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

Fernando Tochio Ichiba

Algoritmo para prospecção multirrelacional de dados espaciais

São José do Rio Preto
2013

Fernando Tochio Ichiba

Algoritmo para prospecção multirrelacional de dados espaciais

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", Câmpus de São José do Rio Preto.

Orientador: Prof. Dr. Carlos Roberto Valêncio

São José do Rio Preto
2013

Fernando Tochio Ichiba

Algoritmo para prospecção multirrelacional de dados espaciais

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", Câmpus de São José do Rio Preto.

Banca Examinadora

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

Prof. Dr. José Márcio Machado
UNESP – São José do Rio Preto

Prof. Dr. Pedro Luiz Pizzigatti Corrêa
USP – São Paulo

São José do Rio Preto
22 de fevereiro de 2013

RESUMO

As pesquisas acerca de spatial data mining - ou prospecção de dados espaciais - tem avançado no sentido de melhorar a qualidade dos resultados obtidos pelos algoritmos da área e aprimorar as técnicas utilizadas, na tentativa de apresentar soluções que contornam os principais problemas e desafios: custo computacional elevado e baixa eficiência dos algoritmos. Neste trabalho, é apresentado um algoritmo desenvolvido para prospecção de dados espaciais, que introduz uma abordagem multirrelacional para suportar o agrupamento de dados por similaridade de características espaciais e não espaciais com possibilidade de agregação semântica nessa tarefa. Aplicável a bases de dados volumosas, o algoritmo desenvolvido apresentou resultados com qualidade superior nos experimentos realizados, se comparado com alguns dos mais tradicionais de spatial data mining, sem que houvesse perda semântica no levantamento das informações - muitas vezes ocasionada pelas junções de dados exigidas na aplicação de algoritmos tradicionais - e com um desempenho otimizado por meio do uso de multithreading.

Palavras-chave: *Spatial data mining*. Agrupamento de dados espaciais. Análise mutirrelacional de dados espaciais.

ABSTRACT

Researches involving spatial data mining have advanced in order to improve the quality of results obtained with algorithms and techniques, aiming to present solutions which avoid the main problems and challenges in this research area: high computational cost and low efficiency of the algorithms. In this work, an algorithm for spatial data mining is presented, based on techniques introduced by the VDBSCAN algorithm, which introduces a multi-relational approach to support spatial clustering by similarity of spatial and non-spatial characteristics with the possibility of semantic aggregation in this process. The developed algorithm is able to deal with voluminous databases and it presented better results than some of the most traditional spatial data mining algorithms, avoiding semantic losses in data joining required by traditional algorithms and performing an optimised execution time due to the use of multithreading.

Keywords: *Spatial data mining. Spatial clustering. Multi-relational spatial data mining*.

Índice

Lista de figuras

Lista de tabelas

Lista de siglas

GKD - Geographic Knowledge Discovery

KDD - Knowledge Discovery in Database

MER - Modelo Entidade-Relacionamento

SGBDE - Sistema Gerenciador de Banco de Dados Espaciais

SIG - Sistemas de Informação Geográfica

SIVAT - Sistema de Informação e Vigilância de Acidentes de Trabalho

Capítulo 1 Introdução

1.1 Considerações iniciais

O aprimoramento dos sistemas de gerenciamento de dados, assim como a evolução dos dispositivos de armazenamento, tem contribuído de maneira significativa na captação e na formação de grandes volumes de dados. No entanto, tal cenário é caracterizado por riqueza de dados e pobreza de informações [HAN_06] [SHE_11].

Frente a essa situação, os esforços passaram a se concentrar no desenvolvimento de ferramentas e técnicas para prospecção desses dados, na tentativa de obtenção de informações valiosas e não observáveis explicitamente. Além das próprias, objetiva-se obter também, pela correlação entre elas, um apanhado de novas informações e conhecimentos, até então, desconhecidos.

Um tipo particular de dado é denotado dado espacial, que tem ganhado espaço e atenção na comunidade científica, assim como a busca por padrões, correlações e indicativos por meio de análise de dados espaciais e não espaciais – conhecida como prospecção dados espaciais ou *spatial data mining* – tem se tornado uma pesquisa promissora para os próximos anos [JIN_10].

Diante desse panorama, um levantamento bibliográfico da área foi realizado e fundamentou-se a necessidade de algoritmos e técnicas capazes de elevar a qualidade dos resultados retornados pela aplicação dos mesmos, sem que sejam limitados à aplicação em grandes repositórios de dados com variação de densidade entre os objetos espaciais, nem a restrições de conjuntos de dados.

1.2 Motivação e escopo

Em se tratando de dados espaciais, a prospecção dos mesmos resulta em informações que levam em consideração a localidade. Nesse sentido, surge o conceito de *Geographic Knowledge Discovery* - GKD, ou seja, o *Knowledge Discovery in Database* - KDD a partir de dados espaciais e *spatial data mining*, cujo objetivo é revelar relações e tendências contidas nos dados e, de maneira automatizada, prover informações que contribuam para as tomadas de decisões [WAN_09a].

Diversos são os trabalhos fundamentados na aplicação de técnicas de *spatial data mining* propostos recentemente, tais como para suporte a tomada de decisão referente a navegação [LIG_10], vigilância e prevenção de acidentes do trabalho [VAL_11], gerenciamento ambiental [VAL_12] e outros.

No entanto, a prospecção de dados espaciais apresenta vários limitantes e problemas devido à sua complexidade e dificuldade na construção de algoritmos, aplicação e análise de dados. Alguns dos principais problemas e desafios são [JIN_10]:

1) A maioria dos algoritmos de *spatial data mining* são derivados ou adaptados de algoritmos de prospecção de dados convencionais e não consideram o armazenamento, o processamento e as características específicas dos dados espaciais. Esses são diferentes dos convencionais e, sendo assim, muitas vezes, as técnicas de prospecção de dados convencionais não são apropriadas para se analisar corretamente os fenômenos e correlações entre objetos espaciais;

2) A eficiência dos algoritmos de *spatial data mining* é baixa; os padrões detectados não são refinados. Isso faz com que muitas vezes não sejam encontrados padrões realmente relevantes e úteis ao usuário. Deve-se então haver a interação humana com o algoritmo para se conseguir um refinamento maior e então obter dados mais concretos. É importante também que sejam implementados algoritmos que levem em consideração a semântica para, de forma automatizada, refinar os padrões encontrados;

3) Não há um consenso geral sobre a padronização da linguagem de manipulação de dados (LMD) espaciais. Uma das razões seria o rápido desenvolvimento da tecnologia de banco de dados, que é contínuo. É importante para o desenvolvimento organizado e padronizado da tecnologia de prospecção de dados espaciais desenvolver uma linguagem de consulta dos dados;

4) A interação entre sistemas de descoberta de conhecimento e prospecção de dados espaciais não é forte. É difícil utilizar as informações de domínio de conhecimentos já obtidos no processo de descoberta de conhecimento em dados espaciais;

5) Os métodos e tarefas de prospecção de dados espaciais são únicos, essencialmente destinados a um problema específico, portanto, o conhecimento que pode ser identificado é limitado;

6) A integração da prospecção de dados espaciais com outros sistemas não é suficiente, o que ignora o papel dos Sistemas de Informação Geográfica - SIG à descoberta de conhecimento espacial. O escopo da aplicação de um sistema de prospecção de dados espaciais, cujos métodos e funções são únicos, está sujeito a muitas restrições. Os sistemas de descoberta de conhecimento desenvolvidos atualmente são limitados a um banco de dados. Se a descoberta de conhecimento for aplicada em um campo mais amplo, haverá a necessidade de integração de fontes de dados distintas, o que envolve bancos de dados com esquemas diferentes, bases de conhecimentos distintas, sistemas específicos, sistemas de apoio à decisão, ferramentas de visualização, rede e muitas outras tecnologias.

Os problemas descritos tornam mais difícil a tarefa de se extrair conhecimento de um banco de dados espaciais em relação aos tradicionais bancos de dados relacionais, o que traz desafios à pesquisa na área de prospecção de dados espaciais. Portanto, há uma diversidade de conteúdo a ser estudado para o desenvolvimento futuro da prospecção de dados espaciais e, apresentar contribuições nesse sentido, foi objetivo do trabalho desenvolvido.

1.3 Objetivos

O foco deste trabalho foi o desenvolvimento de um algoritmo com abordagem multirrelacional para prospecção de dados espaciais, cujo objetivo é realizar a atividade de agrupamento de dados - uma das técnicas para extração de conhecimento - a fim de proporcionar informações que considerem a relação entre atributos espaciais e não espaciais por meio de agrupamento dos dados por similaridade, levantar características dificilmente identificadas numa análise sem o suporte computacional e até mesmo com a prospecção de dados convencional, além de também proporcionar uma estrutura que

suporte a seleção de dados multirrelacional, o que dispensa a preparação dos dados numa única relação.

As contribuições originais pretendidas na implementação do trabalho não se limitam a isso; objetivou-se também a busca por resultados com teor de qualidade elevado e otimizado em relação aos mais tradicionais algoritmos da área, além de garantir a conservação da semântica da base de dados na análise, obtida pelo suporte à seleção de dados multirrelacional que, da maneira como foi implementado, também possibilita agregação semântica no cálculo de similaridade entre os objetos georreferenciados.

1.4 Organização do trabalho

A estrutura dessa dissertação está organizada da seguinte maneira:

- No Capítulo 2 é apresentada a fundamentação teórica da área de *spatial data mining* para entendimento das técnicas existentes e embasamento do trabalho desenvolvido;
- No Capítulo 3 é apresentado o algoritmo desenvolvido, com o detalhamento das abordagens criadas e utilizadas;
- No Capítulo 4, os experimentos realizados e os resultados obtidos são descritos e discutidos;
- Finalmente, no Capítulo 5 são apresentadas as conclusões, assim como as possíveis frentes de continuidade ao trabalho proposto.

Capítulo 2 Fundamentação teórica

2.1 Considerações iniciais

Neste capítulo, os conceitos fundamentais referentes à prospecção de dados espaciais e obtenção de conhecimento por meio de algoritmos computacionais e de maneira automatizada são discorridos, assim como alguns trabalhos propostos no estado da arte são apresentados.

Para iniciar a seção, é importante afirmar que um banco de dados espaciais é caracterizado pelo gerenciamento de dados espaciais, isto é, armazenamento, indexação, manipulação e recuperação desses dados. Os tipos de atributos são bastante distintos de um banco de dados convencional (relacional), assim como suas consultas e formas de organização.

Um dado espacial, por sua vez, possui estrutura diferente de um dado convencional, como topologia e informações de distância, geralmente caracterizados por dados complexos multidimensionais, estruturas de indexação espacial, que muitas vezes necessitam de cálculos geométricos e outras operações complexas para sua manipulação [WAN_09a].

Por meio desses conceitos, as tecnologias de captação de informações georreferenciáveis resultaram em grandes repositórios de dados espaciais [SHU_09]. No entanto, devido à complexidade desse tipo de dado, as técnicas convencionais de *data mining* foram substituídas pelas de *spatial data mining* [SHE_03].

2.2 Dados espaciais e spatial data mining

Com relação aos dados espaciais, pode-se afirmar que são considerados complexos por apresentar as seguintes características [PEI_01] [LUO_03] [MIL_01] [CHE_06] [LEE_07]:

1) **Quantidade massiva de dados -** Muitas vezes a aplicação de algoritmos não é possível diante da dificuldade ou da quantidade excessiva de cálculos necessários e, portanto, uma das tarefas da área de prospecção de dados espaciais é criar estratégias computacionais e desenvolver algoritmos novos e eficientes para superar as dificuldades técnicas causadas pela quantidade massiva de dados;
2) **A ambiguidade da informação espacial -** Ambiguidade existe em quase todos os tipos de dados espaciais, como a ambiguidade de localização espacial, de correlação espacial, etc;
3) **Escala -** As características de um dado geográfico podem não ser as mesmas se observadas em diferentes níveis de *zoom*, o que faz com que a escala represente uma complexidade maior dos dados geográficos. A escala pode ser utilizada para explorar mudanças graduais das características do processo de generalização e refinamento de informações;
4) **Escassez de dados espaciais -** É difícil obter alguns atributos de dados espaciais, pois muitos deles dependem de fontes externas restritas e inacessíveis. As características dos dados podem ser modificadas com o tempo e a atualização das informações torna-se uma das dificuldades no tratamento da complexidade dos dados espaciais;
5) **Relação não-linear entre atributos espaciais -** Trata-se de uma característica significativa da complexidade de sistemas espaciais, reflete mecanismos complexos de uma função interna de um sistema e é um dos pontos-chaves da prospecção de dados espaciais;
6) **Aumento da dimensão espacial -** As propriedades dos objetos espaciais sofrem alterações rapidamente, como na área de sensoriamento remoto, devido ao avanço tecnológico de sensores.

Spatial data mining, por sua vez, pode ser definida como extração de conhecimento implícito a partir de relações espaciais ou de outros modos explícitos de um banco de dados espacial. Sua realização necessita da integração da prospecção de dados e da

tecnologia de um banco de dados espacial que gerencia estruturas de dados espaciais, relações espaciais, relações de dados convencionais com dados espaciais, consultas envolvendo espaço, distância e outros [HAN_06]. É também objetivo da prospecção de dados espaciais descobrir relações entre os atributos não espaciais e espaciais [CAM_95] [BAE_09].

Um exemplo clássico de padrão encontrado, considerando a localização geográfica, foi o ocorrido em Londres no século XIX, quando um especialista em epidemias plotou em um mapa a localização das casas das vítimas dos casos de cólera. Com isso, um agrupamento foi formado e, no centro deste, localizava-se uma bomba de água. Por conta do desligamento desta bomba, a cólera foi controlada [CHA_00].

Outro exemplo é o ocorrido em 1909, quando um grupo de dentistas descobriram que a população de Colorado Springs, nos Estados Unidos, tinha melhor saúde bucal e, por conta disso, atribuíram este fato ao nível elevado de flúor na água que abastecia a cidade. Tempos depois, pesquisadores confirmaram o papel positivo do flúor no controle da cárie dentária [CHA_00].

Em cada um desses exemplos, a exploração dos dados espaciais resultou num conjunto de padrões ou hipóteses inesperados confirmados mais tarde pelos especialistas. Esse é o objetivo da prospecção de dados espaciais: automatizar descoberta de tais padrões, considerados indicativos para análise e validação dessa característica.

No entanto, a prospecção de dados espaciais é considerada complexa devido ao elevado grau de complexidade dos dados espaciais. Diferentemente da prospecção de dados convencionais, esse processo lida com tipos de dados que vão além de *strings*, inteiros, datas e outros, mas também com linhas, polígonos e pontos. Além disso, nas relações entre objetos espaciais devem ser consideradas também distância, direção e topologia.

Outra característica a ser considerada na aplicação do *spatial data mining* é a capacidade de determinados atributos dos objetos espaciais influenciarem os valores dos objetos vizinhos. Isso é bastante aplicado quando se trata de análise de tendências como, por exemplo, na economia, meio ambiente e outros [SHE_02] [CHA_00].

2.3 O processo de spatial data mining

O processo de *spatial data mining* pode ser dividido em três camadas [WAN_09a], como mostrado na Figura 1, sendo a primeira camada a corresponde à fonte de dados que devem ser gerenciados por um sistema gerenciador de banco de dados espaciais – SGBDE. Os dados manipulados nessa camada incluem índices, armazenamento e otimização de consultas.

O processo de prospecção de dados fica na segunda camada, em que são realizadas análises dos dados abstraídos, por meio de métodos, algoritmos e técnicas de prospecção de dados espaciais.

A terceira camada é a *interface* do usuário, na qual as informações obtidas e o conhecimento são refletidos para que o usuário analise, valide e obtenha correlações entre os dados.

Em todo o processo de prospecção de dados espaciais, os usuários podem controlar cada passo. De modo geral, uma série de etapas desse processo e da descoberta de conhecimento estão interligadas e, de modo iterativo, a interação humano-máquina é necessária para se alcançar resultados finais satisfatórios.

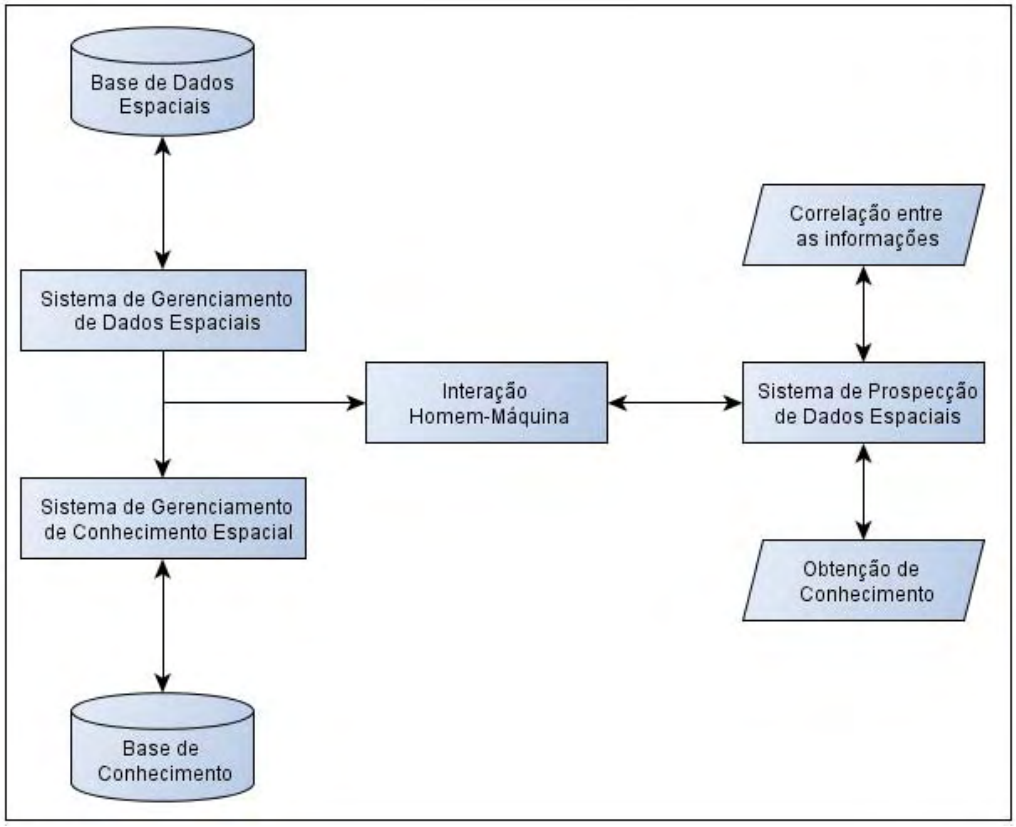


Figura 1 - Arquitetura do Processo de Prospecção de Dados Espaciais (Adaptado de [WAN_09a])

2.4 Pré-processamento de dados espaciais

A etapa de pré-processamento de dados espaciais pode ser considerada como uma das mais importantes no processo de prospecção, assim como no processo de KDD com dados convencionais.

Devido à grande quantidade de informações armazenadas, a maioria delas apresentam inconsistências, ausência de dados, duplicações e demais problemas. A presença de tais anomalias nos repositórios é refletida nos resultados que se obtém com a análise dos dados, o que pode configurar um padrão, uma correlação ou um indicativo errado [LIM_10].

Sendo assim, essa etapa inclui limpeza de dados, cujo objetivo é detectar casos de anomalias semelhantes aos já citados e corrigir esses problemas; integração de dados de múltiplas fontes, que visa obter maior quantidade de dados por meio da obtenção dos que estejam fisicamente armazenados em banco de dados diferentes; indução, que define o escopo da prospecção e evita o processamento de informações que não fazem parte de um mesmo contexto semântico alvo; por fim, a tomada de decisão de qual método de prospecção utilizar [WAN_09b] [CHA_00].

2.5 Métodos dos algoritmos de spatial data mining

Devido à grande quantidade de dados armazenados e à alta complexidade das características dos dados espaciais e das técnicas de armazenamento e recuperação desses dados, as pesquisas em tecnologias e métodos de prospecção de dados tem sido bastante intensas [WAN_09a].

Nesta seção, os métodos de prospecção de dados espacias são abordados, com base na divisão em quatro categorias propostas na literatura pelos autores Jin e Miao [JIN_10]: método de classificação espacial, método de agrupamento espacial (*spatial clustering*), regras de associação espacial e análise de tendência espacial. Apesar de serem encontradas na literatura outras categorias, esta é uma abordagem recente e, por isso, apresentada neste trabalho.

2.5.1 Método de classificação espacial

A classificação espacial é realizada pela análise de modelos de dados de acordo com suas características espaciais, como área de uma região, estrada ou rio. Esse método

também é aplicado para se obter regras de classificação entre atributos espaciais e não espaciais de objetos de um banco de dados espaciais.

Ester *et al* [EST_97] propuseram um primeiro método de classificação de objetos espaciais por meio do algoritmo ID3 [SHE_02]. O algoritmo considera propriedades não espaciais dos objetos classificados e de seus objetos adjacentes, isto é, aqueles que satisfazem qualquer tipo de relação de adjacência. No entanto, o algoritmo não oferece análise de propriedades não espaciais dos valores dos objetos adjacentes. Também não leva em conta o conceito de relação hierárquica que pode existir entre atributos não espaciais e atributos espaciais.

Kopersiki e Han [KOP_95] propuseram uma política de duas etapas de classificação de dados espaciais. Numa primeira etapa, obtém-se um predicado espacial aproximado com baixo custo computacional para cálculos espaciais e analisa-se a primeira correlação obtida; na segunda etapa, realiza-se outro cálculo e obtém-se uma correlação mais refinada, o que resulta então numa árvore de decisão mais precisa e menor [SHE_02].

Yan-hui e Hao [YAN_09] apresentaram um modelo para entidades espaciais baseado na teoria de banco de dados, em que a modelagem espacial hierárquica é proposta como uma extensão do modelo entidade-relacionamento (MER). Ainda nesse sentido, outro trabalho proposto descreve a formalização do modelo de dados para suportar a abstração de dados espaciais [JIM_10].

2.5.2 Método de agrupamento espacial

Clustering é um importante tema de pesquisa no campo de prospecção de dados. *Cluster* é o agrupamento de objetos com base na similaridade em que é encontrada a distribuição de características de dados espaciais, o que faz com que os dados de um agrupamento tenham semelhanças muito elevadas entre si e, em relação ao de diferentes agrupamentos, sejam bastante distintos [HAN_06] [MAR_97].

O método de agrupamento espacial consiste em organizar os objetos de um banco de dados espaciais em diferentes subclasses de acordo com alguns características, de forma que os objetos da mesma subclasse possuam tais características com alta similaridade. A principal vantagem do método de agrupamento espacial é a capacidade de se encontrar semelhança entre objetos espaciais diretamente a partir dos próprios dados, sem a necessidade de seu conhecimento semântico [JIN_10]. A qualidade dos resultados obtidos pela técnica de *Spatial Clustering* está relacionada com algumas características a serem definidas, tais como: algoritmo de agrupamento, número de *clusters, e*scolha de atributos,

homogeneização das variáveis, medidas de dissimilaridade e critérios de agrupamento [NEV_01].

Na literatura, foram levantados alguns métodos de agrupamentos espaciais que são brevemente discutidos nos subtópicos seguintes.

2.5.2.1 Método de agrupamento espacial baseado em particionamento

Os métodos baseados em particionamento – ou em segmentação – incluem o método K-avarage, o K-center e o CLARANS [NGR_02]. Eles adotam uma tecnologia iterativa de reposicionamento que tenta utilizar a movimentação dos grupos de objetos para melhorar o agrupamento. Uma vez que esse método é adequado para identificação de grupos de tamanhos similares, é frequentemente usado em aplicações como análise de localizações. No estado da arte são encontradas outras propostas de algoritmos de agrupamento espacial baseado na segmentação, brevemente descritos a seguir.

Karmaker e Rahman propuseram um trabalho, baseado nos algoritmos PAM, CLARA e CLARANS, cujo objetivo é detectar e minimizar os dados isolados de agrupamento, tratados como *outliers*, a fim de garantir a integridade e validação dos dados em grandes repositórios [KAR_09].

He Bingquan, Jiubin Wang e Chao Li propuseram uma melhoria ao algoritmo K-MEANS. Tal algoritmo classifica n objetos em k *clusters* – e apresenta melhor desempenho do que o original [HEB_10].

2.5.2.2 Método de agrupamento espacial baseado em hierarquia

Este tipo de abordagem consiste apenas na decomposição de uma coleção de objetos. De acordo com o modelo de decomposição hierárquica, esses métodos podem ser divididos em dois tipos: coesão e divisão. Dentre os algoritmos que implementam o método de agrupamento hierárquico, os mais conhecidos são: BIRCH [ZHA_96] e CURE [GUH_98].

Alguns trabalhos neste tema tem sido propostos no estado da arte, como o de Gui-Fen *et al.* [CHE_09] que apresenta um algoritmo de agrupamento espacial dinâmico baseado em lógica *fuzzy* e uso de pesos que calibram o algoritmo. Com uma abordagem semelhante, Li, Shi e Liu [LIB_10] desenvolveram um algoritmo baseado no conceito de incerteza para a classificação de objetos, em que um mesmo grupo pode ser subgrupo de outros, dependendo da semântica das informações neles contidas.

2.5.2.3 Método de agrupamento espacial baseado em densidade

Por meio deste método, a busca por agrupamentos é realizada com base na densidade dos dados e respectivos vizinhos, sendo um dos pioneiros em *data mining*. Geralmente, algoritmos deste método são utilizados para identificação de grupos de forma arbitrária e filtros de "obstáculos", embora os mais tradicionais apresentem baixa eficácia quando a densidade é bastante variada. Alguns deles são: DBSCAN [EST_96], GDBSCAN [SAN_98], DENCLUE [HIN_98], OPTICS [ANK_99] e DBRS [XIN_03], VDBSCAN [LIU_07].

2.5.2.4 Método de agrupamento espacial baseado em grid

Neste método utiliza-se uma *grid* de dados de múltiplas estruturas para dividir o espaço em um número limitado de unidades, nas quais realizam-se as operações de agrupamento. Os algoritmos mais conhecidos são: STING [WAN_97], WaveCluster [SHE_98] e CLIQUE [AGR_98].

Fan e Luo propuseram a utilização de um ambiente *grid* a fim de integrar dados de diferentes fontes numa única fonte virtualizada sob estratégia de partição de dados [FAN_09].

2.5.3 Regra de associação espacial

As regras de associação espacial são derivadas das regras de associação de prospecção de dados convencionais. A fórmula que as definem é A → B [s%; c%], em que **A** corresponde à coleção de predicados espaciais e **B** à coleção de predicados não espaciais; **s%** indica o grau de suporte das regras e **c%** corresponde à confiança das regras. Tal método foi proposto por Agrawal *et al.* em um estudo para prospecção em grandes repositórios de dados [AGR_93] e Koperski e Han aplicaram esse conceito para banco de dados espaciais [KOP_95].

Há diversos tipos de predicados espaciais que podem constituir uma regra de associação espacial, tais como: relações de topologia, orientações espaciais, informações de distância e outros.

Devido à prospecção de regras de associação espacial envolver muitos cálculos de uma variedade de relações espaciais entre muitos objetos espaciais, é conveniente que antes de sua aplicação seja realizado um refinamento inicial, por meio de algum algoritmo

para se efetuar uma prospecção inicial na base de dados espaciais e então efetuar uma nova na porção obtida da base.

Esen Kacar *et al* [KAC_02] propuseram um algoritmo para prospecção espacial com regras de associação *fuzzy*. Liu Junqiang *et al* [LIU_03] projetaram um algoritmo de prospecção de regras de associação em uma única camada do tipo *booleana* com base na PREFIX TREE (FPT-Generate).

Maddox e Shin [MAD_09] propuseram um algoritmo que detecta automaticamente, por meio de regras heurísticas, as dependências entre dados espaciais e outras variáveis. O trabalho é conveniente, pois o algoritmo efetua uma varredura por toda a base de dados em busca de regras de associação entre as informações espaciais de cada tupla.

2.5.4 Análise de tendências espaciais

Tendências espaciais referem-se às mudanças de atributos não espaciais quando se encontram cada vez mais distantes de um objeto espacial. Por exemplo, a tendência de mudança da situação econômica de um município acentua-se quando se está mais longe do centro da cidade [EST_97].

Normalmente, para se analisar a tendência espacial de uma estrutura de dados espaciais e métodos de acesso, é necessária a utilização de métodos de análise de regressão e outros métodos semelhantes. Devido à particularidade das características de objetos espaciais, o modelo de regressão tradicional não é apropriado.

2.6 Sistemas computacionais aplicados à prospecção de dados espaciais

Neste tópico são apresentados alguns dos sistemas computacionais propostos no estado da arte que implementam recursos de prospecção de dados espaciais, além da descrição de alguns trabalhos focados nos algoritmos.

Wang *et al* [WAN_09b] propuseram um sistema cuja arquitetura é modularizada e divida em três camadas. Na camada responsável pela prospecção, diversos módulos auxiliam neste processo para obtenção de conhecimento e o formato modularizado apresenta vantagens como maior eficiência, velocidade, portabilidade, além de alta interação com o usuário durante cada etapa do processo de prospecção.

O sistema Visual Geo-Classify System (VGCS), proposto por Zelu Jia e Yaolin Liu é um protótipo para classificação de dados espaciais e utiliza ferramentais visuais e inteligência artificial [JIA_09].

SD-Miner é o nome dado a outro sistema proposto na literatura, organizado em três grandes módulos: *interface* gráfica, prospecção de dados espaciais e gerenciamento de dados convencionais e espaciais. O módulo de prospecção de dados contempla funcionalidades de agrupamento espacial, classificação espacial, caracterização espacial e regras de associação espaço-temporais [BAE_09].

O VegaMinerPOI [PEN_09] é integrado com a plataforma VegaGIS 3.0 e oferece três técnicas de prospecção de dados espaciais. Sua característica diferencial é a possibilidade de processamento *multithread* que, segundo o autor, pode ser facilmente adaptado para contemplar modo de execução em paralelo.

O sistema GeoKSGrid [JIA_10], cuja arquitetura é dividia em 5 camadas – recursos de *Grid*, camada de dados da *Grid*, camada central de conhecimento, camada de conhecimento avançado e portal da *Grid* – propõe um processo distribuído de prospecção de dados espaciais.

Os autores He YueShun e Li Xiang [HEY_09] propuseram um sistema cuja arquitetura *Web* contempla as principais funcionalidades necessárias para a prospecção de dados espaciais, o que inclui suporte à integração de dados de múltiplas fontes por meio de arquivos XML.

Outro trabalho nesta área é o desenvolvido por Valêncio *et al.* [VAL_11], cuja ferramenta Web desenvolvida implementa o algoritmo CLARANS e o K-MEANS e foi aplicada à análise de dados de acidentes de trabalho.

Kondaveeti *et al* [KON_11] desenvolveram um SIG combinado com técnicas de *spatial data mining* e regras de associação a fim de suportar análise de dados provenientes da fronteira dos Estados Unidos com o México e identificar informações implícitas relacionadas a imigração ilegal, tráfico de armas e drogas e outros.

A extensão PostGIS do sistema gerenciador de banco de dados PostgreSQL é detalhadamente explorada por Jiehai e Wei [JIE_10] e se mostra uma boa alternativa para gerenciamento e manipulação de informações espaciais. Suas principais vantagens caracterizam-se por ser um sistema gratuito e consolidado no mercado corporativo.

2.7 Algoritmos no estado da arte

Alguns trabalhos tem sido propostos na literatura com o objetivo de abordar o problema de desempenho dos algoritmos de prospecção de dados espaciais conhecidos, que são apresentados na sequência.

Um desses trabalhos é o algoritmo VDBSCAN, proposto em 2007, que tem como característica a eliminação dos parâmetros de entrada exigidos pelo DBSCAN. A ideia consiste na definição automática dos parâmetros e na sensibilidade a variações de densidade de objetos no conjunto analisado, uma vez que o DBSCAN não apresenta desempenho satisfatório nessa situação [LIU_07].

Uma continuação desse trabalho foi proposta em 2010, com a ideia de otimizar uma abordagem do VDBSCAN na definição de um dos parâmetros a partir da análise das características do conjunto de dados [CHO_10].

Também em 2010, um algoritmo com a otimização do tempo de execução demandado pelo VDBCAN foi proposto. Pelos resultados obtidos, foi comprovada a eficiência do trabalho, uma vez que, enquanto o VDBSCAN requisitou 50,688 segundos para processar um conjunto de dados com 600 registros, com a nova abordagem, o tempo caiu para 10,5144 segundos [VIJ_10].

Outro algoritmo do estado da arte é o ST-DBSCAN, que também foi confeccionado a partir de uma modificação do DBSCAN, o que o tornou capaz de realiza a prospecção de dados espaciais com base nos valores não espaciais, espaciais e temporais dos objetos. O ST-DBSCAN efetua o agrupamento de dados com base nos três tipos de dados, diferentemente dos demais algoritmos que propõem agrupamento baseado em densidade. O mesmo foi aplicado num *data warehouse* que armazena grande volume de dados espaciais e demonstrou resultados que comprovaram a eficiência no que foi proposto [BIR_07].

Frank, Ester e Knobbe [FRA_09] apresentaram em 2009 uma abordagem multirrelacional para classificação espacial, com a qual uma nova maneira de extração de regras de classificação foi criada, chamada Unified Multi-relational Aggregation-based Spatial Classifier (UnMASC). Por meio dessa abordagem, relações entre vizinhança são criadas de forma a explicitá-las, uma vez que, em se tratando de dados espaciais, as relações muitas vezes são implícitas. A execução em conjuntos de dados espaciais reais apresentou ganhos em termos de precisão, se comparada a técnica proposta com outras

existentes, já que os experimentos sem a inclusão das informações levantadas pela técnica resultaram em baixa precisão.

Outro trabalho é o desenvolvido por Xue Jing-Sheng [XUE_10], que propõe um algoritmo, com base no CLARANS melhorado e paralelizado. O comportamento do algoritmo é o mesmo do CLARANS com poucos dados. No entanto, com grande volume de dados, o tempo de execução do CLARANS cresce exponencialmente, enquanto que do PCLARANS, como é chamada a versão paralelizada, cresce linearmente, tornando-se consideravelmente superior em relação ao anterior.

Por fim, um trabalho que também apresenta uma abordagem otimizada é o algoritmo CCDD [HUA_10], proposto em 2010, que consiste no aprimoramento do DBSCAN – algoritmo do método de agrupamento baseado em densidade. Uma das dificuldades desse algoritmo se configura quando os agrupamentos apresentam densidades diferentes. Isso devido ao DBSCAN utilizar parâmetros globais - distância máxima entre dois objetos para estarem em um mesmo agrupamento e a quantidade de objetos em um agrupamento - o que o torna pouco eficiente quando há no conjunto de dados diferentes densidades.

Para contornar este problema, o algoritmo CCDD utiliza uma abordagem na qual a definição dos parâmetros é dinâmica, de acordo com o contexto da densidade em que se formarão os agrupamentos. Os resultados evidenciam melhor desempenho em relação ao DBSCAN tanto em qualidade dos agrupamentos retornados, quanto ao tempo de execução. No entanto, uma das propostas futuras é a implementação da escolha dos parâmetros automaticamente.

2.8 O futuro das pesquisas relacionadas à prospecção de dados espaciais

A prospecção de dados espaciais é uma área promissora, no entanto há muitos problemas que requerem um estudo mais aprofundado. Devido aos dados espaciais possuírem características de massa, não-linearidade, múltiplas escalas e ambiguidade, a tarefa de extração de conhecimento a partir de um banco de dados espaciais é mais difícil do que extrair conhecimento de um banco de dados relacional tradicional.

Esse panorama traz desafios à área de prospecção de dados espaciais e algumas direções para pesquisas envolvem [JIN_10]:

1) **Desenvolvimento de algoritmos e tecnologias de prospecção de dados espaciais -** Algoritmos de prospecção de regras de associação espaciais, técnicas de prospecção de dados eficientes, algoritmos específicos para atributos espaciais, novas técnicas de classificação espacial, algoritmos para detecção de *outliers*, que representam uma das grandes frentes da área de prospecção de dados espaciais devido à relação com eficiência dos algoritmos de prospecção de dados espaciais;
2) **Pré-processamento de múltiplas fontes de dados espaciais -** Dados espaciais incluem dados de linha digital, imagem e modelos de elevação digital e dados de características de superfície. Devido à própria complexidade estrutural da informação espacial e suas relações, é inevitável que seus atributos, dados sobre obstáculos e inconsistências oriundas de múltiplas fontes exijam um esquema de pré-processamento eficiente;
3) **Pesquisa relacionada a outros tipos de prospecção de dados espaciais e tecnologias relacionadas -** Há algumas tendências nas pesquisas em teoria e metodologias de prospecção de dados espaciais, as quais valem ser citadas: prospecção de dados espaciais em ambiente de rede, prospecção de dados visuais, integração de prospecção de dados espaciais com redimensionamentos e vetorizações, geradores automáticos de árvore de *background knowledge*, prospecção de dados com base em incertezas espaciais (local, atributos temporais, etc), prospecção de dados incrementais, de multi-resolução e multi-níveis, prospecção de dados em paralelo, prospecção de dados com base no banco de imagens de sensoriamento remoto, descoberta de conhecimento multimídia em bancos de dados espaciais, linguagem de consulta de prospecção de dados espaciais e expressão visual de regras.

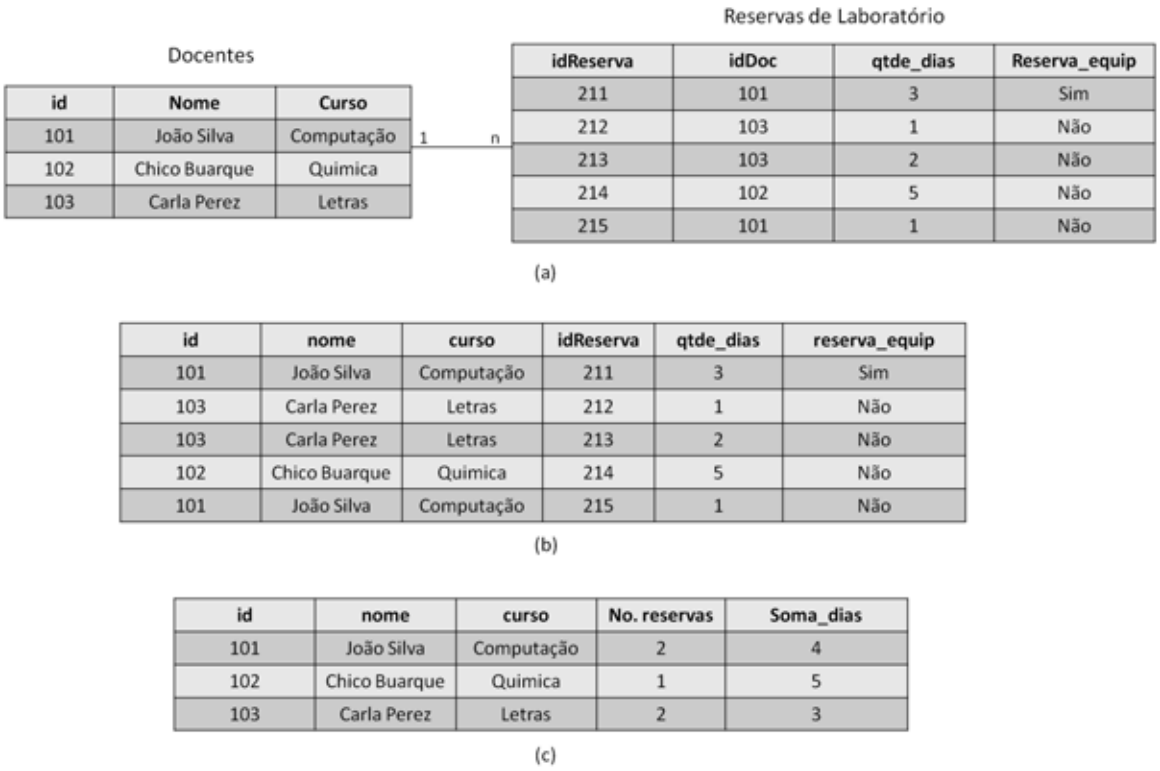
Para a implementação de um sistema de prospecção de dados espaciais, deve-se haver uma junção de sistemas de informações geográficas, sensoriamento remoto, sistema de interpretação técnica, integração de dados espaciais e integração de sistemas de apoio à decisão espaciais.

2.9 Abordagem multirrelacional e os dados espaciais

Nas abordagens tradicionais, considera-se que os dados a serem analisados estão dispostos numa única estrutura. Sendo assim, a aplicação de algoritmos que atuam dessa maneira exige a junção dos dados, quando os mesmos estão dispostos em várias tabelas

semanticamente relacionadas – no caso de um banco de dados relacional. Essa operação acarreta, muitas vezes, em perda de informação ou em geração de imprecisões que não seriam retratados se os dados fossem analisados numa abordagem que conserve a natureza da forma como são armazenados, além de exigir um alto custo computacional e poder implicar em grande quantidade de dados replicados [TSE_99] [DEZ_03].

Um exemplo da perda semântica e erros que podem ser gerados pela junção de dados é o exemplo retratado na Figura 2, em que estão representadas tabelas e relações de uma base de dados fictícia que armazena dados sobre reservas de laboratórios.



Docentes

id	Nome	Curso
101	João Silva	Computação
102	Chico Buarque	Quimica
103	Carla Perez	Letras

Reservas de Laboratório

idReserva	idDoc	qtde_dias	Reserva_equip
211	101	3	Sim
212	103	1	Não
213	103	2	Não
214	102	5	Não
215	101	1	Não

(a)

id	nome	curso	idReserva	qtde_dias	reserva_equip
101	João Silva	Computação	211	3	Sim
103	Carla Perez	Letras	212	1	Não
103	Carla Perez	Letras	213	2	Não
102	Chico Buarque	Quimica	214	5	Não
101	João Silva	Computação	215	1	Não

(b)

id	nome	curso	No. reservas	Soma_dias
101	João Silva	Computação	2	4
102	Chico Buarque	Quimica	1	5
103	Carla Perez	Letras	2	3

(c)

Figura 2 – Exemplo de junção de dados e perda semântica

Na Figura 2(a), é representada uma relação da tabela de docentes com a de reservas de laboratório efetuadas pelos mesmos. Nesse exemplo, um docente pode ter uma ou mais reservas, no entanto, uma reserva pode ser de apenas um docente.

Na Figura 2(b), por sua vez, está ilustrado um exemplo de junção de dados de docentes e reservas de laboratório, para fins de prospecção. Nessa figura é importante observar a redundância de dados - no exemplo, dos docentes João Silva e Carla Perez - o que remete à distorção dos resultados obtidos a partir dessa tabela resultante da junção.

Finalmente, na Figura 2(c) é ilustrado um exemplo de perda semântica, sendo que uma função de agregação foi utilizada para contabilizar o número de reservas realizadas para cada docente, assim como outra para realizar a soma das quantidades de dias reservados de laboratório, o que significa eliminação de redundância de dados e perda do

número de dias de cada reserva. Nesse caso, também não foi possível incluir o atributo “reserva_equip”, que indica se o docente fez a reserva dos equipamentos do laboratório ou não, por não poder ser sumarizado nessa relação, o que poderia influenciar nos resultados de uma prospecção desses dados.

Em se tratando de prospecção de dados espaciais, as replicações e perda semântica rementem à representação espacial dos objetos de forma errônea, o que influencia diretamente nos resultados obtidos. A abordagem multirrelacional é mais recente e consiste na análise dos dados diretamente nas relações em que estão armazenados, sem que o processo de junção seja realizado como preparação para a prospecção [DOM_03].

A prospecção de dados convencionais com abordagem multirrelacional é mais difundida e as técnicas consistem, em sua maioria, na adaptação dos algoritmos tradicionais para o contexto em que foram propostos. Na prospecção de dados espaciais, essa abordagem é bastante recente e promissora [FRA_09], uma vez que diante do levantamento bibliográfico realizado, foi encontrado apenas um trabalho que propôs a implementação que utiliza uma estrutura multirrelacional para definição de regras de classificação em agrupamentos.

2.10 Considerações finais

Neste capítulo foram apresentados um levantamento bibliográfico dos conceitos básicos e o estado da arte na área de prospecção de dados espaciais, também conhecida por base de conhecimento espacial.

As técnicas aplicadas à prospecção de dados espaciais evoluíram rapidamente, mas ainda há várias teorias e tecnologias que precisam a ser melhoradas e ampliadas em termos de aplicação. Notou-se também que a prospecção de dados espaciais tem avançado significativamente por ser aplicável a diversas áreas da ciência da computação e em outros temas, uma vez que possibilita uma maior compreensão do mundo por meio do conhecimento descoberto.

Capítulo 3 Algoritmo para prospecção multirrelacional de dados espaciais

3.1 Considerações iniciais

Neste capítulo, o algoritmo proposto, nomeado MR-Clustering, é descrito desde a sua concepção à sua implementação. Para isso, foram criados alguns fluxogramas para facilitar o entendimento da ideia, assim como do comportamento do mesmo durante sua execução.

Inicialmente, é importante ressaltar que o algoritmo proposto foi concebido com base em algumas abordagens encontradas na literatura pelos algoritmos bastante utilizados na área – o CLARANS, de agrupamento baseado em particionamento; o DBSCAN, de agrupamento baseado em densidade; e o VDBSCAN, versão melhorada do anterior. O resultado é uma abordagem diferenciada do que se tem no estado da arte, uma vez que agrupamentos otimizados e análise de dados espaciais em ambiente multirrelacional são contemplados, com aplicabilidade em grandes repositórios de dados.

3.2 Os algoritmos bases para o MR-Clustering

Diversas são as categorias dos algoritmos para prospecção de dados espaciais, dentre elas a de agrupamento. Em relação às demais, a técnica de *clustering* é considerada uma das mais utilizadas, pela sua característica de identificar estruturas a partir dos dados sem que seja necessário conhecimento prévio dos mesmos.

Por conta do melhor desempenho apresentado, o método de particionamento tem sido o mais utilizado e investigado pelas pesquisas. Os mais comuns são os baseados em ponto central – nomeado de K-MEANS – ou em objeto representativo para o agrupamento – K-MEDOID.

O K-MEDOID consiste na utilização de um objeto representativo do agrupamento localizado ao centro do mesmo, diferentemente do K-MEANS, que utiliza um centro médio. Embora o K-MEDOID seja menos sensível ao ruído, exige um tempo de processamento maior do que o K-MEANS.

Dentre os algoritmos baseados no K-MEDOID destaca-se o CLARANS, criado para melhorar a qualidade e a escalabilidade do CLARA, que já apresentava melhor desempenho em relação ao PAN, outro algoritmo bastante conhecido desse método. Devido ao fato do CLARANS realizar a busca por um conjunto ótimo de K-MEDOIDS num subconjunto das possíveis combinações de objetos de forma contínua e com certa aleatoriedade, o mesmo apresenta melhores resultados em relação ao CLARA e mais rápido que o PAN e, por isso, é considerado por Ng e Han um método eficiente para *spatial data mining* [NGR_02] [KAR_09].

O DBSCAN, por sua vez, dentre os algoritmos de agrupamento baseado em densidade, é o que apresenta maior popularidade. No entanto, a exigência de parâmetros de entrada muitas vezes compromete os resultados obtidos, por conta da influência que provocam na formação dos agrupamentos.

Finalmente, o VDBSCAN, também da categoria de agrupamento baseado em densidade, é uma evolução do DBSCAN, que propõe uma abordagem para eliminação dos parâmetros exigidos de entrada e capaz de identificar agrupamentos em diversos níveis de densidade [PAR_11].

Por conta das características dos algoritmos apresentadas nesse tópico, o MR-Clustering foi concebido a partir de algumas abordagens que tem contribuído para a evolução das técnicas de prospecção de dados espaciais, juntamente com a proposição de novas abordagens para tal.

3.3 Abordagem multirrelacional

Conforme já anunciado no Capítulo 2, nas abordagens tradicionais, considera-se que os dados a serem analisados estão dispostos numa única estrutura, o que faz com que se tenha abertura para perda de informação ou para geração de imprecisões facilmente contornadas por uma abordagem multirrelacional.

Nesse contexto, diante da carência de algoritmos que abordam os dados espaciais numa perspectiva multirrelacional, uma estratégia para seleção de dados foi concebida para que a prospecção dos mesmos seja realizada sem a necessidade de junção de dados, o que contorna as desvantagens do processo com essas condições. Dessa maneira, o leque de aplicações é ampliado, uma vez que possibilita a análise de dados de forma a preservar a semântica contida, muitas vezes, nas relações entre as tabelas em banco de dados relacional.

Embora as vantagens apresentadas sejam de grande relevância, a implementação do algoritmo deve considerar o fato dos dados alvos de análise estarem dispostos em múltiplas relações e, por isso, uma estrutura para absorver essa situação foi criada, conforme esquema da Figura 3. Na estrutura definida, cada uma das características multivaloradas do objeto georreferenciado é armazenada em uma lista e indexada por meio de um *hash*. As características que assumem um único valor, por sua vez, são armazenadas em uma única lista.

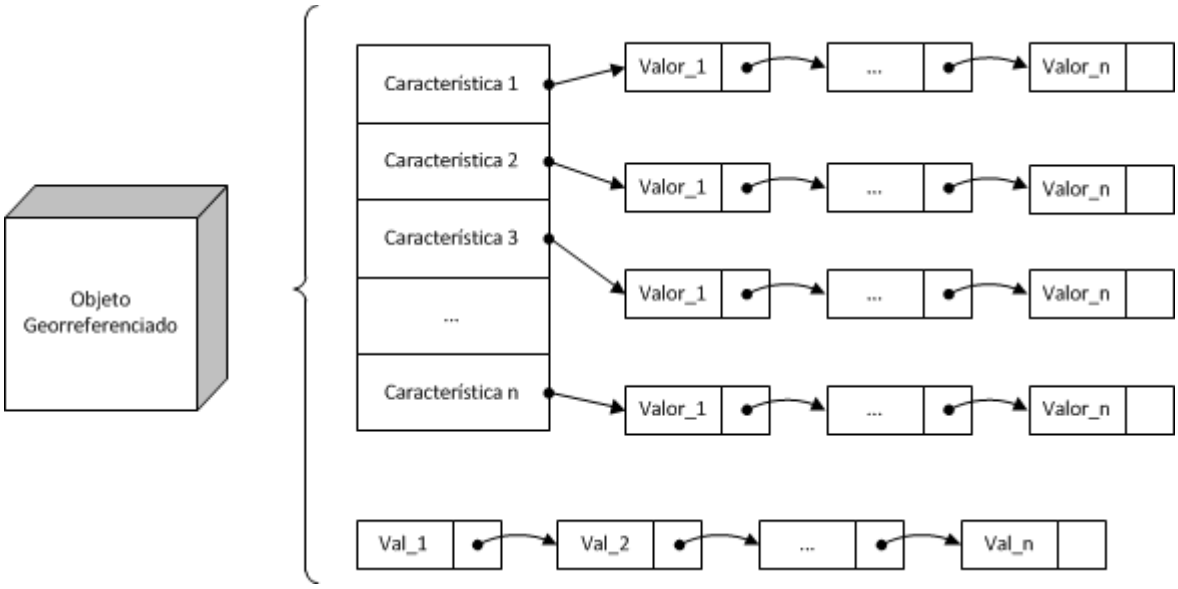


Figura 3 – Estrutura de dados para mapeamento dos atributos a serem analisados na prospecção de dados multirrelacional

A partir dessa estrutura, o algoritmo acessa os valores das características de cada objeto georreferenciado e efetua a comparação com outro objeto para determinar se há similaridade entre eles ou não. Essa comparação é feita característica a característica, seja multivalorada ou monovalorada, para então verificar se atendem ao índice similaridade dado como entrada no algoritmo.

A maneira como é organizada a representação dos objetos espaciais possibilita a agregação semântica no processo de agrupamento, já que a estrutura de dados utilizada é capaz de absorver o enriquecimento das características a serem analisadas por meio de uma inserção simples nas listas que as armazenam. Essa agregação de semântica pode ser viabilizada por ontologias ou outras fontes de dados, o que representa uma contribuição bastante relevante introduzida pelo algoritmo, uma vez que pode resultar em maior qualidade das informações obtidas com classificação semântica.

3.4 O MR-Clustering

O algoritmo proposto é da categoria de agrupamento com base na densidade, porém com combinação de técnicas para promover agrupamento de dados com base na similaridade de características não espaciais monovaloradas ou multivaloradas e da proximidade geográfica entre os objetos georreferenciados, de forma a não considerar os objetos isolados - ruídos ou *noises*, que distorcem a formação dos agrupamentos e não estabelecem um padrão ou nível de densidade de objetos - e ser capaz de identificar agrupamentos com robustez aos níveis de densidade.

Diferentemente da maioria dos trabalhos, o algoritmo pode produzir melhores resultados, uma vez que o único parâmetro de entrada não impede que o mesmo desempenhe a melhor organização dos agrupamentos, sendo que influencia apenas no grau de similaridade entre os objetos agrupados. No CLARANS, por exemplo, dependendo dos parâmetros definidos para a execução, os resultados obtidos podem não ser ótimos, já que podem apresentar diferenças significativas conforme a variação dos valores desses parâmetros.

No algoritmo proposto, como já afirmado, o único parâmetro de entrada é um índice de similaridade de características não espaciais na formação dos agrupamentos. Para que isso seja obtido, foram combinadas algumas técnicas dos algoritmos já citados. A ideia geral do mesmo pode ser melhor entendida a partir do fluxograma representado pela Figura 4.

Figura 4 – Fluxograma geral do algoritmo

Observa-se que o algoritmo pode ser dividido em três etapas:

- Inicialização – Toda a fase de preparação para que seja iniciada a formação dos agrupamentos é realizada. Isso inclui a formação do k-dist set, que é um conjunto de distâncias aferidas de cada objeto a seu k-ésimo vizinho, e a definição dos valores referentes aos níveis de densidade da base de dados, nomeados de Eps, que também podem ser definidos como medida de proximidade;
- Refinamento – Nesta etapa, uma técnica para melhorar a qualidade dos resultados obtidos é implementada com base na eliminação dos objetos isolados – também chamados de *noises* ou *outliers* – na definição dos agrupamentos e consiste no ajuste dos valores de Eps definidos na etapa de inicialização;

- Agrupamento – Depois de estabelecidos os inicializadores do algoritmo de forma automática, assim como o refinamento dos mesmos, o processo de agrupamento é iniciado e os resultados são obtidos de forma iterativa.

Apesar do fluxo do algoritmo se assemelhar ao do VDBSCAN, diversas diferenças podem ser notadas na descrição das etapas, além da nomeada etapa de refinamento. Nas seções seguintes, cada uma das etapas é detalhada.

3.4.1 Etapa de inicialização

Conforme descrito anteriormente, a primeira etapa do algoritmo consiste na definição automatizada dos parâmetros utilizados na formação dos agrupamentos. Poucos são os algoritmos que apresentam essa abordagem, uma vez que muitos exigem a entrada de parâmetros, que influenciam diretamente nos resultados retornados e, muitas vezes, por conta disso, o usuário é obrigado a ajustá-los por meio de tentativas até que se obtenha uma convergência para valores próximos do ótimo, por isso a referida baixa qualidade dos resultados como um fator a ser trabalhado, de acordo com o panorama da área prospecção de dados espaciais levantado na literatura.

Na etapa de inicialização, tomou-se como base a abordagem do VDBSCAN, por conta da estratégia de definição automática dos parâmetros exigidos pelo DBSCAN para agrupamento, no entanto diferencia-se do mesmo em alguns aspectos para fins de otimização de tempo de execução, qualidade dos resultados e ampliação do domínio de aplicação. Antes de descrever como isso é operado, é relevante apresentar os detalhes da etapa num fluxograma, conforme a Figura 5.

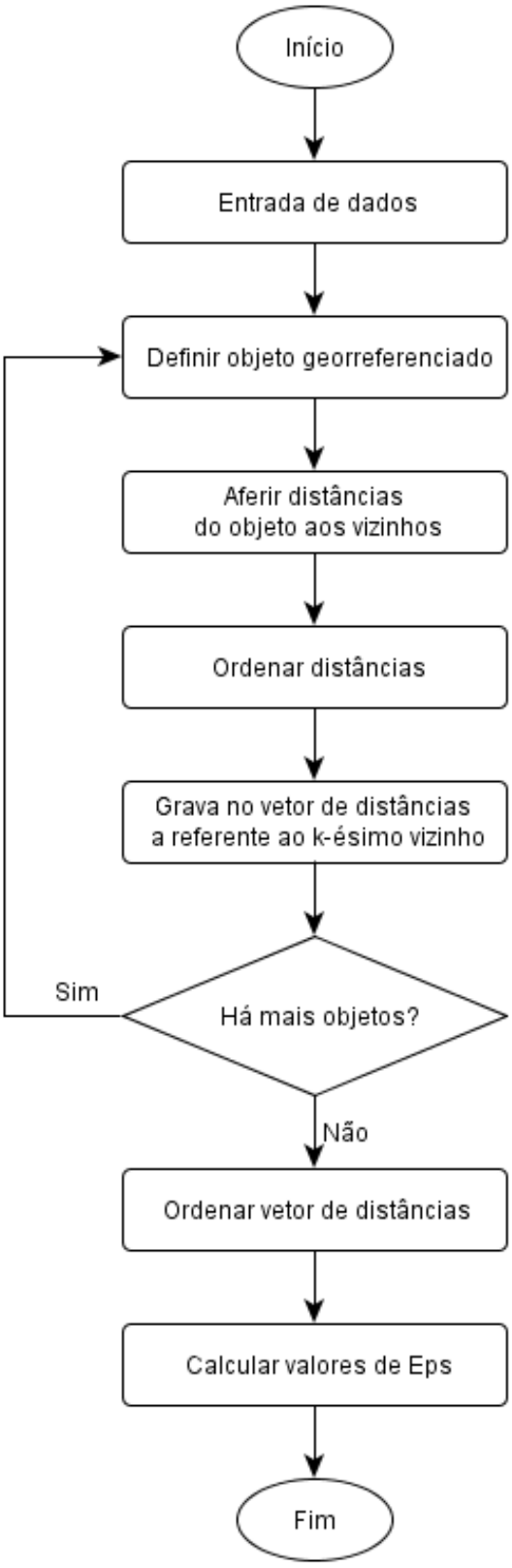


Figura 5 – Fluxograma detalhado da etapa de inicialização

Assim como proposto pelo VDBSCAN, inicialmente efetiva-se a definição dos valores de Eps ideais que, na terceira etapa, equivalem ao número mínimo de objetos dentro de um agrupamento. Para isso, um processo de aferição de distâncias geográficas entre os objetos deve ser iniciado, o que pode ser descrito como a obtenção do k-dist set, que contempla a distância de um objeto até o seu k-ésimo vizinho mais próximo, para cada objeto. Sendo assim, a primeira tarefa executada é a aferição dessas distâncias para todos os objetos georreferenciados da base de dados a ser analisada e, após isso, a ordenação dessas distâncias é efetuada. Para facilitar a compreensão por meio de recursos visuais, um gráfico é exibido na Figura 6, o qual evidencia a distribuição dos objetos e os níveis de densidade formados.

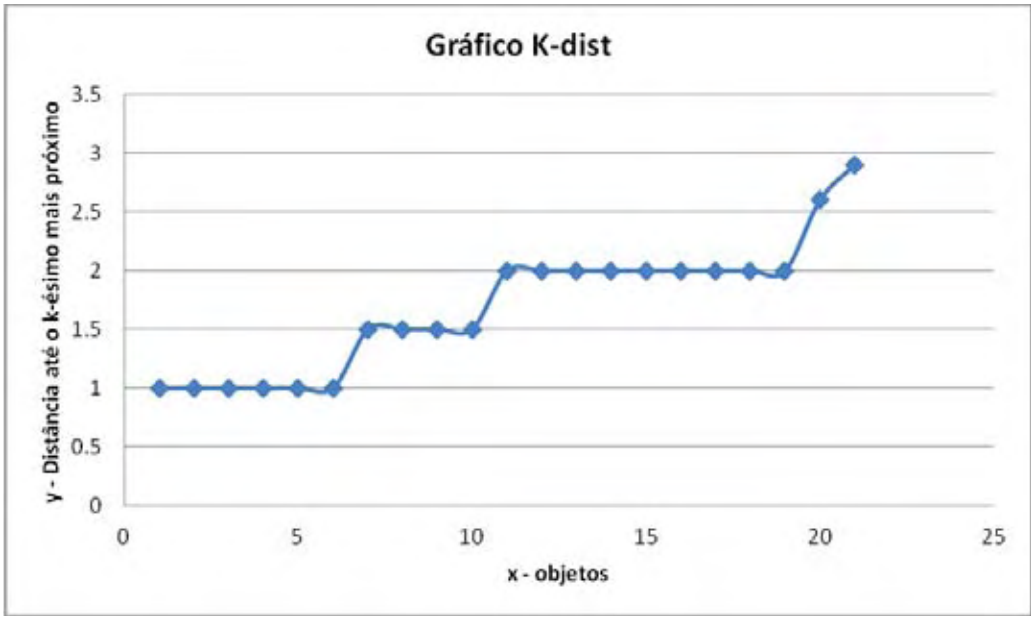


Figura 6 – Gráfico gerado a partir dos valores do k-dist set

No gráfico da Figura 6, o eixo x representa cada um dos objetos e o eixo y, as distâncias de cada um até seu k-ésimo vizinho mais próximo. Supondo que, no caso do exemplo citado, k=10, a distância aferida é entre um objeto e seu décimo vizinho mais próximo.

Nessa representação visual, os Eps correspondem aos valores nos pontos em que o tracejado entre as marcações apresenta uma curva mais acentuada, por exemplo, entre a distância 1 e 1,5; 1,5 e 2; e entre 2 e 3 – que são os valores determinantes de variação de densidade. Portanto, nesse caso há 3 valores de Eps a serem definidos.

Originalmente, o valor de k é aleatório, uma vez que por meio dos experimentos realizados pelos pesquisadores que propuseram a abordagem para definição automática dos

parâmetros do DBSCAN constatou-se que os resultados não apresentam grandes variações a medida que k muda. O processo se resumo basicamente aos passos seguintes e também representados na Figura 5:

Entradas

k - parâmetro para cálculo de distâncias

Procedimento

Para cada objeto:

Aferir a distância a todos os objetos

Organizar as distâncias em ordem crescente

Escolher o k-ésimo menor valor

Saída

dist[] – distâncias até o k-ésimo vizinho de cada objeto

Depois de obtida a distância até o k-ésimo vizinho mais próximo para cada objeto, os valores são ordenados crescentemente a fim de que se obtenha algo semelhante ao gráfico da Figura 6, quando plotados. A obtenção dos valores de Eps então é feita conforme os passos a seguir:

Entradas

dist[] - vetor de distâncias em ordem crescente

Procedimento

Faça min = dist[1] - dist[0]; max = min; threshold = 0; i = 0

Para i = 2 até a última posição no vetor dist[]

aux = dist[i] - dist[i-1];

Se aux maior que max, então max = aux;

Se aux menor que min, então min = aux;

Faça threshold = (min + max) / 2;

Para i=1 até a última posição no vetor dist[]

Se dist[i] - dist[i-1] é maior que threshold, então

Eps[j] = dist[i-1]

Saída

Eps[] - vetor de valores de Eps obtidos

Sendo assim, ao final dessa etapa, os parâmetros iniciais são definidos e preparados para a próxima etapa, de refinamento dos valores, para então iniciar a formação dos agrupamentos.

3.4.2 Etapa de refinamento

Depois de finalizada a primeira etapa, inicia-se a intermediária nomeada de refinamento, importante para que os valores definidos automaticamente na fase de inicialização sejam validados e, se for o caso, recalculados, o que garante uma precisão que contribui de maneira significativa na qualidade dos resultados obtidos.

Na etapa anterior, os valores de Eps foram definidos a partir de uma análise realizada sobre as distâncias aferidas e disponibilizadas no k-dist set. Como já esclarecido anteriormente, visualmente, em cada elevação do gráfico k-dist é definida uma mudança do nível de densidade, a qual é aproximada num valor de Eps obtido ao final da etapa de inicialização. No entanto, a presença dos objetos isolados influencia nos resultados desse cálculo, uma vez que tal característica pode provocar um desvio se os valores mais elevados de distâncias forem considerados níveis diferentes de densidade, e não como representação dos objetos isolados.

Na Figura 7, o referido caso pode ser melhor visualizado. Sabe-se que a variável que determina se uma elevação pode ser considerada uma mudança de nível de densidade – tratada como *threshold* no algoritmo – é obtida a partir de uma média das diferenças de distâncias aferidas, ou seja, visualmente, corresponde a uma média das elevações na curva do gráfico k-dist e, por isso, o trecho destacado em vermelho é determinante nesse cálculo, pois é nele que os objetos isolados podem estar presentes e não caracterizar níveis de densidade a serem considerados e, portanto, não devem influenciar no valor de *threshold* que determina os valores de Eps resultantes nas variações anteriores.

Sendo assim, nessa etapa o referido refinamento é realizado sobre os valores de Eps calculados a partir da eliminação dos objetos isolados na definição do *threshold*. Com isso, novos valores de Eps podem ser obtidos e a etapa posterior – de agrupamento – pode retornar melhores resultados, já que novos valores de Eps implicam em novos níveis de densidade ou definição mais precisa das variações dos mesmos e, portanto, resultados refinados e com melhor qualidade.

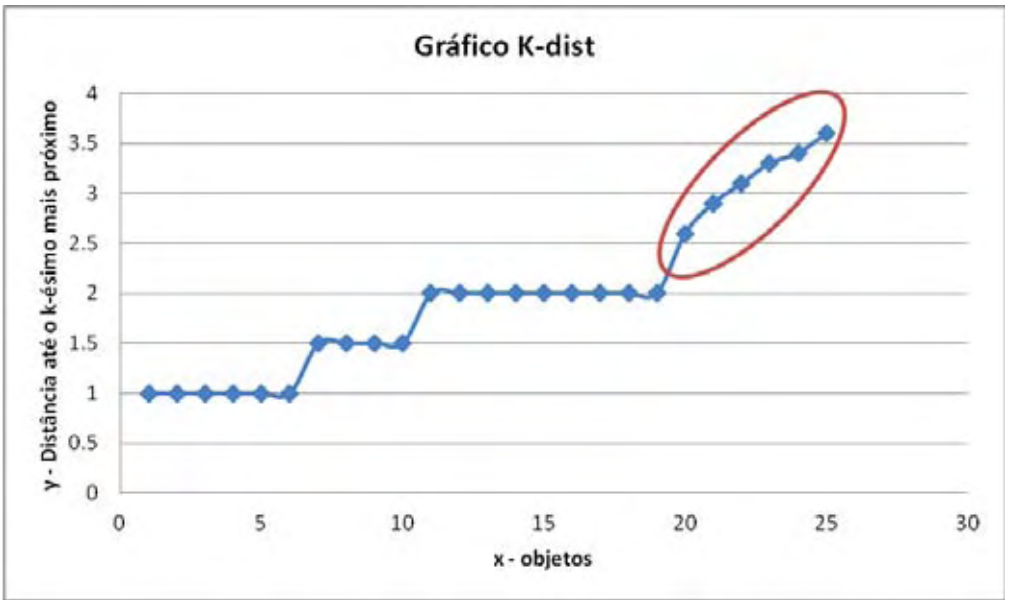


Figura 7 – Pontos isolados que podem influenciar nos valores de Eps

A partir dessa descrição geral da etapa, inicia-se então o detalhamento da implementação dessa ideia. Para isso, é relevante considerar que o algoritmo DBSCAN propõe como passo inicial a eliminação de objetos isolados no processo de agrupamento e, portanto, nessa etapa de refinamento o mesmo é adotado como parte desse processo. Para que isso seja melhor entendido, apresenta-se o fluxograma detalhado da Figura 8.

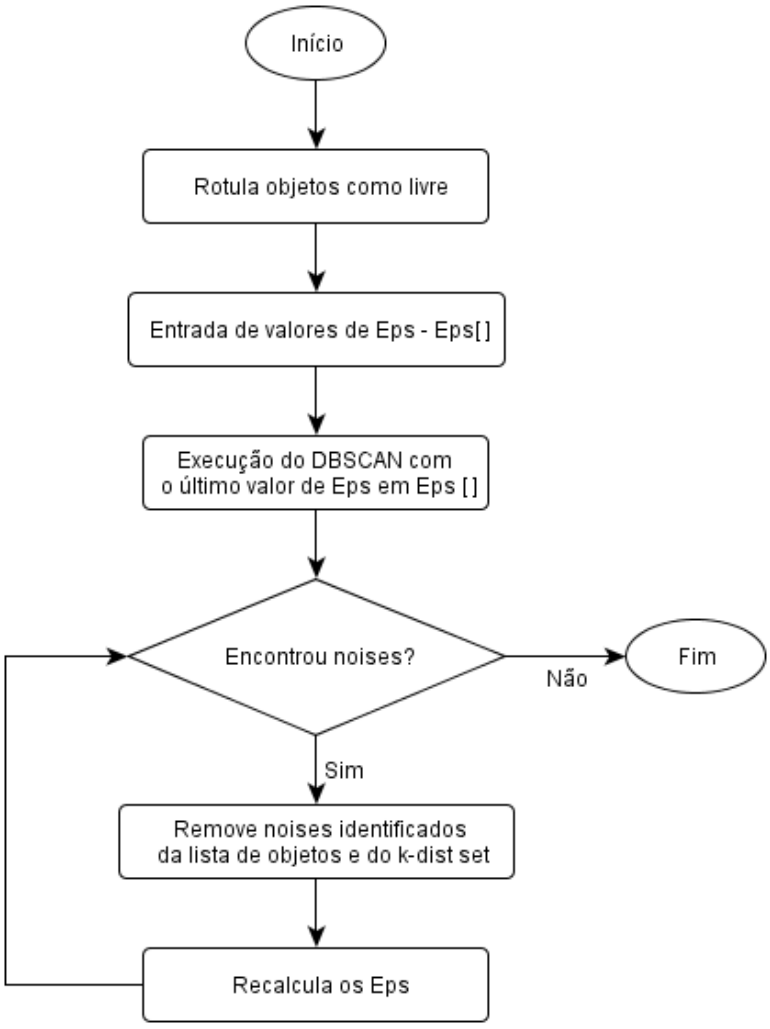


Figura 8 – Fluxograma detalhado da etapa de refinamento

Observa-se pelo fluxograma que ao iniciar essa etapa todos os objetos georreferenciados são primeiramente rotulados como livres. O passo inicial é a execução do DBSCAN para o último valor de Eps definido, ou seja, o que se refere à região crítica de elevação da curva do gráfico k-dist, para então disparar a busca pelos objetos isolados e demarcá-los como excluídos do cálculo dos valores de Eps, conforme os seguintes passos:

Entradas

k – valor utilizado no cálculo das distâncias

Eps[] – vetor com valores de Eps definidos na etapa anterior

n – número de posições de Eps[]

Procedimento

Executar DBSCAN (k, Eps[n])

Enquanto houver objetos georreferenciados livres e não rotulados como excluídos do cálculo de Eps

Recalcular valores de Eps sem considerar os objetos demarcados como noise

Executar o DBSCAN com número mínimo de objetos igual a k e Eps equivalente ao último valor definido recentemente

Saída

Eps[] refinado

É possível concluir que essa etapa se assemelha ao calibre manual dos valores de Eps para execução no DBSCAN, que, originalmente, utiliza um único valor como medida de proximidade e, por isso, a baixa qualidade nos agrupamentos em situações em que a densidade é bastante variada. Sendo assim, embora o MR-Clustering atue com a definição dos níveis de densidade do conjunto de dados espaciais, assim como proposto pelo VDBSCAN, o mesmo apresenta como contribuição à formação de melhores resultados o referido refinamento de valores de Eps, o que influencia diretamente nos agrupamentos em diferentes níveis de densidade.

Finalizada essa etapa, as variáveis necessárias para a execução do agrupamento estão definidas de forma automática, refinadas em seus valores para melhor absorver o contexto de variação de densidade e prontas para se iniciar a última etapa do algoritmo.

3.4.3 Etapa de agrupamento

Após as duas etapas preliminares, nesta a formação de agrupamentos de objetos georreferenciados com base na similaridade de características não espaciais e na localização geográfica é iniciada a partir dos parâmetros definidos nas etapas de inicialização e de refinamento – o k passa a corresponder ao número mínimo de objetos num agrupamento e os valores de Eps determinarão se um objeto é próximo geograficamente ou não, em cada iteração do algoritmo. Além desses, é utilizado também

o índice de similaridade escolhido inicialmente para a verificação das características não espaciais dos objetos a serem agrupados.

Nessa fase é que a abordagem multirrelacional idealizada foi implementada, a fim de que se possibilite a análise de dados sem a necessidade de junção dos mesmos numa única relação, o que apresenta, dentre outras vantagens, a conservação semântica.

O propósito dessa etapa consiste na execução iterativa e com os parâmetros definidos inicialmente do que foi chamado de DBSCAN modificado – eficiente em identificar agrupamentos, uma vez que herda as características do DBSCAN original, e diferenciado do mesmo por implementar a análise das características não espaciais dos objetos a serem agrupados a partir da estrutura de dados utilizada para suportar o contexto multirrelacional.

Os passos a seguir resumem o processo executado nessa última etapa, numa visão macro:

Entrada

k – valor utilizado no cálculo das distâncias
Eps[] – vetor de valores de Eps definidos
Índice de similaridade

Procedimento

Para cada Eps de Eps[]
Executar o DBSCAN modificado no conjunto de objetos ainda não agrupados
Marcar os objetos agrupados

Saída

Agrupamentos gerados

Para melhor descrever o DBSCAN modificado, um fluxograma que detalha os passos seguidos nessa abordagem é apresentado na Figura 9. Observa-se que, para cada objeto georreferenciado na base de dados, o MR-Clustering realiza uma análise dos seus vizinhos, que consiste na demarcação dos similares.

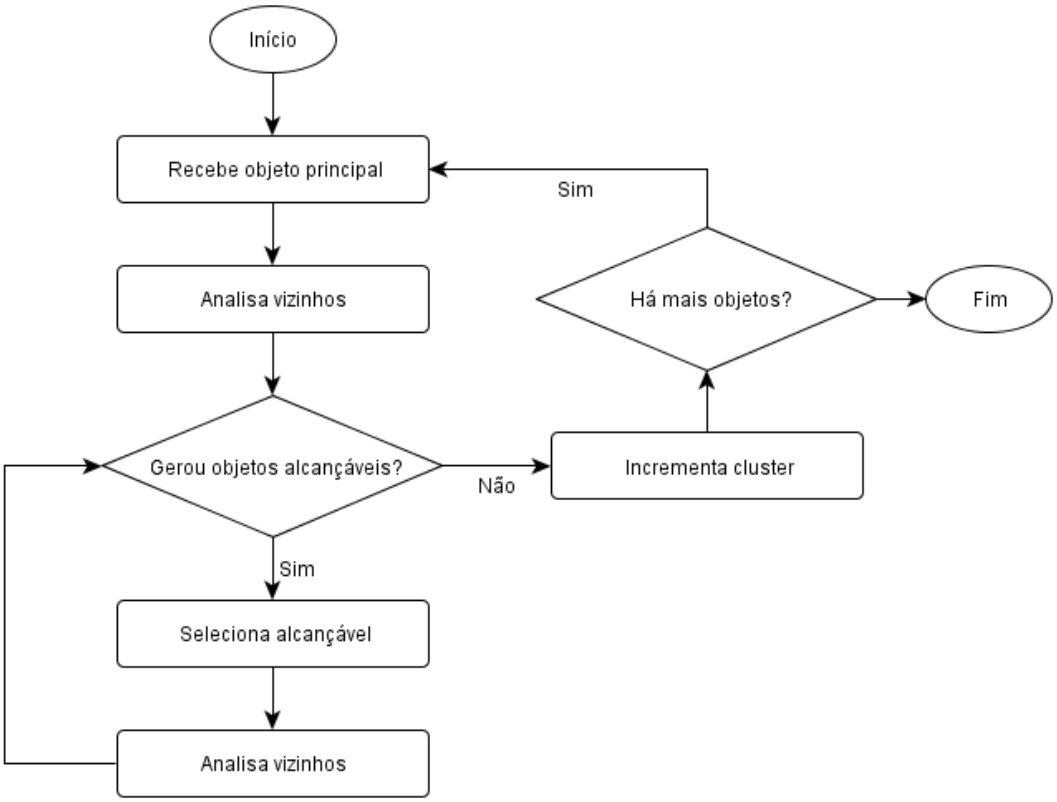


Figura 9 – Algoritmo DBSCAN modificado executado na etapa de agrupamento

O cálculo de similaridade de valores não espaciais é realizado no passo “analisa vizinhos” a partir da comparação das características dos objetos que compõem as listas da estrutura de dados implementada para suportar a abordagem multirrelacional, na qual verifica-se a porcentagem de características idênticas, que passa a ser considerada se maior ou igual ao indicado como entrada pelo usuário.

Para o cálculo da similaridade geográfica, na Figura 10, um esquema visual do procedimento de classificação dos objetos é ilustrado. Nessa etapa, somente os similares em características não espaciais são considerados para essa classificação. Os objetos em roxo são centro, pois numa circunferência de raio Eps em torno de si o número de objetos abrangidos é maior ou igual ao número mínimo exigido num *cluster*. Os objetos em azul são alcançáveis, pois estão numa região de raio Eps em torno do centro; os em verde são borda, pois são alcançáveis, porém num raio de Eps não atingem um número mínimo de pontos exigidos pelo *cluster*; finalmente, o vermelho é *noise*, pois não é alcançável por nenhum dos objetos.

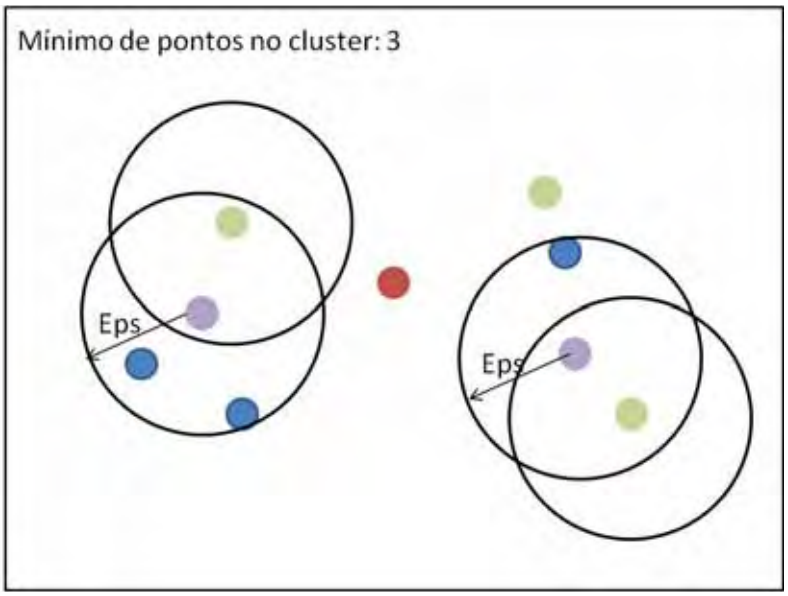


Figura 10 – Comportamento do DBSCAN na classificação dos pontos

Para entender de forma mais detalhada como funciona a análise dos objetos e a formação dos agrupamentos, no fluxograma da Figura 11 os passos que compõem a etapa "analisa vizinhos", da Figura 9, são ilustrados. Na fase de agrupamento, propriamente dita, a comparação é realizada objeto a objeto. Caso o objeto esteja disponível – marcado como livre – ou pertença ao mesmo *cluster*, é verificado se o mesmo encontra-se na região determinada pelo Eps utilizado como parâmetro da iteração do DBSCAN modificado, que é executado iterativamente para cada valor de Eps definido. É importante destacar que essa classificação consiste no seguinte: em torno do objeto selecionado como principal e a partir do qual será iniciada a análise dos vizinhos, são marcados como alcançáveis os objetos que estiverem num raio Eps de distância e, para cada um deles, a análise do vizinho é realizada da mesma maneira.

A partir dessa premissa – objetos similares, livres ou do mesmo *cluster* e localizadas geograficamente num raio Eps do objeto principal – realiza-se então a verificação da quantidade de objetos interna à região determinada por Eps em torno do tomado como principal. Caso essa quantidade atenda ao número mínimo de objetos para formação de um agrupamento, o principal é marcado como centro e os vizinhos como alcançáveis, disparando então o passo "analisa vizinhos" para cada dos alcançáveis. Caso contrário, ou seja, não atingido o número mínimo de objetos num *cluster*, os alcançáveis são marcados como borda e não é realizada a analise dos seus vizinhos.

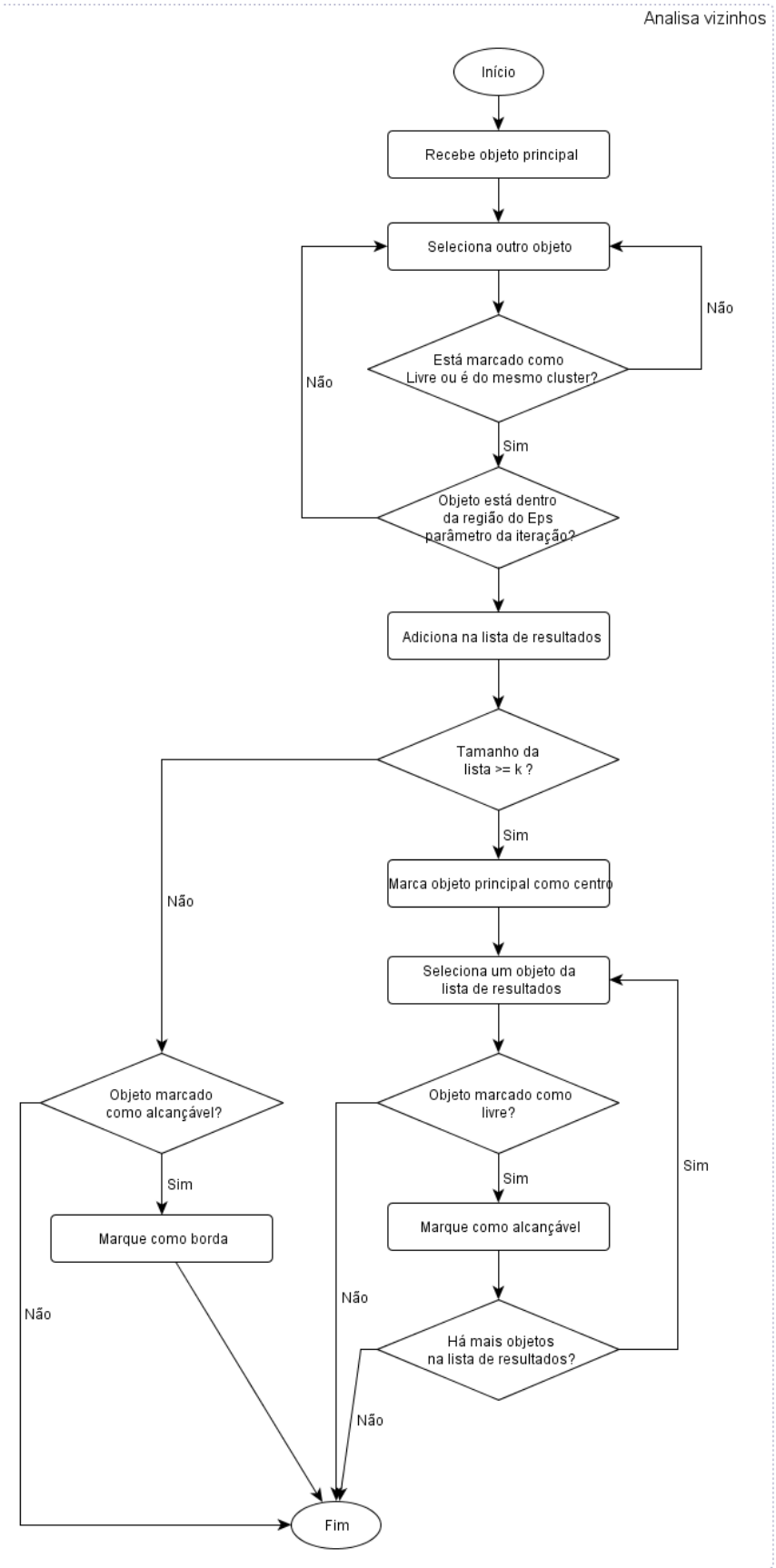


Figura 11 – Detalhamento do passo “analisa vizinhos” do DBSCAN modificado

Após o término da análise dos vizinhos e a extinção dos objetos identificados como alcançáveis – ou seja, todos os inicialmente alcançáveis agora marcados como borda ou centro – inicia-se então a formação do próximo *cluster*, até que todos sejam analisados, para então iniciar o mesmo procedimento com os demais valores de Eps.

Na segunda iteração do DBSCAN modificado, com o segundo valor de Eps, os objetos analisados são os que ainda não foram agrupados. E isso se repete até o último valor de Eps definido pelas etapas de inicialização e refinamento.

3.5 Otimização por multithreading

Além de promover melhor qualidade de resultados, foi também objetivo do trabalho apresentar otimização de desempenho em termos de tempo de execução do algoritmo, por meio do uso de *multithreading*.

Essa técnica foi aplicada nas três etapas do MR-Clustering, com a finalidade de contribuir para o desempenho no cálculo das distâncias entre objetos – etapa de inicialização; na execução do DBSCAN na etapa intermediária – etapa de refinamento; e na execução do DBSCAN modificado na última etapa – agrupamento.

Na etapa de inicialização, a operação executada por meio de uso dos núcleos de processamento disponíveis e que otimizam o tempo de execução é detalhada no fluxograma da Figura 12.

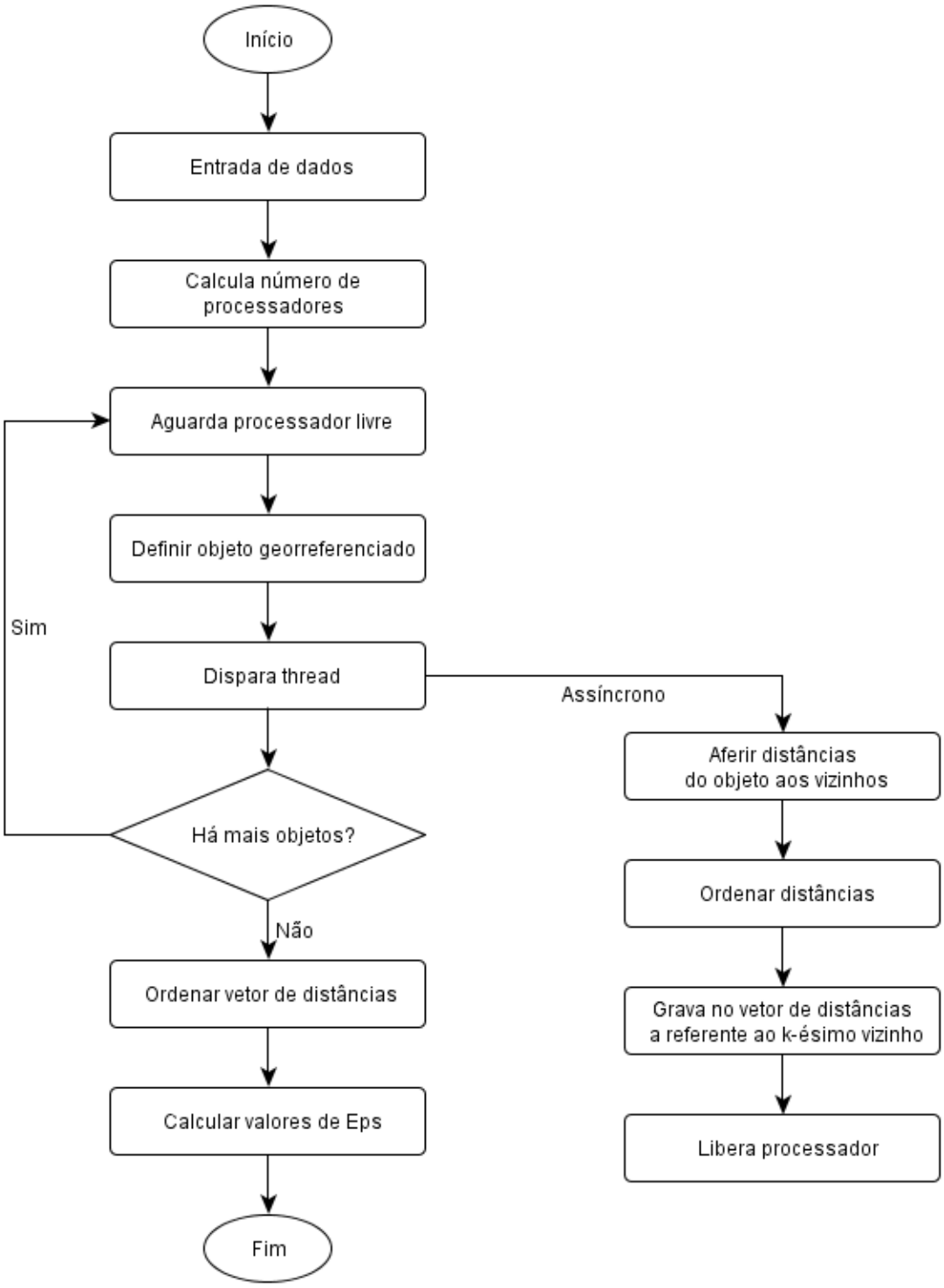


Figura 12 – Fluxograma da etapa de inicialização com uso de *multithreading* para otimização

Observa-se que é realizada uma verificação dos núcleos disponíveis para alocar a um deles a tarefa de aferição das distâncias de um objeto de referência em relação aos demais, no qual é disparada uma *thread* de forma assíncrona à alocação da mesma tarefa para os outros objetos nos outros núcleos de processamento livres. Cada *thread* é responsável então por aferir as distâncias do objeto de referência até os demais, ordenar os referidos valores e inseri-los no k-dist set.

Nas etapas de refinamento e agrupamento, conforme ilustrado na Figura 13, no momento em que as análises dos vizinhos para cada objeto alcançável são iniciadas, o processo é paralelizado e, portanto, executado por *threads* em processadores disponíveis. A alteração do procedimento sequencial foi realizada somente na distribuição dos objetos alcançáveis, o que significa que os passos da análise de vizinhos não foram alteradas da etapa de agrupamento, sendo portanto o mesmo ilustrado na Figura 11.

Diferentemente do “analisa vizinhos” da etapa de agrupamento, o da etapa de refinamento não considera a similaridade de características não espaciais dos objetos, sendo apenas levada em conta a densidade dos mesmos na distribuição espacial. O detalhamento dos passos então não inclui o passo “verifica se a porcentagem de atributos iguais atendem ao requisitado”, conforme ilustrado na Figura 14.

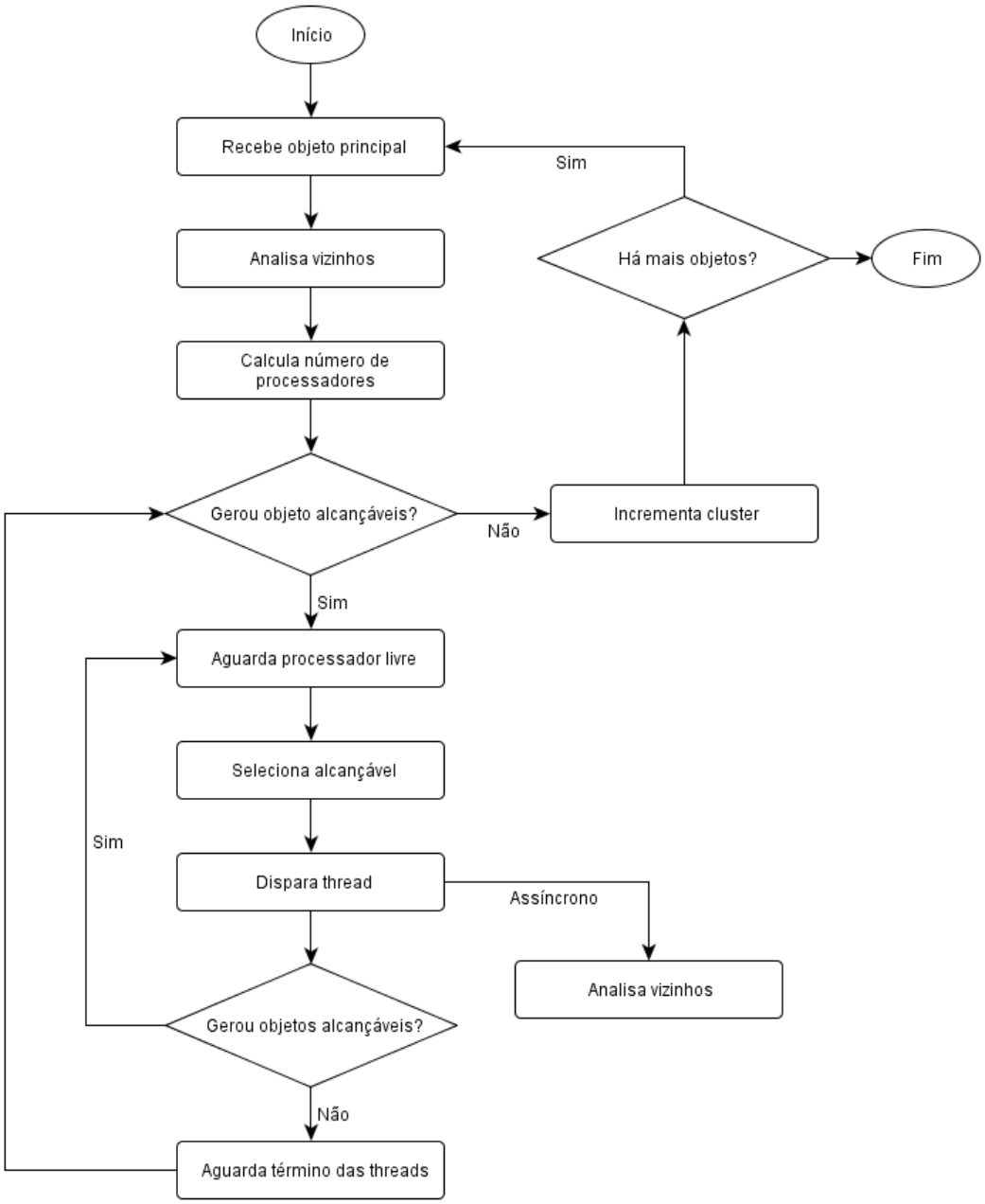


Figura 13 – Fluxograma da etapa de refinamento e agrupamento com uso de *multithreading* para otimização

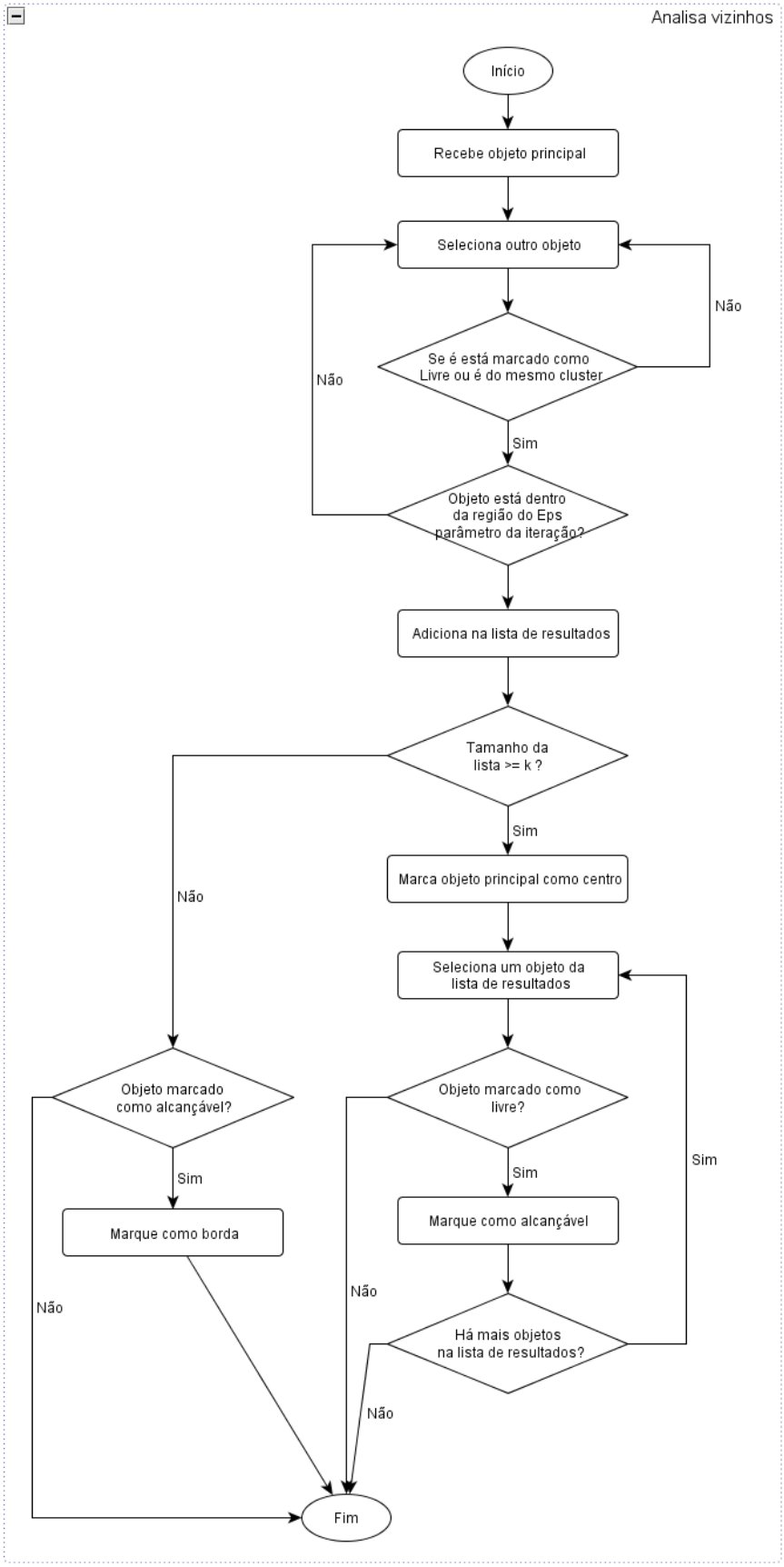


Figura 14 - Detalhamento do passo “analisa vizinhos” da etapa de refinamento

3.6 Considerações finais

Neste capítulo foram apresentados os detalhes de implementação do algoritmo criado. Todas as decisões com relação ao uso e criação das técnicas foram descritas, assim como os procedimentos que cada uma das etapas realiza. Nota-se que o algoritmo leva em consideração a análise de densidade dos objetos georreferenciados, por isso, considera-se que o algoritmo tem as raízes nos algoritmos DBSCAN e VDBSCAN, para cálculo de proximidade geográfica, e no CLARANS, por conta da análise das características não espaciais. No entanto, o mesmo diferencia-se das origens e dos demais algoritmos do estado da arte ao implementar: uma abordagem multirrelacional para o conjunto de dados a ser analisado; o agrupamento por similaridade de características não espaciais com abertura para possibilidade de agregação semântica nessa classificação, por conta da maneira como as características dos objetos georreferenciados são dispostas e analisadas; a otimização por *multithreading* nas três etapas que o compõem; o refinamento dos parâmetros definidos dinamicamente, sem a necessidade de entrada de dados para disparar o algoritmo e com melhora na qualidade dos resultados apresentados; dentre outros.

No capítulo 4, são descritas a execução do algoritmo e a análise dos resultados obtidos pelas técnicas implementadas.

Capítulo 4 Experimentos e resultados

4.1 Considerações iniciais

Neste capítulo são apresentados os experimentos executados com o algoritmo proposto, MR-Clustering, a fim de demonstrar a eficiência da aplicação do mesmo em contexto multirrelacional com a garantia da qualidade e otimização dos resultados obtidos e com tempo de execução inferior a alguns trabalhos da literatura, que não contemplam as características que o diferenciam positivamente.

Para isso, escolheu-se uma base de dados de acidentes do trabalho registrados em uma região do interior de São Paulo com mais de 100 municípios. A alimentação desses dados é efetivada por meio do Sistema de Informação e Vigilância de Acidentes de Trabalho – SIVAT, sendo que as notificações armazenadas ultrapassam 70 mil registros – destes, mais de 17 mil são georreferenciados – o que possibilita a aplicação do algoritmo proposto e análise das informações resultantes.

Na Figura 15, uma parte do esquema da base do SIVAT é ilustrada. Observa-se que a tabela “ficha” faz referências a outras tabelas para as características monovaloradas – “ocupacao” e “maquina_causadora”. As características multivaloradas são representadas por tabelas específicas – “ficha_parte_corpo” e “ficha_cid_10” – que fazem referência tanto à “ficha”, quanto às tabelas das características – “parte_corpo” e “cid_10”. Nota-se que um dos dados armazenados na tabela “ficha” é a coordenada geográfica – atributo “geom”, para cada notificação registrada.

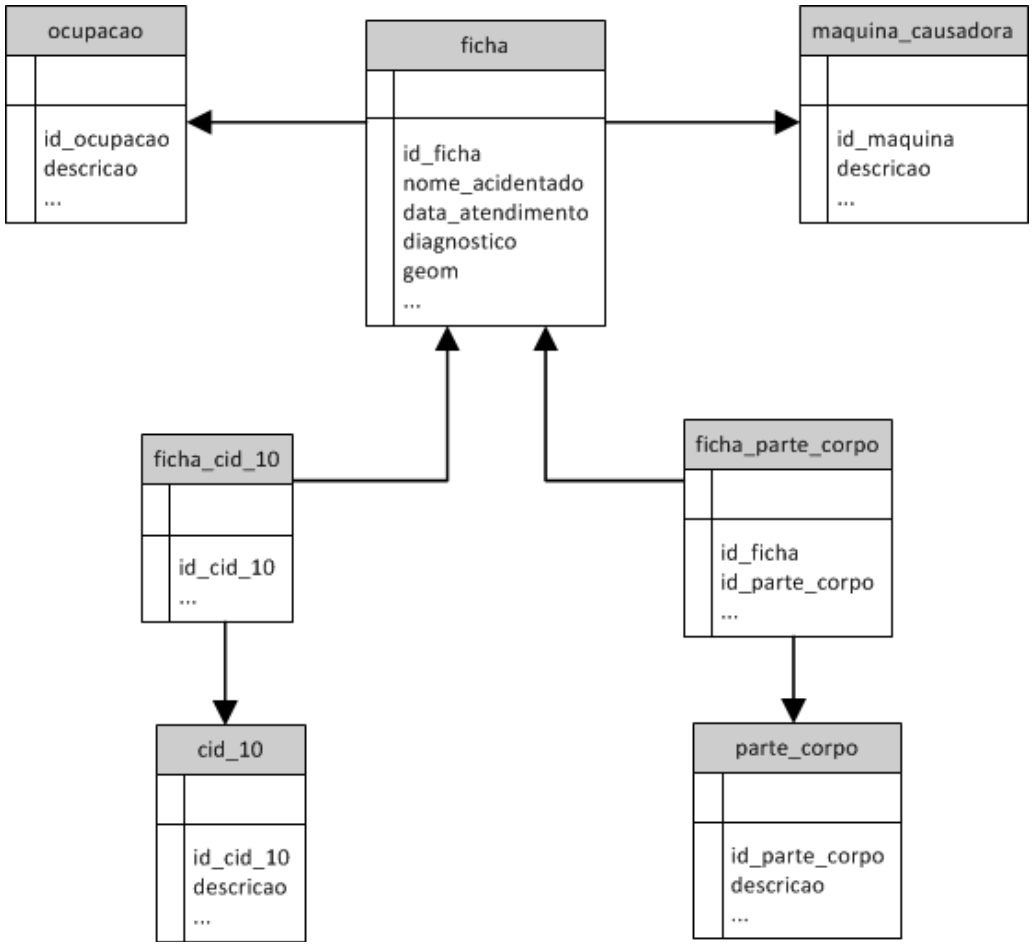


Figura 15 – Parte do esquema da base de dados SIVAT

4.2 Aplicação do MR-Clustering

No primeiro experimento realizado, objetivou-se comprovar a eficiência do algoritmo ao analisar um conjunto de dados georreferenciados com o retorno de agrupamento de objetos que apresentem determinado nível de similaridade e que, ao serem comparados com objetos pertencentes a outros agrupamentos, não apresentem a mesma característica.

O MR-Clustering foi então aplicado na base de dados SIVAT, num conjunto de mais de 17 mil registros georreferenciados, com refinamento dos valores de Eps. A tabela alvo escolhida foi a “ficha”, que conforme já anunciado, armazena todos os dados dos acidentes de trabalho, faz referências às demais tabelas que compõem a base SIVAT e é referenciada por elas, o que exige o processo de junção de dados na execução da prospecção, caso o algoritmo não ofereça suporte ao contexto multirrelacional. Nessa relação alvo, as características utilizadas para análise dos dados foram o tipo de acidente (“tipo_acidente”) e o dia da semana em que ocorreram os acidentes. Foi também escolhida

a relação de partes do corpo afetadas com os acidentes que, da maneira como é modelada a base de dados, encontram-se armazenadas na tabela “ficha_parte_corpo”, que é composta pelos atributos “id_ficha” e “id_parte_corpo”, sendo o primeiro chave estrangeira da tabela “ficha” e o último, a chave estrangeira para a tabela que armazena as partes do corpo afetadas numa ocorrência de acidente.

Na Figura 16 são exibidos os agrupamentos mais relevantes e resultantes da primeira aplicação do MR-Clustering no domínio de objetos georreferenciados descrito. Para entendimento da representação desses resultados, é importante destacar:

- O mapa sob as marcações está focado na cidade de São José do Rio Preto, estado de São Paulo;
- Os quadrados vermelhos representam uma localização geográfica, ou seja, nesse caso da base SIVAT, esses referem-se a um local em que houve um ou mais acidentes do trabalho;
- Os polígonos de bordas pretas e preenchimento verde delimitam uma região em que foi encontrado um agrupamento de objetos similares;
- Por fim, os quadrados azuis representam os objetos que pertencem ao agrupamento delimitado pelo polígono verde correspondente. No caso analisado, pelos números apresentados na Tabela 1, é possível concluir que mais de um acidente do trabalho pode ter ocorrido numa mesma localização geográfica, já que o número de quadrados azuis nos agrupamentos é menor do que o número de acidentes pertencentes ao agrupamento.

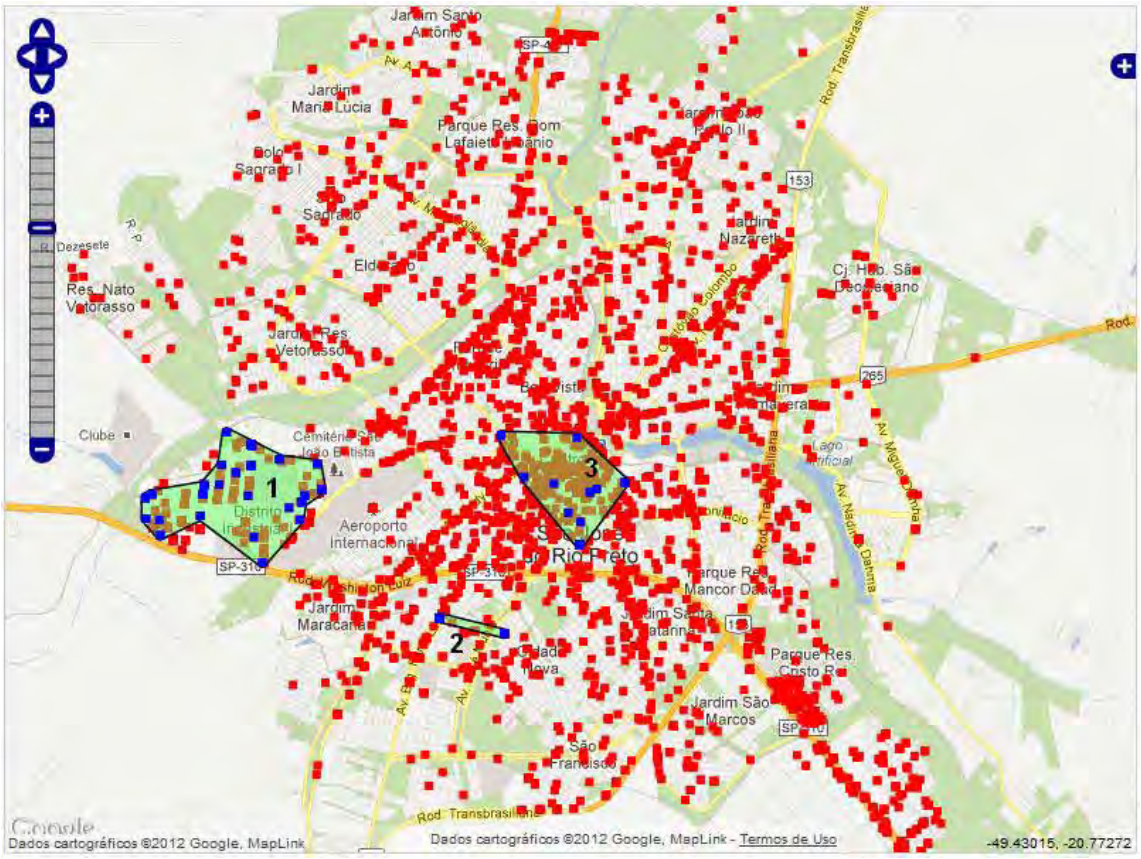


Figura 16 – Experimento 1: aplicação do MR-Clustering para validação

Os números que sumarizam as informações obtidas nesse primeiro experimento e representadas na Figura 16 são os que seguem na Tabela 1.

Tabela 1 – Resultados do primeiro experimento realizado

Agrupamento	*No. de notificações*	*Tipo de acidente*	*Dia da semana*	*Parte do corpo afetada*
1	138	Típico	Quinta-feira	Olho
2	13	Típico	Domingo	Membro superior
3	17	Típico	Domingo	Membro superior

Para melhor entendimento das informações, algumas considerações devem ser feitas:

- Acidente típico refere-se àqueles que ocorreram durante o expediente de serviço e que não são os ocasionados por doença do trabalho, nem que tenha ocorrido durante o percurso de ida ou volta ao local de trabalho;
- Uma vez que uma notificação representada no mapa tem alto grau de similaridade em relação às demais pertencentes ao mesmo agrupamento e,

considerando que um acidente pode afetar uma ou mais partes do corpo – situação que exige o processo de junção de dados numa prospecção tradicional ou uma abordagem multirrelacional pelo algoritmo que se aplica – é importante destacar que o MR-Clustering agrupou as ocorrências de acidentes notificadas com parte do corpo afetadas em comum, devido à comparação característica a característica para verificar se o grau de similaridade atende ao requisitado. No caso do agrupamento 1, todos tiveram "membro inferior" como parte do corpo afetada, o que não exclui a possibilidade de haver acidentes que afetaram outras partes do corpo além dessa.

Pelo que pode ser observado na Figura 16 e analisado na Tabela 1, de imediato é possível concluir que diversas ocorrências de acidentes do trabalho se concentraram em pequenas regiões do perímetro urbano de São José do Rio Preto e com alto grau de similaridade entre eles, o que possibilitou que tais agrupamentos fossem identificados por conta dos atributos que os rotulam e da proximidade geográfica das ocorrências similares. Destaca-se que a aplicação de prospecção de dados convencionais dificilmente identificaria os referidos agrupamentos, uma vez que o elemento que permitiu tal resultado foi a relevância da localidade na análise dos dados.

Para comprovação da eficácia da abordagem multirrelacional, um levantamento das partes do corpo afetadas nos acidentes, bem como a quantificação de cada uma delas nos agrupamento 1 foi realizado e disposta na Tabela 2.

Tabela 2 – Partes do corpo afetadas no agrupamento 1

Agrupamento	***Parte do corpo afetada***	***Quantidade***
1	Olho	138
	Cabeça	2
	Membro superior	1
	Mão	1

Observa-se que, de fato, todos os 138 acidentes que compõem o agrupamento 1 afetaram o olho dos acidentados. Entretanto, graças à abordagem multirrelacional, constatou-se que além do olho ter sido afetado, alguns acidentes afetaram também outras partes do corpo – cabeça, membro superior e mão. Numa abordagem convencional, uma tabela de junção dos dados dos acidentes com as partes do corpo deveria ser criada para

que se pudesse obter a informação apresentada, o que acarretaria em redundâncias e problemas semânticos que comprometeriam os resultados.

Uma consideração complementar que pode ser feita em relação à relevância do resultado é que o padrão geral da base de dados referente à parte do corpo afetada nos acidentes é a mão e, sendo assim, o fato da prospecção de dados considerar a localização geográfica das ocorrências fez com que os 138 acidentes que afetaram o olho fossem levantados como um agrupamento relevante, uma vez que ocorreram em locais próximos.

Outra observação importante a se fazer é que os 138 casos de olhos afetados nos acidentes do agrupamento 1 ocorreram no Distrito Industrial da cidade, o que pode justificar o elevado número de ocorrências similares e próximas geograficamente e constituir um indicativo para vigilância, executada por setores públicos responsáveis pela garantia da saúde do trabalhador.

Para enriquecer a demonstração da prospecção de dados multirrelacional, na Tabela 3 a quantidade acidentes para cada parte do corpo afetada no agrupamento 3 é sumarizada.

Tabela 3 - Partes do corpo afetadas no agrupamento 3

Agrupamento	***Parte do corpo afetada***	***Quantidade***
3	Membro superior	17
	Cabeça	4
	Pescoço	1
	Tórax	1

Pela análise do agrupamento 3, é possível concluir que todos os acidentes que o formam afetaram algum membro superior dos acidentados, entretanto, houve cabeças, pescoço e tórax afetados em alguns dos acidentes, além de membro superior.

Outro exemplo de aplicação do MR-Clustering foi com a análise das características monovaloradas representadas nos atributos “tipo_acidente” e “maquina_causadora” – que armazenam a descrição da causa ou do objeto causador do acidente – e a característica multivalorada referente a partes do corpo afetadas no acidente – a fim de fazer uso da seleção de dados de diferentes tabelas sem a necessidade de junção. Na Figura 17, os principais agrupamentos retornados são exibidos e, na Tabela 4, um resumo das informações para análise é apresentado.

Figura 17 – Exemplo de aplicação do MR-Clustering considerando o tipo de acidente, causa/causador do acidente e partes do corpo do acidentado afetadas

Tabela 4 – Sumarização dos resultados obtidos a partir dos agrupamentos retornados pelo segundo experimento de aplicação do MR-Clustering

Agrupamento	***No. de notificações***	***Tipo de acidente***	***Causador do acidente***	***Parte do corpo afetada***
1	56	Típico	Fagulha	Olho
2	17	Típico	Agressão física	Membro superior
3	37	Típico	Agulha	Olho
4	41	Típico	Sangue	Olho
5	17	Típico	Chapa	Membro superior

Pelos resultados obtidos, de acordo com o agrupamento 1, em 100% das ocorrências agrupadas os acidentes foram causados por fagulha, sendo os acidentados notificados com o olho afetado. Assim como no agrupamento 1 do experimento anterior, que teve a mesma parte do corpo afetada nos 138 acidentes, as ocorrências foram registradas no Distrito Industrial da cidade, portanto, um indicativo para vigilância pode se

resumir à alta concentração de acidentes nessa região por fagulha, afetando o olho do trabalhador e às quintas-feiras.

Outro resultado obtido devido à relevância da localização geográfica na prospecção dos dados é o agrupamento 3, uma vez que na base SIVAT completa foram notificados 1.601 acidentes com agulha. Considerando que apenas 23,5% da base toda é georreferenciada e que somente 881 notificações desse conjunto ocorreram com agulha, os 37 acidentes agrupados no *cluster* 3 correspondem a 4% das notificações com agulha da base completa – pouco relevante. No entanto, a atenção para esse caso aumenta quando visualmente no mapa todos ocorreram no mesmo lugar, o que leva a mais um indicativo para vigilância, já que se todas as 881 notificações com agulha fossem georreferenciadas, essa concentração poderia ser ainda maior.

Nesse sentido, o agrupamento 4 também merece atenção, já que na base toda foram registradas 174 notificações de acidentes do trabalho com sangue – ou seja, profissionais que tiveram contato com sangue – sendo que desses, 97 são georreferenciados e foram analisados na prospecção desse experimento. Dessa forma, conclui-se que o agrupamento 4 representa 23,5% de todos os acidentes ocorridos com sangue e, se trazido esse número para somente o universo dos acidentes georreferenciados, esse percentual cresce para 42,2%, bastante expressivo.

Para finalizar, na Tabela 5 as partes do corpo afetadas no agrupamento 4 são quantificadas, sendo elas recuperadas pela abordagem multirrelacional sem a necessidade do processo de junção de dados.

Tabela 5 – Partes do corpo afetadas nos acidentes notificados do agrupamento 4

Agrupamento	***Parte do corpo afetada***	***Quantidade***
4	Olho	41
	Cabeça	3
	Pescoço	1
	Tórax	1
	Mão	1

4.2.1 MR-Clustering sem refinamento

Uma abordagem para refinamento dos resultados foi concebida no MR-Clustering e, para isso, uma etapa foi criada e nomeada de refinamento, conforme já detalhado anteriormente no Capítulo 3. Nesse tópico e no 4.2.2, a comprovação da eficácia dessa abordagem é demonstrada com experimentos.

Inicialmente, os resultados obtidos com o uso do MR-Clustering sem a etapa de refinamento são descritos. O conjunto de dados desse experimento continha a ocupação do acidentado e o(s) CID-10 (Classificação Internacional de Doenças) do diagnóstico realizado nas unidades de atendimento. Nesse caso, conforme pode ser visualizado na Figura 15, uma notificação pode ter mais de um CID-10 relacionado, assim como um CID-10 armazenado na base de dados SIVAT pode ser relacionado a mais de uma notificação, portanto o uso da abordagem multirrelacional mais uma vez se faz necessário.

Na Figura 18 e na Figura 19, o resultado retornado pelo algoritmo a partir da análise do referido conjunto de dados é apresentado. Observa-se na Figura 18 que algumas notificações se localizam em regiões mais afastadas da grande concentração ao centro do mapa. Além disso, algumas representações estão em vermelho, sendo essas as que não foram agrupadas pelo algoritmo, diferentemente dos azuis, que fazem parte de algum dos agrupamentos representados em verde. Considerando isso, é importante notar que, nesse exemplo, alguns agrupamentos extrapolaram o elemento proximidade entre as ocorrências agrupadas e buscaram objetos bastante distantes para si, como no caso do que considerou um que se encontra em outro município.

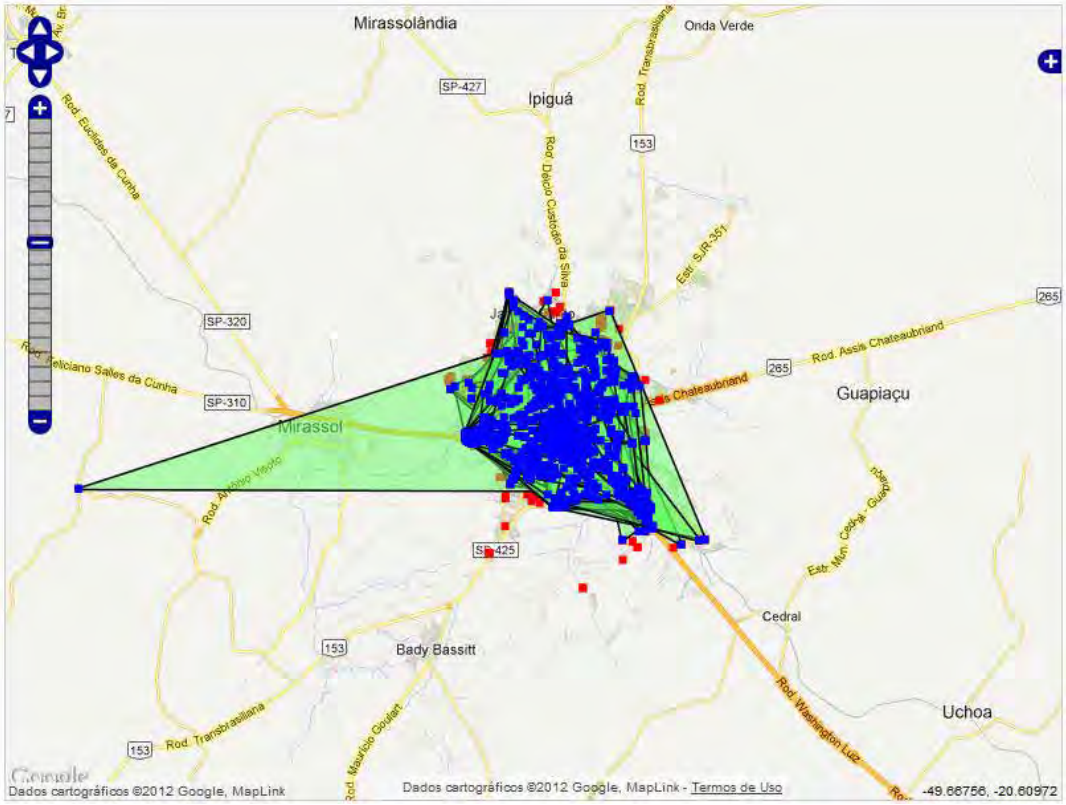


Figura 18 – Agrupamentos formados pelo MR-Clustering sem refinamento dos valores de Eps

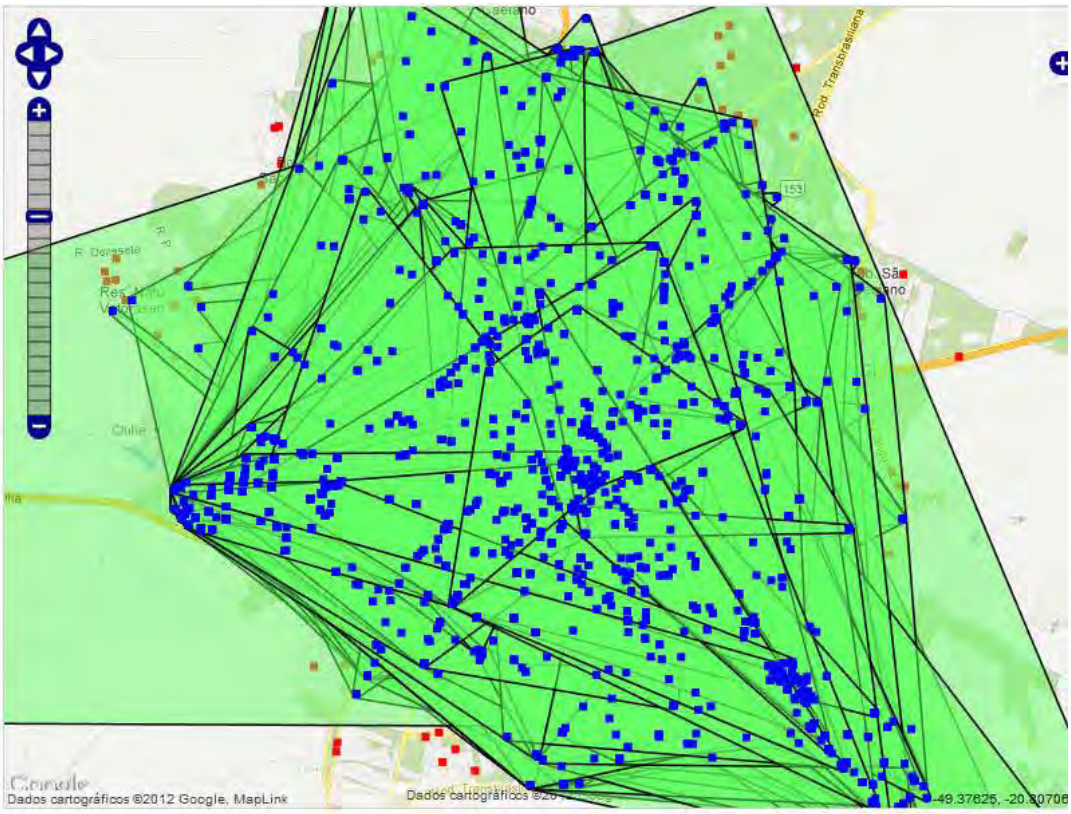


Figura 19 – Visão aproximada dos agrupamentos formados pelo MR-Clustering sem refinamento dos valores de Eps

A justificativa do refinamento dos valores de Eps definidos pela etapa de inicialização então é visualmente comprovada. A situação de agrupamentos grandes e que considerem objetos que deveriam ser classificados como ruído ou *noises* ocorre por conta dos valores de Eps não muito precisos, resultantes do cálculo realizado com os *noises* no vetor k-dist – Figura 7 – que causam distorção dos valores, uma vez que não representam uma mudança do nível de densidade do conjunto de objetos a serem analisados. Por conta disso, muitas vezes os valores de Eps acabam sendo superestimados e, consequentemente, afetam o resultado final da etapa de agrupamento – *clusters* grandes e pouco informativos.

4.2.2 MR-Clustering com refinamento

Nesse tópico, são descritas as considerações pertinentes à aplicação do MR-Clustering com a etapa de refinamento ativa, sendo que ao se comparar com o que foi apresentado no tópico 4.2.1, é possível verificar que houve uma determinada otimização na qualidade dos resultados obtidos. Isso pode ser comprovado pelas figuras Figura 20 e Figura 21.

De imediato é possível perceber que grande parte das notificações foi desconsiderada dos agrupamentos devido à dissimilaridade que apresentam em relação aos objetos agrupados, principalmente aquelas que se localizam a uma distância maior da região mais densa de objetos – *noises*. Isso é benéfico para a qualidade da prospecção de dados espaciais, uma vez que a localização é um elemento relevante e o agrupamento de um objeto muito distante não representa similaridade geográfica.

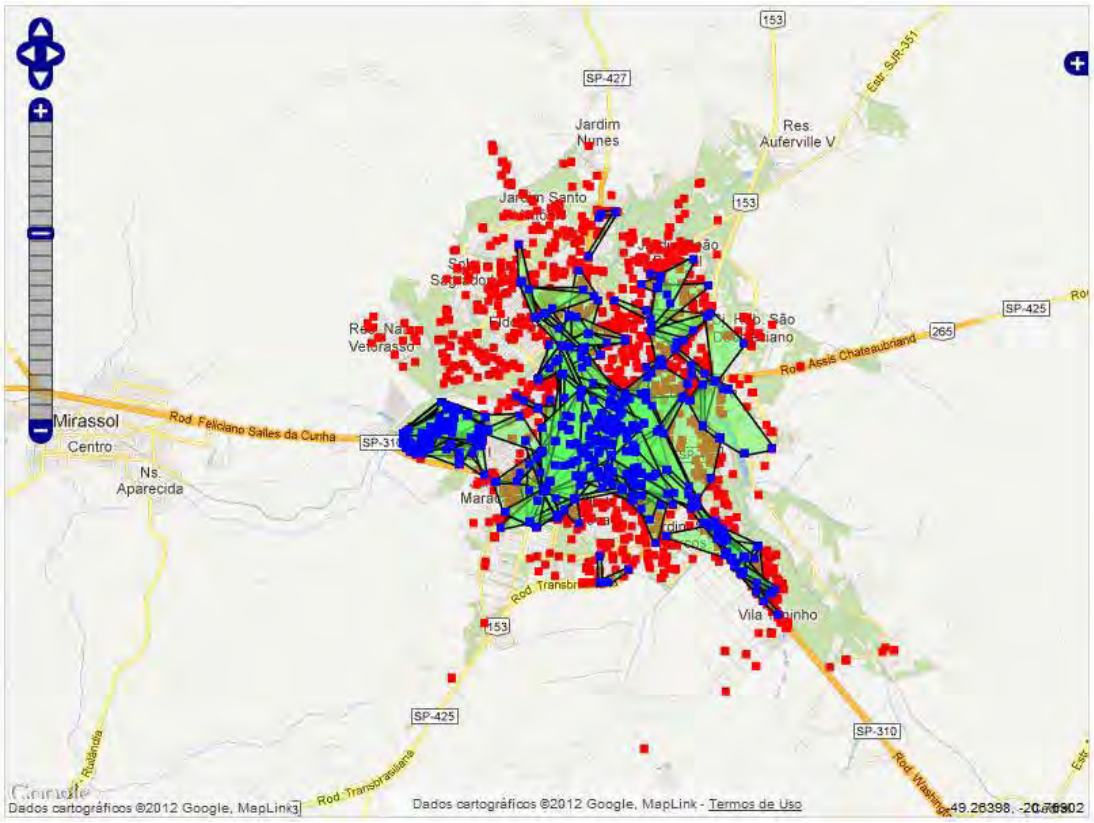


Figura 20 – Agrupamentos formados pelo MR-Clustering com refinamento dos valores de Eps

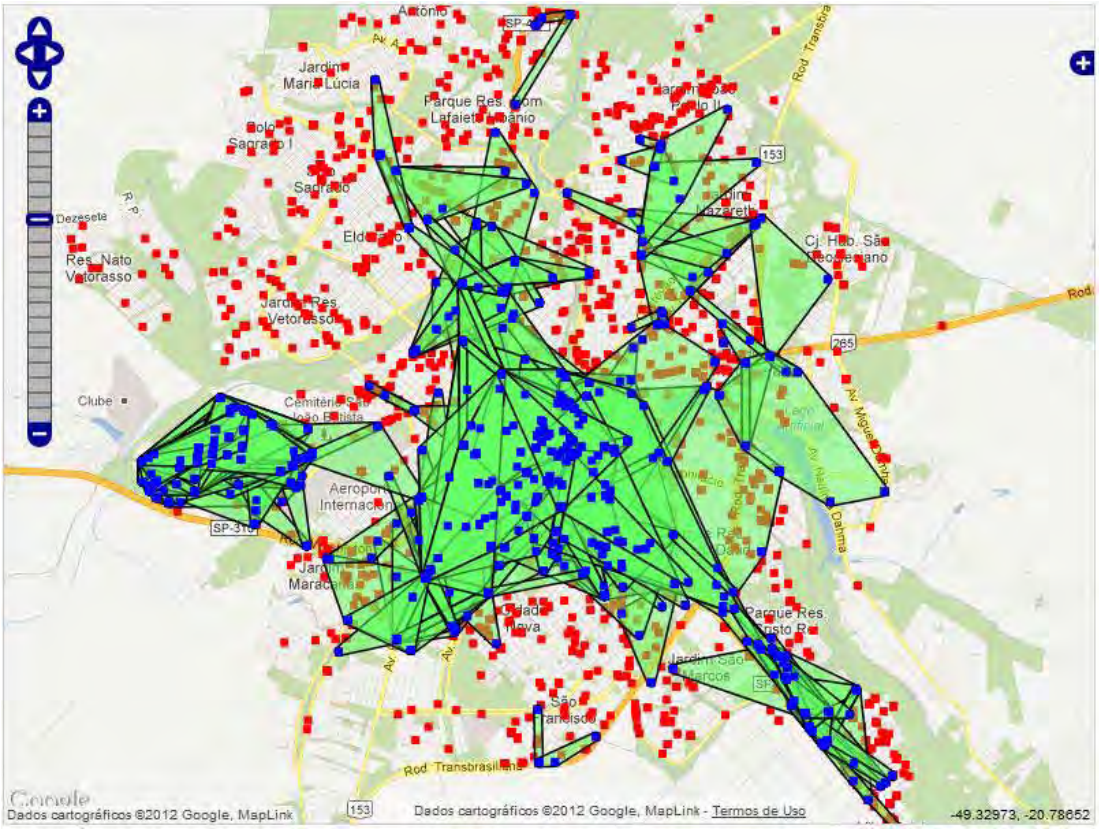


Figura 21 – Visão aproximada dos agrupamentos formados pelo MR-Clustering com refinamento dos valores de Eps

Outra observação que deve ser feita é o particionamento de agrupamentos grandes, que sem o refinamento dos valores de Eps envolveram objetos bastante distantes geograficamente, embora similares em termos de características não espaciais. Esse particionamento é justificado pela desconsideração dos *noises* no cálculo dos valores de Eps a partir das distâncias entre os objetos. Com isso, os Eps refinados representam fielmente a mudança de densidade no conjunto de objetos a serem analisados e, portanto, o algoritmo se torna mais ajustável às variações de densidade e à formação de agrupamentos melhores representativos.

4.3 Algoritmo tradicional *versus* multirrelacional

A segunda etapa de experimentos consiste na comparação do MR-Clustering com a abordagem tradicional de prospecção de dados espaciais. Para isso, o algoritmo CLARANS foi executado a fim de confrontá-lo com o MR-Clustering, uma vez que, além de apresentar um propósito de agrupamento de objetos geográficos similares semelhante ao do algoritmo proposto, o CLARANS é um dos mais tradicionais.

Tanto o CLARANS, quanto o MR-Clustering foram executados também sobre a base de dados SIVAT a fim de analisar os dados correspondentes à ocupação e CID-10 dos acidentados, da mesma maneira dos experimentos dos tópico 4.2.1 e 4.2.2. Sendo assim, os resultados apresentados pelo MR-Clustering foram os mesmos da Figura 20 e Figura 21. No total, foram gerados 520 *clusters* e, por conta da capacidade em absorver o contexto multirrelacional, o algoritmo possibilitou que a semântica da base de dados fosse mantida, além de não necessitar do processo de junção e preparação dos dados.

O CLARANS, por sua vez, foi executado para encontrar 520 agrupamentos, já que requisita essa definição como parâmetro de entrada. No entanto, devido ao elevado número de objetos georreferenciados, agrupamentos, número de interações realizadas e à ineficiência do algoritmo em aplicações a grandes repositórios de dados, o mesmo não viabilizou a aplicação da prospecção com essa configuração, então a entrada para o número de agrupamentos a ser retornado foi reduzida para 50 e 100.

Além disso, o CLARANS exigiu que os dados de interesse para análise fossem dispostos numa única relação, o que a princípio já acarretou na identificação de infidelidade com a realidade que, posteriormente, influenciou nos resultados. Isso devido ao fato das notificações georreferenciadas terem sido replicadas nos casos de relação a mais de um CID-10, o que causa algumas distorções tanto na formação dos agrupamentos, quanto na validação dos resultados. Na Figura 22 é possível observar a redundância de

dados de uma mesma notificação (ficha) no caso citado. Além de custo computacional e recursos de memória desperdiçados, o atributo que armazena a representação das coordenadas geográficas, nomeado “geom”, também é replicado, o que gera redundância ao serem analisados e provocam a formação de agrupamentos distorcidos da realidade.

	id_ficha integer	id_ocupacao integer	id_cid integer	geom geometry
1	17	2088	8787	0101000020E6100000B1975C7B55B048C04F7D7B325AD034C0
2	17	2088	9126	0101000020E6100000B1975C7B55B048C04F7D7B325AD034C0
3	17	2088	13487	0101000020E6100000B1975C7B55B048C04F7D7B325AD034C0
4	56	496	9143	0101000020E61000009AF4ADC55DAD48C09065C1C41FD134C0
5	25	2241	9045	0101000020E61000008E33976599B048C0F3CCCB61F7D134C0
6	25	2241	10923	0101000020E61000008E33976599B048C0F3CCCB61F7D134C0
7	56	496	9051	0101000020E61000009AF4ADC55DAD48C09065C1C41FD134C0
8	56	496	10202	0101000020E61000009AF4ADC55DAD48C09065C1C41FD134C0
9	62	2088	8722	0101000020E610000080C63B7606B148C033F55F8CEDD234C0
10	62	2088	9143	0101000020E610000080C63B7606B148C033F55F8CEDD234C0
11	62	2088	10923	0101000020E610000080C63B7606B148C033F55F8CEDD234C0
12	65	1663	8791	0101000020E6100000B83A00E2AEB448C0068AB37D23D334C0
13	65	1663	10978	0101000020E6100000B83A00E2AEB448C0068AB37D23D334C0
14	44	61	8786	0101000020E610000049827005 14AD48C013DB38BDE6D034C0
15	44	61	9142	0101000020E610000049827005 14AD48C013DB38BDE6D034C0
16	44	61	10956	0101000020E610000049827005 14AD48C013DB38BDE6D034C0

Figura 22 – Parte dos dados numa tabela de junção para aplicação do CLARANS

Uma vez que uma notificação está relacionada a quatro CID-10 diferentes, por exemplo, a representação espacial contemplará quatro notificações idênticas com apenas CID-10 diferentes, por conta da leitura de quatro tuplas na relação alvo do CLARANS, quando somente uma deveria ser contemplada. Dessa forma, a realidade fica comprometida e, além disso, diversos agrupamentos podem ser gerados por conta da concentração redundante de objetos georreferenciados em determinadas regiões, o que compromete a formação dos agrupamentos e qualidade dos resultados.

Nas Figuras 23 e 24, os resultados apresentados pela aplicação do CLARANS com 50 e 100 agrupamentos são ilustrados, respectivamente.

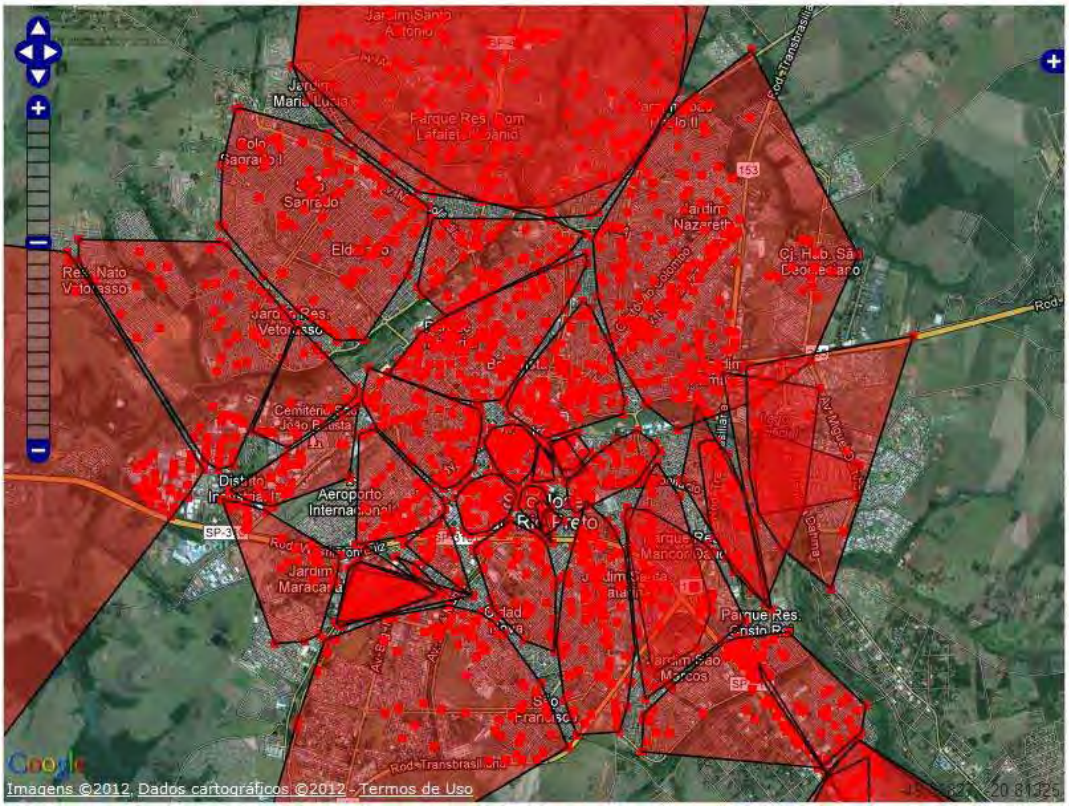


Figura 23 – Cinquenta agrupamentos retornados pelo CLARANS a partir da prospecção do mesmo conjunto de notificações analisado pelo MR-Clustering

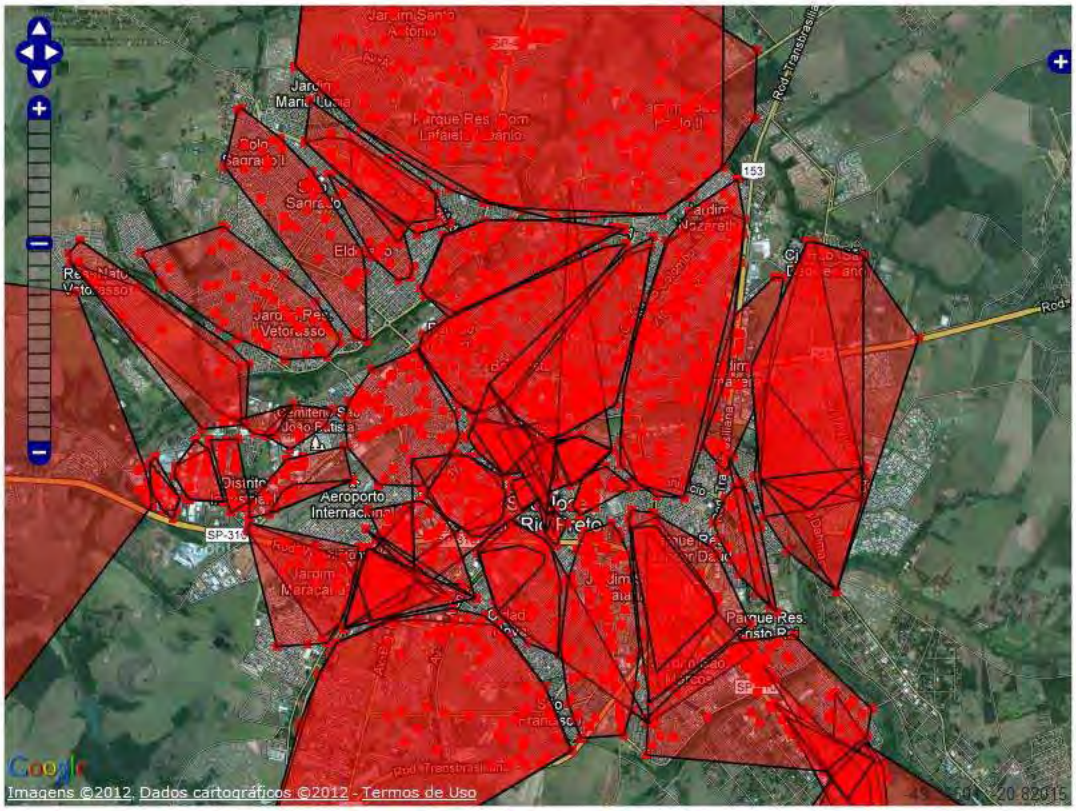


Figura 24 - Cem agrupamentos retornados pelo CLARANS a partir da prospecção do mesmo conjunto de notificações analisado pelo MR-Clustering

Para comprovar que o processo de junção prejudicou os resultados retornados pelo algoritmo, na Figura 25, é exibido um trecho da tabela de resultados obtidos pela execução do CLARANS. Nota-se que a replicação de objetos georreferenciados num mesmo agrupamento é recorrente em várias delas, o que confirma a perda semântica nos resultados e, o que é mais grave, muitas vezes agrega semântica de maneira errônea – como no caso demonstrado.

	oid	cluster integer	the_geometry geometry	id_ficha double precis	atributo0 double precis	atributo1 double precis
5917	1399111	7	[illegible]	[illegible]	58	12177
5918	1399112	7	0101000020E6100000FECBFFF51B8448C05706D50627C034C0	9554	2088	5260
5919	1399113	7	0101000020E6100000FECBFFF51B8448C05706D50627C034C0	9554	2088	10967
5920	1399114	7	0101000020E61000004CDD3AA4BDB248C091BAF83658C934C0	9762	2500	8855
5921	1399115	7	0101000020E61000004CDD3AA4BDB248C091BAF83658C934C0	9762	2500	12177
5922	1399116	7	0101000020E6100000FDCE407628B448C02E32A605D4C634C0	10002	3537	8786
5923	1399117	7	0101000020E6100000FDCE407628B448C02E32A605D4C634C0	10002	3537	9062
5924	1399118	7	0101000020E6100000FDCE407628B448C02E32A605D4C634C0	10002	3537	10923
5925	1399119	7	0101000020E61000009C919CA795B248C08A5F0C40FEC334C0	10259	1883	9045
5926	1399120	7	0101000020E61000009C919CA795B248C08A5F0C40FEC334C0	10259	1883	10901
5927	1399121	7	0101000020E6100000DCBA9BA73AB148C0925CFE43FAC334C0	10436	2088	8796
5928	1399122	7	0101000020E6100000DCBA9BA73AB148C0925CFE43FAC334C0	10436	2088	10973
5929	1399123	7	0101000020E61000004CDD3AA4BDB248C091BAF83658C934C0	10584	2088	8787
5930	1399124	7	0101000020E61000004CDD3AA4BDB248C091BAF83658C934C0	10584	2088	10813
5931	1399125	7	0101000020E61000004CDD3AA4BDB248C091BAF83658C934C0	10934	2088	9154
5932	1399126	7	0101000020E61000004CDD3AA4BDB248C091BAF83658C934C0	10934	2088	11561
5933	1399127	7	0101000020E610000019DE510889B348C063F5A2D178C334C0	11192	2088	8779
5934	1399128	7	0101000020E610000019DE510889B348C063F5A2D178C334C0	11192	2088	13487
5935	1399129	7	0101000020E61000004CDD3AA4BDB248C091BAF83658C934C0	11280	2088	9153
5936	1399130	7	0101000020E61000004CDD3AA4BDB248C091BAF83658C934C0	11280	2088	11561
5937	1399131	7	0101000020E610000078B240608AB048C0A20106932CC334C0	11893	3470	8779
5938	1399132	7	0101000020E610000078B240608AB048C0A20106932CC334C0	11893	3470	10978
5939	1399133	7	0101000020E61000002B3D89BE71B048C05D56BCEC32C334C0	16281	2088	9157
5940	1399134	7	0101000020E61000002B3D89BE71B048C05D56BCEC32C334C0	16281	2088	11165
5941	1399135	7	0101000020E610000096A3117D8BB248C001A19A37F3C234C0	20564	134	8365

Figura 25 – Agrupamento 7 com redundância de notificações georreferenciadas resultante da análise sobre a relação de junção

Além da incapacidade em se tratar contexto multirrelacional com fidelidade dos resultados com a semântica da base de dados, os agrupamentos que se formam pelo CLARANS dependem diretamente dos parâmetros de entrada exigidos pelo algoritmo, o que remete a qualidade dos agrupamentos ao calibre manual das entradas. O comportamento do MR-Clustering, por sua vez, se diferencia do CLARANS por conta da independência de entradas manuais, ou seja, o algoritmo desempenha a formação dos agrupamentos sem influência do usuário e converge para o resultado ótimo.

Para comprovar essa deficiência do CLARANS, nos exemplos das Figuras 23 e 24, foram definidos como entrada:

- Número de agrupamentos: 50 e 100;
- Número máximo de interações: 50;
- Número mínimo de vizinhos: 10.

Observa-se que a definição de parâmetros faz com que o algoritmo induza a formação do número de agrupamentos de entrada e com a limitação das interações que faz ao comparar os objetos com os respectivos vizinhos. Dessa forma, os resultados apresentados muitas vezes se diferem dos que seriam obtidos se tais exigências não influenciassem diretamente no comportamento do algoritmo.

Diferentemente do CLARANS, os resultados apresentados pelo MR-Clustering demonstraram um nível de qualidade maior, uma vez que foram retornados agrupamentos de objetos similares e próximos geograficamente e, por isso, os agrupamentos retornados são menores e desconsideram os ruídos.

4.4 Desempenho do MR-Clustering

Uma vez confirmada a eficácia do MR-Clustering em se processar dados espaciais na busca de informações, é relevante demonstrar o desempenho do algoritmo em relação a tempo de execução.

No artigo de Vijayalakshmi e Punithavalli [VIJ_10], é descrito o trabalho realizado no sentido de otimizar o tempo demandado pelo VDBSCAN para execução da prospecção de dados espaciais. Os resultados são exibidos por meio de uma comparação de tempo para processamento de conjuntos de 200, 400 e 600 registros. Em todos os casos, as melhorias implementadas pelo trabalho se mostraram eficazes na otimização do desempenho do algoritmo.

Sendo assim, para analisar o desempenho do MR-Clustering em relação ao que tem sido proposto no estado da arte, foi feita uma comparação dos tempos requeridos para processamento dos conjuntos de registros entre o algoritmo proposto e o VDBSCAN melhorado. O resultado das aferições encontra-se representado no gráfico da Figura 26.

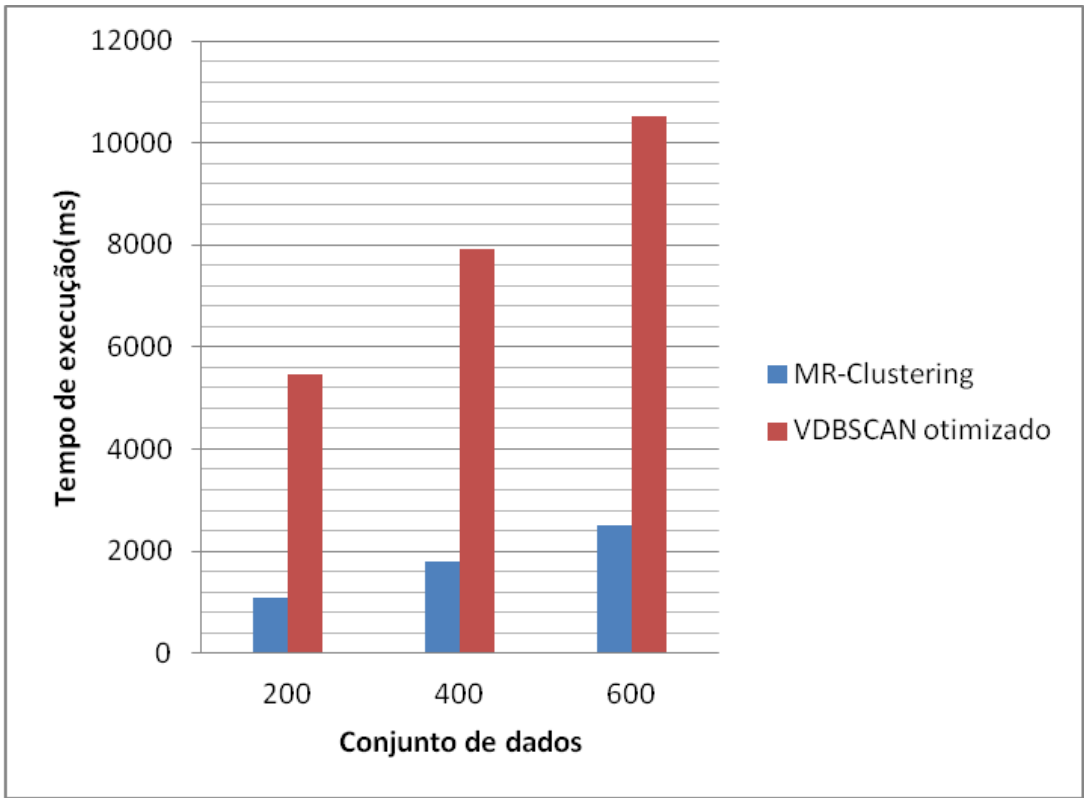


Figura 26 – Gráfico comparativo de tempo de execução exigido pelo MR-Clustering e VDBSCAN otimizado para processamento de conjuntos de 200, 400 e 600 registros

Outro experimento realizado com foco no desempenho do MR-Clustering foi a comparação do mesmo com o CLARANS, como já afirmado, bastante utilizado da categoria de agrupamento por particionamento. No exemplo do tópico 4.3, a aplicação do algoritmo levou 624 segundos, com a formação de menos *clusters*, enquanto que o MR-Clustering, 287 segundos – 54% a menos, sendo que pelo CLARANS houve limitação dos parâmetros imposta para suportar a aplicação nos mais de 17 mil registros do conjunto de dados, diferentemente do MR-Clustering.

Para finalizar, no último experimento realizado foram considerados todos os registros georreferenciados da base SIVAT, com análise dos atributos “ocupacao” e das características multivaloradas das notificações referentes às partes do corpo afetadas no acidente, com a finalidade de analisar o uso de *multithreading* no processamento dos registros. Para isso, foram realizadas cinco execuções do algoritmo com o mesmo conjunto de dados, a fim de aferir o tempo de execução requerido com 1, 2, 3, 4 e 5 *threads*, respectivamente.

No gráfico da Figura 27, o resultado desse experimento é contemplado e deve-se considerar que o processador utilizado é composto por quatro núcleos de processamento. O tempo de execução com apenas uma *thread* é bastante superior ao desprendido pelo teste

com uso de quatro *threads*, conforme pode ser observado no gráfico. Além disso, nota-se também que a execução do algoritmo com mais *threads* do que o total de núcleos de processamento tende a estabilizar o tempo de execução, como pode ser visto o teste com 5 *threads*.

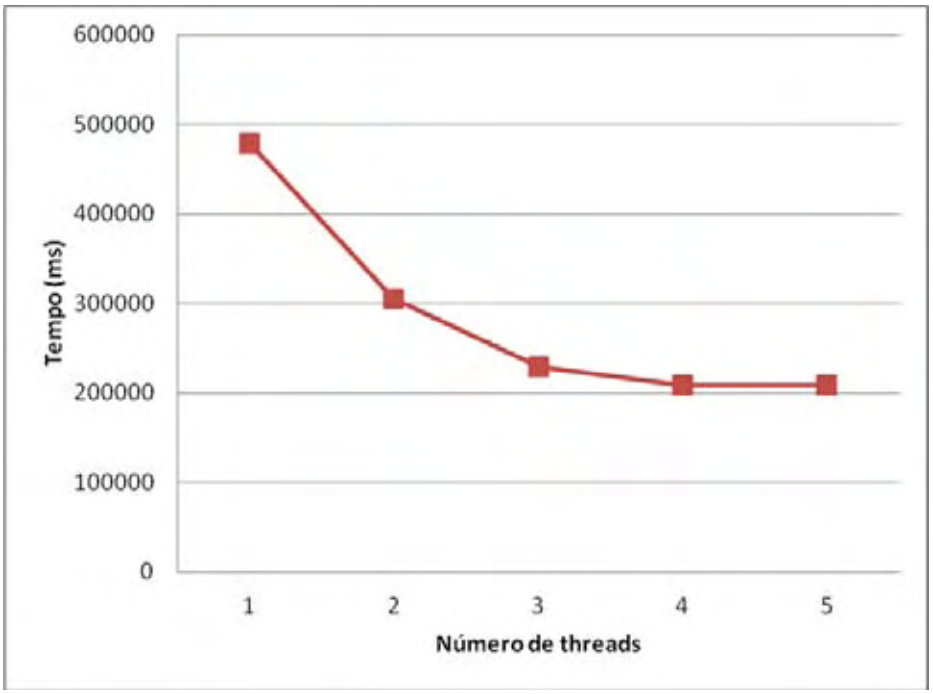


Figura 27 – Teste de desempenho do algoritmo com variação do número de *threads*

4.5 Considerações finais

Neste capítulo foram apresentados os experimentos realizados para a comprovação das características implementadas no algoritmo desenvolvido, bem como a eficácia das técnicas e abordagens propostas.

Pelos resultados obtidos, foram exemplificadas algumas aplicações do algoritmo em uma base de dados reais da saúde pública, com um volume de dados numa dimensão em que o CLARANS, um dos principais algoritmos de agrupamento por particionamento, se mostrou inviável para a tarefa de prospecção, diferentemente do algoritmo desenvolvido, que além de ter sido enriquecido com abordagens inovadoras que contribuíram para a melhoria da qualidade dos resultados, executou as atividades com um tempo inferior nos experimentos realizados.

Capítulo 5 Conclusões

Neste trabalho foram apresentadas algumas informações sobre a área de *spatial data mining*, que possibilitaram estabelecer um panorama do que tem sido proposto na literatura. Nesse contexto, o algoritmo apresentado neste trabalho foi concebido com a finalidade de introduzir uma abordagem inovadora para efetivar a análise de dados espaciais, com foco na qualidade dos resultados e na otimização do tempo de execução nas aplicações em bases de dados volumosas.

Para isso, alguns dos algoritmos mais tradicionais da área foram utilizados como base para a criação da nova abordagem – aplicável a contexto multirrelacional e com recursos para prospecção de grande volume de dados, sem que a qualidade dos resultados seja comprometida; pelo contrário, a maneira como as características não espaciais dos objetos georreferenciados são organizadas possibilita a agregação semântica para orientar, de forma ainda mais eficiente, o agrupamento de objeto similares com base em ontologias, por exemplo. Além disso, a restrição de parâmetros de entrada que influenciam diretamente no resultado final promovido pelas técnicas tradicionais foi eliminada, sendo unicamente aberta a escolha do grau de similaridade entre objetos de um mesmo agrupamento que seja deseja obter.

Por meio dos experimentos realizados, foi possível confirmar a eficácia do algoritmo no desempenho das atividades que se propõe a executar, principalmente na abordagem desenvolvida para absorver o contexto de prospecção multirrelacional, o que possibilitou que fosse evitada a perda semântica ocasionada pela junção de dados e demais

tarefas de pré-processamento para uni-los todos numa única relação, além de ter contornado o alto custo computacional requerido por essas operações.

Destaca-se também que em todos os casos analisados nos exemplos de aplicação do algoritmo desenvolvido os resultados somente foram levantados por conta da localidade ter sido considerado como um elemento de importância na prospecção dos dados, uma vez que, ao se aplicar *data mining* convencional, as características dos agrupamentos muitas vezes seriam reportadas com pouca expressividade diante do universo da base de dados. No entanto, ao se considerar a localidade como um fator relevante, a proximidade geográfica onde as referidas características ocorrem faz com que esses agrupamentos sejam identificadas e, portanto, informações implícitas na base de dados passam a ser explícitas após a prospecção com essa abordagem.

5.1 Comparação do algoritmo com os correlatos

Na Tabela 6, uma breve comparação do algoritmo desenvolvido em relação a alguns dos mais tradicionais propostos na literatura é apresentada, o que destaca as contribuições deste trabalho.

Tabela 6 – Comparativo do MR-Clustering com os principais algoritmos de agrupamento de dados espaciais

	VDBSCAN (densidade)	DBSCAN (densidade)	OPTICS (densidade)	ST-DBSCAN (densidade)	CCDD (densidade)	CLARANS (particionamento)	PCLARANS (particionamento)	K-MEANS (particionamento)	**MR-Clustering (densidade)**
Sem exigência de parâmetros de entrada	✔	✖	✔	✖	✖	✖	✖	✖	✔
Agrupamento por similaridade de características não espaciais e espaciais	✖	✖	✖	✖	✖	✔	✔	✖	✔
Robusto à variação de densidade	✔	✖	✔	✔	✖	✖	✖	✖	✔
Aplicável a contexto multirrelacional sem preparação dos dados	✖	✖	✖	✖	✖	✖	✖	✖	✔
Multithreading no processamento dos dados	✖	✖	✖	✖	✖	✖	✖	✖	✔
Suporta agregação de semântica na formação de agrupamentos	✖	✖	✖	✖	✖	✖	✖	✖	✔

5.2 Trabalhos futuros

Considerando a continuidade do trabalho desenvolvido, uma contribuição que pode ser implementada é a criação de uma quarta etapa do algoritmo, que teria como objetivo a classificação dos agrupamentos retornados por meio de tratamento de obstáculos e de semântica implícita nas relações espaciais, que geralmente não são representadas pelos atributos que descrevem os objetos nas bases de dados geográficos.

A formação de agrupamentos de objetos similares com base na análise das características espaços-temporais também pode ser um indicativo para a continuidade do trabalho, uma vez que, no estado da arte, diversos são os trabalhos que tem direcionado esforços com este propósito.

Ao considerar que nem sempre os dados estão disponíveis em uma única base de dados, algumas contribuições futuras podem ser obtidas com a adaptação do algoritmo para habilitá-lo à prospecção em banco de dados distribuídos.

O uso de *grids* computacionais pode ser uma alternativa para otimização do desempenho do algoritmo quanto ao tempo de execução, sendo também necessárias adaptações para que o mesmo seja executado neste ambiente com o máximo de aproveitamento dos recursos. Essa abordagem é bastante exequível, uma vez que é possível contar com a disponibilidade do gridUNESP, que fornece aproximadamente 3 mil núcleos de processamento interligados em alta velocidade.

Para finalizar, outro trabalho que poderia render dividendos interessantes e contribuições inovadoras ao algoritmo é a implementação das técnicas com conceitos para uso dos recursos de processamento gráfico. Esse tipo de abordagem é recente e tem apresentado eficiência na otimização do desempenho de tarefas que exigem maior poder computacional [DAS_12] [DIA_12].

Referências bibliográficas

[AGR_93] Agrawal, R. et al. 1993. Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93* (New York, New York, USA, 1993), 207-216.

[AGR_98] Agrawal, R. et al. 1998. Automatic subspace clustering of high dimensional data for data mining applications. *ACM SIGMOD Record*. 27, 2 (Jun. 1998), 94-105.

[ANK_99] Ankerst, M. et al. 1999. OPTICS: ordering points to identify the clustering structure. *ACM SIGMOD Record*. 28, 2 (Jun. 1999), 49-60.

[BAE_09] Bae, D.-H. et al. 2009. SD-Miner: A spatial data mining system. *2009 IEEE International Conference on Network Infrastructure and Digital Content* (Nov. 2009), 803-807.

[BIR_07] Birant, D. and Kut, A. 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*. 60, 1 (Jan. 2007), 208-221.

[CAM_95] Câmara, G. 1995. Modelos, linguagens e arquiteturas para bancos de dados geográficos. Tese (Doutorado em Computação Aplicada), Instituto Nacional de Pesquisas Espaciais (INPE) - São José dos Campos, São Paulo.

[CHA_00] Chawla, S.; Shekhar, W. Wu; Ozesmi, U. 2000. Modeling spatial dependencies for mining geospatial data: An introduction. In *H. Miller and J. Han, Geographic data mining and Knowledge Discovery (GKD).*

[CHE_06] Chen, Y.-L. et al. 2006. A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales. *Decision Support Systems*. 42, 3 (Dec. 2006), 1503-1520.

[CHE_09] Chen, G.-F. et al. 2009. Research on Spatially Weighted Fuzzy Dynamic Clustering Algorithm and Spatial Data Mining Visualization. *2009 WRI World Congress on Software Engineering* (2009), 60-66.

[CHO_10] Chowdhury, A.K.M.R. et al. 2010. An efficient method for subjectively choosing parameter "k" automatically in VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise) algorithm. *2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE)* (Feb. 2010), 38-41.

[DAS_12] Dashora, S. and Khare, N. 2012. Implementation of graph algorithms over GPU: A comparative analysis. *2012 IEEE Students' Conference on Electrical, Electronics and Computer Science* (Mar. 2012), 1-8.

[DEZ_03] Džeroski, S. 2003. Multi-relational data mining. *ACM SIGKDD Explorations Newsletter*. 5, 1 (Jul. 2003), 1.

[DIA_12] Diaz, J. et al. 2012. A Survey of Parallel Programming Models and Tools in the Multi and Many-Core Era. *IEEE Transactions on Parallel and Distributed Systems*. 23, 8 (Aug. 2012), 1369-1386.

[DOM_03] Domingos, P. 2003. Prospects and challenges for multi-relational data mining. *ACM SIGKDD Explorations Newsletter*. 5, 1 (Jul. 2003), 80.

[EST_96] Ester, M. et. al. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *In Proceedings of KDD*. 1996, 226-231.

[EST_97] Ester, M., Kriegel, H. P. and Sander, J. 1997. Spatial Data Mining: a Database Approach. *In: Scholl M V, ed. Proceedings of the 5thInternationai Symposium on Spatial Databases (SSD. 97).* Berlin:Springer-Verlag.

[FAN_09] Fan, W. and Luo, W. 2009. The Key Technologies Research of Spatial Data Mining Based on the GIS Grid Services.*2009 International Conference on Computational Intelligence and Software Engineering*. (2009), 1-4.

[FRA_09] Frank, R. et al. 2009. A multi-relational approach to spatial classification. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09* (New York, New York, USA, 2009), 309.

[GUH_98] Guha, S. et al. 1998. CURE: an efficient clustering algorithm for large databases. *ACM SIGMOD Record*. 27, 2 (Jun. 1998), 73-84.

[HAN_06] Han, J.; Kamber, M. 2006. Data mining: concepts and techniques. *2 ed. San Francisco: Morgan Kaufmann Publishers*, 2006. 743 p.

[HEB_10] He, B. et al. 2010. The Research of the Data Mining Based on the Spatial Database Technology. *2010 International Conference of Information Science and Management Engineering* (Aug. 2010), 203-206.

[HEY_09] He, Y. and Li, X. 2009. A Study of Spatial Data Mining Technique Based on Web. *2009 International Conference on Management and Service Science* (Sep. 2009), 1-4.

[HIN_98] Hinneburg, A. and Keim, D.A. 1998. An efficient approach to clustering in large multimedia databases with noise. *Knowledge Discovery and Data Mining* (1998), 58-65.

[HUA_10] Hua, Z. et al. 2010. Clustering algorithm based on characteristics of density distribution. *2010 2nd International Conference on Advanced Computer Control* (2010), 431-435.

[JIA_09] Jia, Z. and Liu, Y. 2009. Visualized Spatial Data Classifying Based on Spatial Data Mining. *2009 First International Workshop on Education Technology and Computer Science* (2009), 133-137.

[JIA_10] Jiaxiang, L. et al. 2010. Distributed Spatial Data Mining in Geospatial Knowledge Service Grid. *2010 Second International Conference on Advanced Geographic Information Systems, Applications, and Services* (Feb. 2010), 80-87.

[JIE_10] Jiehai, C. and Wei, L. 2010. Research on the storage and management of mine spatial data based on PostgreSQL. *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)* (Oct. 2010), V9-493-V9-496.

[JIM_10] Ji, M. et al. 2010. Mine geological hazard multi-dimensional spatial data warehouse construction research. *2010 18th International Conference on Geoinformatics* (Jun. 2010), 1-5.

[JIN_10] Jin, H. and Miao, B. 2010. The research progress of spatial data mining technique. *2010 3rd International Conference on Computer Science and Information Technology* (Jul. 2010), 81-84.

[KAC_02] Kacar, E.; Cicekli, N. K. 2002. Discovery Fuzzy Spatial Association Rules, Data Mining and Knowledge Discovery: Theory, Tools and Technology IV. *In: Dasarathy B V, ed. Proceedings of SPIE*, VoI4730, 2002. 94-102.

[KAR_09] Karmaker, A. and Rahman, S. 2009. Outlier Detection in Spatial Databases Using Clustering Data Mining. *2009 Sixth International Conference on Information Technology: New Generations* (2009), 1657-1658.

[KON_11] Kondaveeti, A. et al. 2011. Extracting geographic knowledge from sensor intervention data using spatial association rules. *Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services* (Jun. 2011), 127-130.

[KOP_95] Koperski, K.; Han, J. 1995. Discovery of Spatial Association Rules in Geographic Information Databases. *In Proceedings of 4th International Symposium on Large Spatial Databases*. pp 47-66, Portland, Maine, Ago. 1995.

[LEE_07] Lee, A.J.T. et al. 2007. Mining spatial association rules in image databases. *Information Sciences*. 177, 7 (Apr. 2007), 1593-1608.

[LIB_10] Li, B. and Liu, J. 2010. Research on spatial data mining based on uncertainty in Government GIS. *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery* (Aug. 2010), 2905-2908.

[LIG_10] Li, G. et al. 2010. Spatial Data Mining and its application in Marine Geographical Information System. *2010 The 2nd Conference on Environmental Science and Information Application Technology* (Jul. 2010), 514-516.

[LIM_10] Lim, S. 2010. Cleansing Noisy City Names in Spatial Data Mining. *2010 International Conference on Information Science and Applications* (2010), 1-8.

[LIU_03] Liu, J.; Yunhe, P. 2003. Design and Implementation of FIPT - Based Spatial Association Rules Mining Algorithm. *Journal of Image and Graphics*, 2003, 8A (4): 476-480.

[LIU_07] Liu, P. et al. 2007. VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise. *2007 International Conference on Service Systems and Service Management* (Jun. 2007), 1-4.

[LUO_03] Luo Zhi-Qing; Li Qi. 2003. Research on Urban Spatial Data Infrastructure. *Geography and Geo-Information Science*, 2003, 19(3): 32-34.

[MAD_09] Maddox, J. and Shin, D.-G. 2009. Applying Relational Dependency Discovery Framework to Geo-spatial Data Mining. *2009 International Conference on Information and Multimedia Technology* (2009), 10-14.

[MAR_97] Maravalle, M., Simeone, B. and Naldini. 1997. R. Clustering on trees. *Computational Statistics & Data Analysis*, v. 24, n., p. 217-234, 1997.

[MIL_01] Miller, H. J. And Han, J. 2001. Geographic data mining and knowledge discovery. *London: Taylor and Francis*, 2001: 3-32.

[NEV_01] Neves, M. C., Freitas, C. C. and Câmara, G. Mineração de Dados em Grandes Bancos de Dados Geográficos. 2001. Disponível em: <http://www.dpi.inpe.br/geopro/modelagem/relatorio_data_mining.pdf >. Acesso em: 05 jun. 2012.

[NGR_02] Ng, R. T. and Han, J. 2002. CLARANS: A Method for Clustering Objects for Spatial Data Mining. *IEEE Transactions on Knowledge and data engineering,* [S.I], v. 14, n. 5, set.2002, p. 1003-1016.

[PAR_11] Parimala, M. et al. 2011. A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases. *Science And Technology*. 31, (2011), 59-66.

[PEI_01] Pei Tao, Zhou Cheng-Hu, Luo Jian-Cheng, Han Zhi-Jun. Review on the Proceedings of Spatial Data Mining Research. *Journal of Image and Graphics*, 2001, 6(9): 854-860.

[PEN_09] Peng, S. et al. 2009. VegaMinerPOI: A spatial data mining system for POI datasets. *2009 17th International Conference on Geoinformatics* (Aug. 2009), 1-4.

[SAN_98] Sander, J. et al. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications. 1998. *Data Mining and Knowledge Discovery*, 1998, 2(2): 169-194.

[SHE_02] Shekhar, S. et al. 2002. What's Spatial About Spatial Data Mining: Three Case Studies. *Training*. Kluwer Academic Publishers. 28.

[SHE_03] Shekhar, S. et al. Trends in Spatial Data Mining. In*: Next Generation Challenges and Future*, Minneapolis, MN, 2003.

[SHE_98] Sheikholeslami, G. et al. 2000. WaveCluster: a wavelet-based clustering approach for spatial data in very large databases. *Very Large Data Bases The International Journal on*. 8, 3-4 (2000), 289-304.

[SHE_11] Shengwu, H. 2011. Method development about spatial data mining and its problem analysis. *Proceedings of 2011 International Conference on Electronics and Optoelectronics* (Jul. 2011), V2-144-V2-147.

[SHU_09] Shun, H.Y. and Wei, X. 2009. A study of spatial data mining architecture and technology. *2009 2nd IEEE International Conference on Computer Science and Information Technology* (2009), 163-166.

[TSE_99] Tsechansky, M.S. and Pliskin, N. 1999. Mining relational patterns from multiple relational tables. *Decision Support Systems*. 27, 1 (1999), 177-177.

[VAL_11] Valêncio, C.R. et al. 2011. Spatial Clustering Applied to Health Area. *2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies* (Oct. 2011), 427-432.

[VAL_12] Valêncio, C.R. et al. 2011. Web Geographic Information System to support environmental resource management. *8th International Conference on Ecological Informatics* (Dec. 2012).

[VIJ_10] Vijayalakshmi, S. and Punithavalli, M. 2010. Improved varied density based spatial clustering algorithm with noise. *2010 IEEE International Conference on Computational Intelligence and Computing Research*. (2010), 1-4.

[WAN_97] Wang, W. and Muntz, R. 1997. STING : A Statistical Information Grid Approach to Spatial Data Mining. (1997).

[WAN_09a] Wang, J. et al. 2009. Research of GIS-based Spatial Data Mining Model. *2009 Second International Workshop on Knowledge Discovery and Data Mining* (Jan. 2009), 159-162.

[WAN_09b] Wang, P. et al. 2009. Research on Logistics Oriented Spatial Data Mining Techniques. *2009 International Conference on Management and Service Science* (Sep. 2009), 1-4.

[XIN_03] Xin, W. and Hamilton, H. J. 2003. DBRS: A Density-Based Spatial Clustering Method with Random Sampling. In: *Proc. of the 7th PAKDD*, Seoul, Korea, 2003. 563-575.

[XUE_10] Xue Jing-Sheng. Parallel CLARANS- improvement and application of CLARANS algorithm. *Computer and Communication Technologies in Agriculture Engineering (CCTAE), 2010 International Conference* On , vol.3, no., pp.248-251, 12-13, 2010.

[YAN_09] Yan-hui, W. and Hao, M. 2009. Spatial Data Mining on Hierarchical Semantic Relation among Multi-scale Geographical Representations. *Education Technology and Computer Science 2009 ETCS 09 First International Workshop on* (2009), 640-644.

[ZHA_96] Zhang, T. et al. 1996. BIRCH : A New Data Clustering Algorithm and Its Applications. 40, (1996).

Autorizo a reprodução xerográfica para fins de pesquisa.

São José do Rio Preto, _____/_____/____

_____
Assinatura