A.; PILOUK, M. **Spatial data modelling for 3D GIS**. New York: Springer Berlin Heidelberg, 2007.

Academy of Natural Sciences: MAL. Disponível em: http://www.gbif.org/dataset/86b50d88-f762-11e1-a439-00145eb45e9a. Acesso em: 02 jan. 2014.

AGRAWAL, R. et al. Automatic subspace clustering of high dimensional data for data mining applications. **SIGMOD Record**, New York, v. 27, n. 2, p. 94-105, 1998.

ANDREOPOULOS, B. et al. A roadmap of clustering algorithms: finding a match for a biomedical application. **Briefings in Bioinformatics**, London, v. 10, n. 3, p. 297-314, 2009.

ANKERST, M. et al. OPTICS. **SIGMOD Record**, New York, v. 28, n. 2, p. 49-60, 1999. Disponível em: <http://portal.acm.org/citation.cfm?doid=304181.304187>. Acesso em: 3 out. 2013.

ANWER, E. **Handling Multiple Processors in Your Code Using RapidMind**, 2007. Disponível em: <http://www.codeguru.com/cpp/misc/misc/plug-insadd-ins/article.php/c14005/Handling-Multiple-Processors-in-Your-Code-Using-RapidMind.htm>. Acesso em: 7 Jan. 2014.

BOYD, D. S.; FOODY, G. M. An overview of recent remote sensing and GIS based research in ecological informatics. **Ecological Informatics**, Amsterdam, v. 6, n. 1, p. 25-36, 2011.

CÂMARA, G. **Modelos, linguagens e arquiteturas para bancos de dados geográficos.** 1995. Tese (Doutorado em Computação Aplicada)-Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 1995. Disponível em: <http://://www.dpi.inpe.br/teses/gilberto/>. Acesso em: 5 abr. 2013.

CARREIRA, J. A. **Ferramenta de apoio à limpeza de dados no processo de extração de conhecimento de bases de dados**. 2008. Monografia (Trabalho de Conclusão de Curso)-Universidade Estadual Paulista, São José do Rio Preto, 2008.

CEREST. **Centro de Referência em Saúde do Trabalhador**. Disponível em: <http://portal.saude.gov.br/portal/saude/visualizar_texto.cfm?idtxt=30427>. Acesso em: 5 jan. 2014.

CHEN, X. J.; LI, X.Y.; ZHAN, Y. S. Design of WebGIS application based on Ajax. **Advanced Materials Research**, Zürich, v. 542/543, p. 1282-1285, 2012.

CHOWDHURY, A. K. M. R. An efficient method for subjectively choosing parameter 'k' automatically in VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise) algorithm. In: INTERNATIONAL CONFERENCE ON COMPUTER AND AUTOMATION ENGINEERING, 2., 2010. **Proceedings**… Singapore, 2010. v. 1, p. 38-41.

CLEVE, C. et al. Classification of the wildland-urban interface: a comparison of pixel and object-based classifications using high-resolution aerial photography. **Computers, Environment and Urban Systems**, New York, v. 32, n. 4, p. 317-326, 2008.

DEAN, J.; GHEMAWAT, S. MapReduce: simplified data processing on large clusters. **Communications of the ACM**, New York, v. 51, n. 1, p. 107-113, 2008.

Deliberação CBH-TG, 2008. Deliberação CBH-TG Nº 141 de 25/03/2008. Disponível em < http://www.sigrh.sp.gov.br/sigrh/ARQS/DELIBERACAO/CRH/CBH-TG/2558/delcbhtg141-08_%20prioridade%20de%20investimentos%20fehidro-2008.pdf >. Acesso em: 04 dez. 2013.

_____, 2010. Deliberação CBH-TG Nº 172/2010 de 30/04/2010. Disponível em <http://www.sigrh.sp.gov.br/sigrh/ARQS/DELIBERACAO/CRH/CBH-TG/3454/delcbhtg172-10_indica%20prioridades%20de%20investimento%20fehidro_2010.pdf>. Acesso em: 10 dez. 2013.

DENG, M. et al. An adaptive spatial clustering algorithm based on Delaunay triangulation. **Computers, Environment and Urban Systems**, New York, v. 35, n. 4, p. 320-332, 2011.

DUQUE, J. C. et al. A computationally efficient method for delineating irregularly shaped spatial clusters. **Journal of Geographical Systems**, New York, v. 13, n. 4, p. 355-372, 2011.

EICHINGER, F.; PANKRATIUS, V.; BÖHM, K. Data mining for defects in multicore applications: an entropy-based call-graph technique. **Concurrency and Computation**: Practice and Experience, Chinchester, v. 26, n. 1, p. 1-20, 2014.

ELMAGARMID, A. K.; IPEIROTIS, P. G.; VERYKIOS, V. S. Duplicate record detection: a survey. **IEEE Transactions on Knowledge and Data Engineering**, New York, v. 19, n. 1, p. 1-16, 2007.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados**. 6. ed. São Paulo: Pearson Addison Wesley, 2011.

ERL, T. **Service-oriented architecture**: concepts, technology, and design. Upper Saddle River: Prentice Hall PTR, 2005.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2., 1996. **Proceedings…** Portland,1996, p. 226-231.

_____. Algorithms for characterization and trend detection in spatial databases. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 4., 1998. **Proceedings**… New York: AAAI - American Association for Artificial Intelligence, 1998. p.1-7.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J. S. Algorithms and applications for spatial data mining. In: MILLER, H. J.; HAN, J. (Ed.). **Geographic data mining and knowledge discovery.** New York: Taylor &Francis, 2001. (Research Monographs in GIS)

_____. Spatial data mining: a database approach. In: SCHOLL, M.; VOISARD, A. (Ed.). **Advances in spatial databases**. Portland: Springer-Verlag, 1997. p.47-66.

ESTIVILL-CASTRO, V.; LEE, I. AMOEBA: Hierarchical clustering based on spatial proximity using Delaunay diagram. In: INTERNATIONAL SYMPOSION ON SPATIAL DATA HANDLING, 9., 2000. **Proceedings…** Beijing, 2000a, p.26-41.

_____. Argument free clustering for large spatial point-data sets via boundary extraction from Delaunay diagram. **Computers, Environment and Urban Systems**, New York, v. 26, n. 4, p. 315-334, 2002.

_____. AUTOCLUST: Automatic clustering via boundary extraction for massive point-data Sets. In: INTERNATIONAL CONFERENCE ON GEOCOMPUTATION, 5., 2000. **Proceedings**… Wuhan, 2000b. p. 1-18.

FAN, B. A hybrid spatial data clustering method for site selection: the data driven approach of GIS mining. **Expert Systems with Applications**, New York, v. 36, n. 2, p. 3923-3936, 2009.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, New York, v. 39, n. 11, p. 27-34, 1996.

GENG, X.; YANG, Z. Data mining in cloud computing. In: INTERNATIONAL

CONFERENCE ON INFORMATION SCIENCE AND COMPUTER APPLICATIONS, 2013. **Proceedings…** Hong Kong: Atlantis Press, 2013. p. 1–7.

GEOSERVER. **Servidor de mapas para web**. Disponível em: <http://geoserver.org/display/GEOS/Welcome>. Acesso em: 8 out. 2013.

GOODMAN, L.; KRUSKAL, W. Measures of associations for cross-validations. **Journal of the American Statistical Association**, New York, v. 49, p. 732-764, 1954.

GOOGLE. **Visualizar dados geoespaciais**. Disponível em: <https://developers.google.com/maps/visualize?hl=pt-br>. Acesso em: 13 dez. 2013.

GUHA, S.; RASTOGI, R.; SHIM, K. ROCK: a robust clustering algorithm for categorical attributes. In: INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 15., 1999. **Proceedings**… Sydney: IEEE, 1999, p. 512–521.

HALEVY, A.; RAJARAMAN, A.; ORDILLE, J. Data integration: the teenage years. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 32., 2006. **Proceedings**… Seoul, 2006. p. 9-16.

HAN, J.; KAMBER, M. **Data mining concepts and techniques**. 3. ed. San Francisco: Elsevier, 2011.

HASAN, R. et al. Extended algorithm for spatial characterization and discrimination rules. In: INTERNATIONAL CONFERENCE ON COMPUTER AND INFORMATION TECHNOLOGY, 11., 2008. **Proceedings**… Sydney, 2008. p. 548-553.

HE, Q.; SHI, Z.-Z.; REN, L.-A. The classification method based on hyper surface. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, 2002. **Proceedings**… Honolulu: IEEE, 2002. p. 1499-1503.

HE, Q.; ZHAO, W.; SHI, Z. CHSMST: a clustering algorithm based on hyper surface and minimum spanning tree. **Soft Computing**, Berlin, v. 15, n. 6, p. 1097-1103, 2011.

HE, Y. et al. MR-DBSCAN: An efficient parallel density-based clustering algorithm using mapreduce. In: INTERNATIONAL CONFERENCE ON PARALLEL AND DISTRIBUTED SYSTEMS, 17., 2011. **Proceedings**… Tainan: IEEE, 2011. p. 473-480.

HINNEBURG, A.; HINNEBURG, E.; KEIM, D. A. An efficient approach to clustering in large multimedia databases with noise. In: INTERNATIONAL CONFERENCE KNOWLEDGE DISCOUVERY AND DATA MINING, 1998. **Proceedings**… New York: AAAI Press, 1998. p. 58-65.

JAIN, A. K. **Algorithms for clustering data**. New Jersey: Prentice Hall, 1988.

JARKE, M. et al. **Fundamentals of data warehouses**. 2. ed. Secaucus: Springer-Verlag, 2003.

JAVED, M.Y.; NAWAZ, A. Data Load Distribution by Semi Real Time Data Warehouse. In: INTERNATIONAL CONFERENCE ON COMPUTER AND NETWORK TECHNOLOGY, 2., 2010. **Proceedings**… Bangkok: IEEE, 2010, p. 556-560.

JIA, Z.; LIU, Y. Visualized geology spatial data classifying based on integrated techniques between GIS and SDM. In: INTERNATIONAL WORKSHOP ON DATABASE TECHNOLOGY AND APPLICATIONS, 1., 2009. **Proceedings**… Wuhan, 2009. p. 165-169.

JIEHAI, C.; WEI, L. Research on the storage and management of mine spatial data based on PostgreSQL. In: INTERNATIONAL CONFERENCE ON COMPUTER APPLICATION AND SYSTEM MODELING, 2010. **Proceedings**… Taiyuan, 2010. p. V9-493-496, 2010.

JIN, H.; MIAO, B. The research progress of spatial data mining technique. In: INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND INFORMATION TECHNOLOGY, 3., 2010. **Proceedings**… Chengdu, 2010. p. 81-84.

JING-SHENG, X. Parallel CLARANS: improvement and application of CLARANS algorithm. In: INTERNATIONAL CONFERENCE ON COMPUTER AND

COMMUNICATION TECHNOLOGIES IN AGRICULTURE ENGINEERING, 2010. **Proceedings**… Beijing, 2010. v. 3, p. 248-251.

KAFURA, D. **MapReduce Concurrency for data-intensive applications**. Disponível em: <http://courses.cs.vt.edu/cs5204/fall10-kafura-BB/Presentations/MapReduce.pdf>. Acesso em 03 jan. 2014.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data**: an introduction to cluster analysis. New York: John Wiley & Sons, 1990.

KOPERSKI, K.; HAN, J. Discovery of spatial association rules in geographic information databases. In: SCHOLL, M.; VOISARD, A. (Ed.). **Advances in spatial databases**. Portland: Springer-Verlag, 1997. p.47-66.

KYUEUN, Y.; GAUDIOT, J.-L. Network applications on simultaneous *multithreading* processors. **IEEE Transactions on Computers**, New York, v. 59, n. 9, p. 1200-1209, 2010.

LEI, G. et al. An incremental clustering algorithm based on grid. In: INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY*,* 8., 2011. **Proceedings**… Shangai, 2011. p. 1100-1103.

LIU, P.; ZHOU, D.; WU, N. VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise. In: INTERNATIONAL CONFERENCE ON SERVICE SYSTEMS AND SERVICE MANAGEMENT, 2007. **Proceedings**… Chengdu, 2007. p. 1-4.

LIU, Q. et al. A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. **Computers & Geosciences**, New York, v. 46, p. 296-309, 2012.

LIU, X.; LU, C.-T.; CHEN, F. Spatial outlier detection. In: SIGSPATIAL INTERNATIONAL CONFERENCE ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS - GIS'10, 2010. **Proceedings**…New York, New York, USA: ACM Press, 2010, p. 370-379.

LLOYD, S. Least squares quantization in PCM. **IEEE Transactions on Information Theory**, New York, v.28, n.2, p.129-137, 1982. Versão original: Technical Report, Bell Labs, 1957.

MAMOULIS, N. **Spatial data management**. San Francisco: Morgan & Claypool Publishers, 2011.

MENNIS, J.; GUO, D. Spatial data mining and geographic knowledge discovery: an introduction. **Computers, Environment and Urban Systems**, New York, v. 33, n. 6, p. 403-408, 2009.

MILLER, H. J.; HAN, J. **Geographic data mining and knowledge discovery**. 2. ed. New York: CRC Press; Taylor & Francis, 2009.

NATARAJAN, R. et al. Effectiveness of compiler-directed prefetching on data

mining benchmarks. **Journal of Circuits, Systems and Computers**, Singapore, v. 21,

n. 2, p. 1240006, 2012.

NEMIROVSKY, M.; TULLSEN, D. M. **Multithreading architecture**. [s. l.]: Morgan & Claypool Publishers, 2013. (Synthesis Lectures on Computer Architecture, 21)

NEVES, M. C.; FREITAS, C. C.; CÂMARA, G. **Prospecção de dados em grandes bancos de dados geográficos**. São José dos Campos: INPE, 2001. Relatório Técnico. Disponível em: <http://www.dpi.inpe.br/geopro/modelagem/relatorio_data_mining.pdf>. Acesso em: 5 abr. 2013.

NG, R. T.; HAN, J. CLARANS: a method for clustering objects for spatial data mining. **IEEE Transactions on Knowledge and Data Engineering**, New York, v. 14, n. 5, p. 1003-1016, 2002.

_____. Efficient and effective clustering methods for spatial data mining. In: VERY LARGE DATA BASE CONFERENCE, 20., 1994. **Proceedings**… Santiago de Chile, 1994. p.144-155.

NOSOVSKIY, G. V.; LIU, D.; SOURINA, O. Automatic clustering and boundary detection algorithm based on adaptive influence function. **Pattern Recognition**, Elmsford, v.41, n.9, p.2757-2776, 2008.

OBRADOVIC, D. D. et al. Modeling and PostGIS implementation of the basic planar imprecise geometrical objects and relations. In: IEEE INTERNATIONAL SYMPOSIUM ON SYSTEMS AND INFORMATICS, 9., 2011. **Proceedings**… Subotica, 2011. p. 157-162.

OPENLAYERS. **OpenLayers**: free maps for the web. Disponível em: <http://openlayers.org>. Acesso em: 20 jun. 2013.

ORACLE. **O que é o Java?** Disponível em: <http://www.java.com/pt_BR/download/whatis_java.jsp>. Acesso em: 13 out. 2013a.

_____. **JavaServer Pages Technology**. Disponível em: <http://www.oracle.com/technetwork/java/javaee/jsp/index.html>. Acesso em: 13 out. 2013b.

_____. **Benefits of Multithreading**. Disponível em: <http://docs.oracle.com/cd/E19455-01/806-3461/6jck06gqj/index.html>. Acesso em 03 jan. 2014.

PATTABIRAMAN, V. et al. A novel spatial clustering with obstacles and facilitators constraint based on edge deduction and k-medoids. In: INTERNATIONAL CONFERENCE ON COMPUTER TECHNOLOGY AND DEVELOPMENT, 2009. **Proceedings**… Kota Kinabalu, 2009. v. 1, p. 402-406.

PEI, T. et al. Detecting feature from spatial point processes using collective nearest neighbor. **Computers, Environment and Urban Systems,** New York, v. 33, n. 6, p. 435-447, Nov. 2009.

PETER, J. H.; ANTONYSAMY, A. Heterogeneous density based spatial clustering of application with noise. **International Journal of Computer Science and Network Security**, Seoul, v.10, n. 8, p. 210-214, 2010.

PIATETSKY-SHAPIRO, G. **KDNUGGETS**: data mining, web mining, text mining, and knowledge discovery. Disponível em: <http://www.kdnuggets.com/>. Acesso em: 16 jun. 2012.

PIVATO, M. A. Prospecção de regras de associação em dados georreferenciados. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional)– Universidade de São Paulo, São Carlos, 2006.

POSTGRESQL. **About**. Disponível em: <http://www.postgresql.org/about>. Acesso em: 8 out. 2013.

PRASAD, R. K.; SARMAH, R. Variable density spatial data clustering. In: INTERNATIONAL CONFERENCE ON COMPUTER AND COMMUNICATION TECHNOLOGY, 2011. **Proceedings**… Allahabad, 2011. p. 53-58.

PRIM, R. C. Shortest connection networks and some generalization. **Bell Systems Technical Journal**, Paris, v. 36, p.1389-1401, 1957.

QIANG, X.; WEI, Y.; HANFEI, Z. Application of visualization technology in spatial data mining. In: INTERNATIONAL CONFERENCE ON COMPUTING, CONTROL AND INDUSTRIAL ENGINEERING, 2010. **Proceedings**… Hong Kong, 2010. p. 153-157.

RAHM, E.; DO, H. H. Data cleaning: problems and current approaches. **IEEE Data Engineering Bulletin**, New York, v. 23, p. 3-13, 2000.

RAMAKRISHNAN, R.; GEHRKE, J. **Sistemas de gerenciamento de banco de dados**. 3. ed. São Paulo: McGraw Hill Brasil, 2008.

REFRACTIONS RESEARCH. **What is PostGIS?** Disponível em: <http://postgis.refractions.net>. Acesso em: 8 out.2013.

ROQUEIRO, D.; FRASOR, J.; DAI, Y. BindSDb: a binding-information spatial database. In: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE WORKSHOPS, 2010. **Proceedings**… Philadelphia: IEEE, 2010, p. 573-578.

SANTOS, M. Y.; AMARAL, L. A. Mining geo-referenced data with qualitative spatial reasoning strategies. **Computers & Graphics**, New York, v. 28, n. 3, p. 371-379, 2004.

SENCHA. **Sencha Ext JS JavaScript Framework for Rich Desktop Apps**. Disponível em: < http://www.sencha.com/products/extjs>. Acesso em: 13 out. 2013.

SHEIKHOLESLAMI, G.; CHATTERJEE, S.; ZHANG, A. WaveCluster : a multi-resolution clustering approach for very large spatial databases. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATABASE, 1998. **Proceedings**... New York, 1998, p. 28-43.

SHEKHAR, S.; ZHANG, P.; HUANG, Y. Trends in spatial data mining. In: KARGUPTA, H. (Ed.). et al. **Next generation challenges and future**. Boca Raton: CRC Press, 2003.

SONG, Y.; JIN, S.; SHEN, J. A unique property of single-link distance and its application in data clustering. **Data & Knowledge Engineering**, Amsterdam, v. 70, n. 11, p. 984-1003, 2011.

WANG, W.; YANG, J.; MUNTZ, R. STING : a statistical information grid approach to spatial data mining. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 23., 1997. **Proceedings…** San Francisco: Morgan Kaufmann Publishers, 1997. p. 186-195.

XI, J. Spatial clustering algorithms and quality assessment. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 2009. **Proceedings**… Barcelona, 2009. p. 105-108.

XIAO-PING, L.; ZHENG-YUAN, M.; JAIN-HUA, L. A spatial clustering method by *means* of field model to organize data. In: WRI GLOBAL CONGRESS ON INTELLIGENT SYSTEMS, 2., 2010. **Proceedings**… Wuhan: IEEE Computer Society, 2010. p. 129-131.

YANG, H.-C. et al. Map-reduce-merge. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2007. **Proceedings…** New York: ACM Press, 2007. p. 1029.

YEUNG, A. K. W.; HALL, G. B. **Spatial database systems**: design, implementation and project management. Amsterdam: Springer, 2007.

YUESHUN, H.; WEI, X. A study of spatial data mining architecture and technology. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND INFORMATION TECHNOLOGY, 2009. **Proceedings**… Beijing, 2009. p. 163-166.

ZHANG, H. et al. Land use information release system based on Google Maps API and XML. In: INTERNATIONAL CONFERENCE ON GEOINFORMATICS, 18., 2010. **Proceedings**... Beijing: IEEE, 2010. p. 1-4.

ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. An Efficient Data Clustering Method for Very Large Databases. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 1996. **Proceedings**… New York: ACM Press, 1996. v. 1, p. 103-114. Disponível em: <http://portal.acm.org/citation.cfm?doid=233269.233324>. Acesso em: 3 out. 2013.

ZHAO, Y. et al. Enhancing grid-density based clustering for high dimensional data. **Journal of Systems and Software**, New York, v. 84, n. 9, p. 1524-1539, 2011.

ZHONG, C.; MIAO, D.; FRÄNTI, P. Minimum spanning tree based split-and-merge: a hierarchical clustering method. **Information Sciences**, New York, v. 181, n. 16, p. 3397-3410, 2011.

ZHONG, C.; MIAO, D.; WANG, R. A graph-theoretical clustering method based on two rounds of minimum spanning trees. **Pattern Recognition**, Elmsford, v. 43, n. 3, p. 752-766, 2010.