UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

WELDER BATISTA DE OLIVEIRA

Operações Espaciais Robustas à Imprecisão nas Coordenadas Geográficas

Goiânia
2017

sistema de bibliotecas ufg

PRPG
PRÓ-REITORIA DE PÓS-GRADUAÇÃO

UFG

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1. Identificação do material bibliográfico: **[X] Dissertação** **[ ] Tese**

2. Identificação da Tese ou Dissertação:

Nome completo do autor: Welder Batista de Oliveira

Título do trabalho: Operações Espaciais Robustas à Imprecisão nas Coordenadas Geográficas

3. Informações de acesso ao documento:

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF da tese ou dissertação.

Welder Batista de Oliveira
Assinatura do(a) autor(a)[2]

Ciente e de acordo:

Kleber Vieira Cardoso
Assinatura do(a) orientador(a)[2]

Data: 29/09/2017

[1]Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.

Casos de embargo:
- Solicitação de registro de patente
- Submissão de artigo em revista científica
- Publicação como capítulo de livro
- Publicação da dissertação/tese em livro

[2]A assinatura deve ser escaneada.

Versão atualizada em maio de 2017.

WELDER BATISTA DE OLIVEIRA

Operações Espaciais Robustas à Imprecisão nas Coordenadas Geográficas

Dissertação apresentada ao Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação.

Orientador: Prof. Kleber Vieira Cardoso

Goiânia
2017

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Batista de Oliveira, Welder
Operações Espaciais Robustas à Imprecisão nas Coordenadas Geográficas [manuscrito] / Welder Batista de Oliveira. - 2017.
lvii, 57 f.: il.

Orientador: Prof. Dr. Kleber Vieira Cardoso; co-orientador Dr. Vagner José do Sacramento Rodrigues.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2017.
Bibliografia.
Inclui siglas, mapas, abreviaturas, símbolos, gráfico, tabelas, algoritmos, lista de figuras, lista de tabelas.

1. Sistemas de Informação Geográfica. 2. incerteza. 3. junção espacial. 4. consulta skyline. 5. coordenadas imprecisas. I. Vieira Cardoso, Kleber, orient. II. Título.

CDU 004

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

UFG

ATA Nº 19/2017

ATA DA SESSÃO DE JULGAMENTO DA DISSERTAÇÃO DE MESTRADO DE WELDER BATISTA DE OLIVEIRA

Aos vinte e um dias do mês de agosto de dois mil e dezessete, às catorze horas, na sala 150 do Instituto de Informática da Universidade Federal de Goiás, Campus Samambaia, reuniu-se a banca examinadora designada na forma regimental pela Coordenação do Curso para julgar a dissertação de mestrado intitulada **"Operações Espaciais Robustas à Imprecisão nas Coordenadas Geográficas",** apresentada pelo aluno Welder Batista de Oliveira como parte dos requisitos necessários à obtenção do grau de Mestre em Ciência da Computação, área de concentração Ciência da Computação. A banca examinadora foi presidida pelo orientador do trabalho de dissertação, Professor Doutor Kleber Vieira Cardoso (INF/UFG), tendo como membros os Professores Doutores Vagner José do Sacramento Rodrigues (INF/UFG - coorientador), Clodoveu Augusto Davis Junior (DCC/UFMG) e Helton Saulo Bezerra dos Santos (EST/UNB). Aberta a sessão, o candidato expôs seu trabalho. Em seguida, o aluno foi arguido pelos membros da banca e:

(X) tendo demonstrado suficiência de conhecimento e capacidade de sistematização do tema de sua dissertação, a banca concluiu pela **aprovação** do candidato, sem restrições.

( ) tendo demonstrado suficiência de conhecimento e capacidade de sistematização do tema de sua dissertação, a banca concluiu pela **aprovação** do candidato, condicionado a satisfazer as exigências listadas na Folha de Modificação de Dissertação de Mestrado anexa à presente ata, no prazo máximo de 60 dias, a contar da presente data, ficando o professor-orientador responsável por atestar o cumprimento dessas exigências.

( ) não tendo demonstrado suficiência de conhecimento e capacidade de sistematização do tema de sua dissertação, a banca concluiu pela **reprovação** do candidato.

Os trabalhos foram encerrados às 17 horas. Nos termos do Regulamento Geral dos Cursos de Pós-Graduação desta Universidade, lavrou-se a presente ata que, lida e julgada conforme, segue assinada pelos membros da banca examinadora.

Prof. Dr. Kleber Vieira Cardoso Kleber Vieira Cardoso

Prof. Dr. Vagner José do Sacramento Rodrigues

Prof. Dr. Clodoveu Augusto Davis Junior

Prof. Dr. Helton Saulo Bezerra dos Santos

Welder Batista de Oliveira

Graduou-se em Matemática pela Universidade Federal de Goiás (UFG) em 2007 e em Estatística em 2014 pela mesma universidade. Durante a gradução em Matemática foi monitor de Geometria Euclidiana Plana e bolsista PROBEC desenvolvendo projetos educacionais relacionados ao uso de *softwares* livres voltados ao ensino e aprendizagem de geometria na Educação Básica. No período em que cursou Estatística, foi bolsista CNPq na modalidade de Desenvolvimento Tecnológico e Industrial DTI-C. Durante o mestrado foi bolsista FAPEG e publicou um artigo na Revista Brasileira de Cartografia entitulado: "*A Method for Location Recommendation via Skyline Query Tolerant to Noised Georeferenced Data*". Foi co-autor no artigo publicado no JIDME 2015, entitulado: *Understanding and Modeling the Behavior of Web Map Users*.

Dedico esta dissertação a meus pais, por todo o suporte prestado em minha trajetória acadêmica, a minha namorada, por sua compreensão e carinho e a todos os meus professores por sua contribuição ao meu desenvolvimento intelectual e pessoal.

Agradecimentos

Agradeço aos meus pais, Helena Aparecida Breve de Oliveira e Valmir Batista Breve, por todo suporte prestado durante minha trajetória acadêmica. Igualmente agradeço ao meu irmão Valmir Lucas de Oliveira por todo apoio recebido. À Daniela, minha namorada, por todo seu amor, carinho, compreensão e apoio. Agradeço também por compartilhar os momentos de alegria e de tensão presentes ao longo de minha jornada durante todo esse período. Ao Prof. Kleber Vieira Cardoso, por sua orientação, amizade, paciência e confiança. Ao Prof. Vagner José do Sacramento Rodrigues, por sua orientação, contribuição e confiança. Aos Profs. Leonardo, Clodoveu e Helton, por aceitarem o convite, pela presença na banca e pelas contribuições à dissertação. Ao Sávio S. Teles de Oliveira e ao Helton Saulo por toda sua contribuição ao trabalho realizado, sobretudo nos trabalhos submetidos para os congressos e revistas. Além disso, também agradeço ao Sávio S. Teles de Oliveira pelas valiosas implementações das junções espaciais probabilísticas propostas neste trabalho, as quais permitiram que a avaliação experimetal fosse realizada. À equipe da secretaria: Mirian e Mariana pela atenção, paciência e suporte operacional. Ao INF/UFG, pelas instalações e equipamentos utilizados. À goGeo, pelas instalações e equipamentos utilizados. Agradeço à Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG), pelo suporte financeiro.

Everything is related to everything else, but near things are more related than distant things

Tobler W,
A computer movie simulating urban growth in the Detroit region.

Resumo

Batista de Oliveira. **Operações Espaciais Robustas à Imprecisão nas Coordenadas Geográficas**. Goiânia, 2017. 57p. Dissertação de Mestrado. Instituto de Informática, Universidade Federal de Goiás.

Os Sistemas de Informação Geográfica revolucionaram a pesquisa geográfica nas últimas três décadas. Esses sistemas comumente disponibilizam uma série de funcionalidades para processar e analisar dados espaciais, como, por exemplo, a junção espacial e a consuta *skyline*. Embora relevantes, a eficácia dessas funcionalidades é impactada pela imprecisão das coordenadas geográficas obtidas pelo método de georreferenciamento empregado. Além disso, o erro contido nas coordenadas pode apresentar diversos padrões distribucionais, o que demanda o desenvolvimento de soluções que sejam generalistas quanto ao padrão de erro que conseguem tratar adequadamente. Por fim, operações espaciais já são computacionamente caras em sua versão determinística, o que se agrava com a introdução do componente estocástico. O presente trabalho apresenta uma estrutura geral para o desenvolvimento de soluções para operações espaciais robustas a coordenadas imprecisas. Além disso, para lidar com os problemas mencionados, a estrutura proposta é projetada para contemplar os requisitos de generalidade, eficácia e eficiência em patamares que viabilizem sua aplicação prática. A estrutura geral de solução é composta pela combinação de versões probabilísticas de heurísticas das versões determinísticas das operações espaciais e por simulações de Monte Carlo. A partir dela, são desenvolvidas as soluções específicas - como estudo de caso - para a *skyline* espacial e da junção espacial. Resultados teóricos e experimentais demonstraram o potencial das soluções desenvolvidas em atender aos três requisitos estabelecidos nesse trabalho.

Palavras–chave

Sistemas de Informação Geográfica, incerteza, junção espacial, consulta *skyline*, coordenadas imprecisas.

Abstract

Batista de Oliveira. **Spatial Operations Robust to Geographic Coordinates Uncertainty**. Goiânia, 2017. 57p. MSc. Dissertation. Instituto de Informática, Universidade Federal de Goiás.

Geographic Information Systems have revolutionized geographic research over the past three decades. These systems commonly provide a number of features for processing and analyzing spatial data, such as spatial join and skyline. Although relevant, the effectiveness of such functionalities is affected by the imprecision of the geographic coordinates obtained by the georeferencing method employed. Moreover, the error contained in the coordinates may present several distributional patterns, which demands the development of solutions that are generalist concerning the error pattern that they can handle properly. Finally, spatial operations are already computationally expensive in their deterministic version, which is aggravated by the introduction of the stochastic component. The present work presents a general structure of spatial operations solutions robust to imprecise coordinates based on the use of simulations and probabilistic adaptations of heuristics in the literature. In addition, to deal with the problems mentioned, the proposed structure is designed to contemplate the requirements of generality, accuracy and efficiency at levels that enable its practical application. The overall solution structure is composed of the combination of probabilistic versions of heuristics of the deterministic versions of the spatial operations and by Monte Carlo simulations. From that structure, specific solutions - as case studies - are developed for the spatial join and skyline. Theoretical and experimental results demonstrated the potential of the developed solutions to meet the three requirements established in this work.

Keywords

Geographic Information System, uncertainty, spatial join, skyline query, imprecise coordinates.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Lista de Códigos de Programas

CAPÍTULO 1

Introdução

Nesse capítulo, abordamos o problema motivador do trabalho: o impacto da imprecisão das coordenadas geográficas nas operações espaciais de forma geral e, especicamente, da consulta *skyline* e da junção espacial. Enfatizamos como a imprecisão nas coordenadas pode prejudicar a eficácia dessas operações, a ponto de comprometer sua utilidade prática. Além disso, apresentamos nossa proposta de estrutura geral de soluções para operações espaciais robustas à imprecisão das coordenadas geográficas.

1.1 O impacto da imprecisão das coordenadas nas operações espaciais

Os Sistemas de Inteligência Geográfica (SIG) revolucionaram a pesquisa geográfica nas últimas três décadas, tendo enormes impactos na pesquisa geográfica e na comunidade científica de maneira geral, sobretudo na forma como os cientistas colaboram e comunicam seus estudos [13]. Pode-se definir tais sistemas como um banco de dados digital no qual um sistema comum de coordenadas espaciais é o principal meio de referência [13]. Um SIG abrangente requer um meio de: a) inserir dados espaciais; b) armazená-los; c) transformá-los, analisá-los e modelá-los; d) reportá-los, por exemplo com o uso de mapas. Dentre as várias funcionalidades que podem ser nativas aos SIGs, pode-se citar: renderização de pontos, linhas, polígonos e imagens de satétite, interseção entre objetos espaciais (junção espacial), geocodificação de endereços, geocodificação reversa, consultas *skyline* e muitas outras.

As operações espaciais demandadas por SIGs, e consequentemente pelas aplicações que delas derivam, normalmente assumem uma precisão absoluta nas coordenadas dos objetos representados. Na realidade, dados espaciais provindos das mais diversas fontes apresentam distintos padrões de erro, devido a uma série de fatores. Por exemplo, [16] aponta que a precisão horizontal (com intervalo de confiança de 95%) do GPS (*Global Positioning System*) com padrão WAAS (*Wide Area Augmentation System*) para a cidade

de Los Angeles é de $0,922$ metros, enquanto que a precisão vertical para Miami é de $1,373$ metros.

Procedimentos de geocodificação - uma forma menos precisa que o GPS de se obter coordenadas geográficas a partir de endereços escritos em linguagem natural - possuem tipicamente erros de pelo menos 100 metros em cerca de 20% a 30% dos casos [11]. Por sua vez, métodos baseados em imagens de satélite podem ser afetados, por exemplo, pela escala usada como referência na produção dos dados.

O impacto que o erro posicional dos objetos possui em uma dada operação espacial depende de dois fatores: 1) a própria magnitude do erro; 2) a robustez da operação a coordenadas imprecisas. Para exemplificar, considere que A é uma camada de dados georreferenciados representando casas e B é uma camada georreferenciada representando os bairros de uma cidade. Uma junção espacial aplicada a essas duas camadas, ambas com precisão nas coordenadas a nível de GPS, não deverá produzir um número expressivo de casamentos equivocados (falsos positivos) ou ausência de casamentos legítimos (falsos negativos). Contudo, caso A tenha sido obtida com um procedimento de geocodificação, por exemplo, a probabilidade de que uma dada casa seja erroneamente atribuída a um bairro X ou a um vizinho passa a ser mais expressiva. A proporção de erros se torna ainda maior caso as áreas georreferenciadas em B sejam menores, por exemplo, caso representem setores censitários (menor unidade geográfica definida pelo IBGE) ou mesmo quadras. As operações espaciais constantes em aplicações ou em plataformas SIG normalmente não levam em consideração a magnitude dos erros associados às coordenadas dos dados a que são aplicadas. Como consequência, para alguns cenários de uso, as operações espaciais podem ter sua eficácia comprometida, e portanto, apresentar uma menor utilidade para o usuário da solução.

Embora a magnitude dos erros possa ser aleatória e em alguma direção arbitrária, esses são, comumente, passíveis de serem modelados com alguma função de densidade de probabilidade (FDP) existente na literatura. As FDPs são funções que permitem que a probabilidade do erro E assumir um valor em um dado intervalo (a, b) de valores reais possa ser calculada. Diferentes técnicas de georreferenciamento comumente apresentarão padrões de erro distintos e requerirão FDPs distintas para que a modelagem do erro seja efetuada de forma efetiva. Logo, o cruzamento de duas camadas de dados distintas deve considerar os dois padrões de erros envolvidos. Por exemplo, uma camada A pode apresentar erro médio de 100 metros, enquanto uma camada B pode apresentar erro médio de 1000 metros. Isso se constitui em um problema para técnicas tradicionais e demanda cuidados adicionais por partes dos analistas ao aplicar ou interpretar os resultados obtidos com essas técnicas. No entanto, ao modelar probabilisticamente os erros nas duas camadas, pode-se desenvolver operações espaciais que tratem adequadamente os diferentes padrões de erros e produzam resultados probabilísticos consistentes, dispensando a ne-

cessidade de maiores cuidados por parte dos analistas quanto à aplicação e interpretação dos resultados.

Para exemplificar, caso seja necessário realizar uma junção espacial nas camadas A e B mencionadas acima, utilizando uma técnica tradicional, o analista deverá ter o cuidado de interpretar uma interseção $(a,b), a \in A, b \in B$ apontada pelo algoritmo não necessariamente como uma interseção, mas como um dos possíveis seguintes cenários: 1) há de fato interseção entre a e b; 2) os objetos reais representados nos dados pelas geometrias a e b podem apenas estar próximos o suficiente para que a e b apresentem interseção e esta seja devida ao erro. Por exemplo, os dois objetos podem ser encontrar a uma distância de 500 metros, o que é inferior ao erro médio de B. Ao utilizar uma solução tradicional, o analista, quando não negligencia completamente o impacto dos erros nas coordenadas, deve ter o cuidado de interpretar os resultados da maneira mais adequada possível. As soluções probabilísticas aqui apresentadas eliminam essa complexidade e permitem ao analista determinar uma probabilidade de corte desejada para se considerar que uma dada condição seja verdadeira, por exemplo, que há interseção entre dois objetos, mesmo quando o padrão de erros nas duas camadas envolvidas é distinto.

O presente trabalho tem como objetivo apresentar soluções para operações espaciais que funcionem de maneira eficaz e computacionalmente eficiente, mesmo diante de dados espaciais com níveis significativos de imprecisão nas coordenadas e com erros aleatórios que possam ser modelados com alguma distribuição de probabilidade integrável disponível na literatura. Em outras palavras, propõe-se soluções para operações espaciais que sejam robustas à imprecisão das coordenadas e que atendam aos seguintes requisitos de projeto: **acurácia**, **eficiência** e **generalidade**. Além disso, o cliente das soluções poderá ajustar, por meio dos parâmetros, o balanço desejado entre acurácia e eficiência, i.e. o quanto aceita perder em uma dessas dimensões para ganhar na outra.

Alguns métodos existentes na literatura permitem que funcionalidades SIG possam ser criadas de modo a atender a um ou dois dos requisitos de projeto considerados neste trabalho de maneira relativamente satisfatória. Por exemplo, Ni et al [28] conseguiram adaptar as etapas de uma solução de junção espacial tradicional para o caso específico em que os erros nas coordenadas seguem uma distribuição Circular Normal. Dessa forma, atingiram resultados teóricos e experimentais comparativamente bastante satisfatórios tanto em termos de acurácia como de eficiência (desempenho computacional). Entretanto, sua solução não contempla erros que seguem um padrão típico de outras FDPs, não sendo, portanto, generalista. Por outro lado, Openshaw [29] recomenda o uso de simulações de Monte Carlo para a junção espacial. Como o Método de Monte Carlo (MMC) pode ser aplicado com qualquer distribuição de probabilidade, essa abordagem consegue ser generalista. Contudo, devido ao alto custo computacional envolvido nas simuações requeridas, as soluções desenvolvidas com o MMC apresentam desempenho computacional

proibitivo para muitas aplicações práticas. Para resolver esse problema, propopomos uma adaptação do MMC, a qual chamamos de Método de Monte Carlo Progressivo (MMCP), além de adaptações de heurísticas empregadas em operações espaciais para o caso probabilístico. A seguir apresentamos a estrutura geral de solução proposta em nosso trabalho.

1.2 Estrutura geral da solução

A Figura 1.1 apresenta a estrutura geral de soluções para operações espaciais robustas à imprecisão nas coordenadas. Essa estrutura é uma das contribuições desse trabalho. De maneira transversal, aplica-se o Método de Monte Carlo, ou seja, a realização de simulações computacionais dos erros nas coordenadas. Contudo, o MMC por si só não é capaz de viabilizar soluções ao mesmo tempo eficazes e eficientes. Para tanto, faz-se necessário que em cada operação espacial, as heurísticas utilizadas em algoritmos tradicionais sejam adaptadas para o caso probabilístico. No presente trabalho, oferecemos duas contribuições nesse sentido: o conceito de retângulo de confiança e o Método de Monte Carlo Progressivo. Esses dois dispositivos quando aplicados a uma dada operação espacial impactam significativamente no tempo de processamento das consultas. O capítulo 3 apresenta a aplicação da estrutura geral, além das contribuições específicas, a duas operações espaciais: consulta *skyline* espacial e junção espacial. As próximas duas seções apresentam essas duas operações.

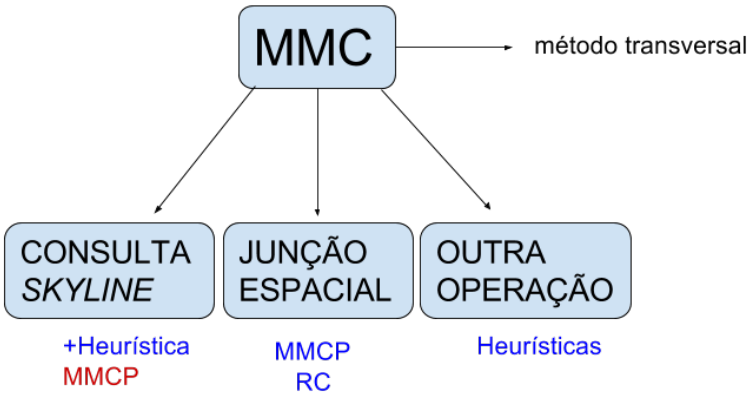

Figura 1.1: *Estrutura geral de soluções para operações espaciais robustas à imprecisão nas coordenadas.*

1.3 Skyline espacial

Algumas aplicações permitem a realização de recomendação de localidade levando em consideração múltiplos, e comumente conflitantes, critérios. Nesses casos pode

não haver uma solução ótima, mas sim várias potencialmente interessantes ao cliente da aplicação. Por exemplo, um turista pode estar interessado em hotéis econômicos com uma classificação razoável (digamos, pelo menos 3 estrelas) e que estejam próximos a um dado ponto turístico na cidade a ser visitada. Nesse exemplo, pode ser interessante retornar ao cliente todas as opções de hotéis que melhor atendam cada um desses critérios (ótimos em cada dimensão), bem como aqueles que apresentam uma boa combinação dos valores nessas dimensões. Uma maneira popular de prover esse tipo de solução é por meio das consultas *skylines* [2].

A consulta *skyline* recupera os objetos que **não são dominados** por qualquer outro no conjunto de dados, Em outras palavras, dado um conjunto S de objetos, a *skyline* de S compreende os elementos Pareto eficientes de S, i.e, elementos $s \in S$ tais que não haja qualquer elemento $r \in S$ com r preferível a s em relação a todos os critérios. No exemplo mencionado anteriormente, caso haja dois hotéis, h_1 e h_2 com a mesma classificação, de modo que h_1 seja ao mesmo tempo mais barato e mais próximo do ponto turístico que h_2, então não há necessidade de retornar $h2$ como opção para o usuário. Nesse caso, é dito que h_1 domina h_2 no conjunto de hotéis considerados e em relação aos critérios avaliados.

A consulta *skyline* que opera sobre critérios e dados espaciais é chamada de *skyline* espacial. Como comentado anteriormente, quanto maior for a magnitude dos erros contidos nas coordenadas dos objetos espaciais, maior será o impacto na operação espacial aplicada sobre tais dados. Nesse contexto, erros advindos de procedimentos menos precisos como a geocodificação, por exemplo, podem comprometer a eficácia dos resultados atingidos pelas aplicações que utilizam a *skyline* como forma de recomendação de localidade. Isso ocorre porque alguns objetos poderão ser erroneamente considerados dominados devido apenas ao erro nas coordenadas.

Há na literatura opções de consultas *skyline* que lidam com dados incertos, a chamada *p-skyline*. Contudo, essas opções não se configuram como *skylines* espaciais ou não fornecem um método baseado na modelagem do erro posicional que permita sua aplicação a cenários de dados espaciais com imprecisão nas coordenadas. Por exemplo, a abordagem de Pei et al [32] leva em consideração instâncias dos objetos espaciais para o cálculo de sua *p-skyline* e não da modelagem de tais objetos por alguma FDP. O presente trabalho apresenta uma solução para a *skyline* espacial com coordenadas imprecisas. A solução proposta será chamada ao longo do texto de **pSkyEMC** (*p-skyline* espacial com Método de Monte Carlo).

A pSkyEMC recebe como entrada: o erro médio μ, o desvio padrão σ, a FDP que modela o erro das coordenadas presentes nos dados, uma probabilidade de corte p requerida para que duas geometrias sejam consideradas casadas, e o número máximo de iterações m de simulações de Monte Carlo usadas para estimar a probabilidade de dominância em cada para um dado par de objetos espaciais e para cada dimensão espacial

considerada. Quanto maior o valor de m, mais eficaz será a avaliação de probabilidade de um dado objeto pertencer à *skyline*, contudo maior será também o custo computacional.

1.4 Junção espacial

De acordo com Patel e DeWitt [31], as aplicações espaciais frequentemente precisam combinar dois conjuntos de dados baseado em alguma relação espacial entre os seus objetos. A essa operação se dá o nome de junção espacial. Por exemplo, dado um conjunto de dados de polígonos que representam queimadas e outro que representam áreas de vegetação, os polígonos que se cruzam nesses dois conjuntos mostrarão quais áreas verdes já são afetadas pelo fogo. Alternativamente, as áreas de vegetação que estão prestes a serem atingidas pelo fogo podem ser identificadas com os mesmos conjuntos de dados, mas usando um predicado de junção diferente, por exemplo, áreas de vegetação que estão a no máximo 1 km de distância dos focos de incêndio.

Embora muito útil, em geral, a avaliação de predicados de junção espacial demanda altos custos computacionais, bem como soluções sofisticadas para alcançar um desempenho satisfatório para grandes volumes de dados como se pode ver em [31]. A tarefa torna-se ainda mais difícil quando precisamos lidar com objetos que apresentam coordenadas imprecisas. Por exemplo, dado uma camada R de rios e uma P de plantações com R apresentando um padrão de erro com média 1000 metros e desvio-padrão de 500 metros e P com erro médio de 100 metros e desvio-padrão de 60 metros, não se pode afimar categoricamente que um objeto $r \in R$ e um $p \in P$ estejam de fato se interceptando ou que estejam a uma distância de, digamos, 20 metros. Dessa maneira, não há como se definir, com certeza, se está havendo inflação ao código florestal, o qual proibe que a plantação esteja a menos de 30 metros de distância do rio (podendo chegar a 500 metros, dependendo da largura do rio). Nesse caso, há a necessidade de se avaliar a probabilidade de que tal evento esteja ocorrendo. Dado a incerteza nas posições dos objetos, a junção pode não ser avaliada de maneira acurada, i.e., concluir não haver casamento quando na realidade esse existe (falso negativo) e apontar aqueles que na realidade não existem (falso positivo). Assim sendo, soluções robustas baseadas em alguma abordagem probabilística são candidatas naturais para melhorar a junção espacial nesses cenários.

A junção espacial com incerteza nas posições pode ser considerada como um problema com objetivos conflitantes. Por um lado, a abordagem probabilística acrescenta acurácia ao custo de adicionar também complexidade computacional. Por outro lado, a junção espacial já possui um alto custo de computação, mas não pode prescindir da eficácia para ser útil. Esses dois objetivos foram simultaneamente alcançados somente em cenários específicos, como em Jinfeng Ni et al (2003) onde se assume que os erros seguem uma distribuição Circular Normal. Logo, há carência de soluções na literatura que

apresentem um bom balanço entre eficiência, precisão e generalidade na modelagem dos padrões de erro. Nosso método - Junção Espacial Probabilística com Método de Monte Carlo Progressivo (JEPMCP) - propõe uma solução que considera esses três objetivos como requisito de seu projeto.

A JEPMCP recebe como entrada: o erro médio μ, o desvio padrão σ, a FDP que modela o erro das coordenadas presentes nos dados, uma probabilidade de corte p requerida para que duas geometrias sejam consideradas casadas, o nível de confiança γ relacionado à precisão da avaliação do predicado de junção e o parâmetro de número máximo de iterações $max.iter$ necessário para garantir a eficiência do algoritmo. Com esses elementos, o algoritmo é capaz de retornar de forma eficiente e acurada todos pares de geometrias (a, b), com $a \in A$ e $b \in B$, tais que a probabilidade de casamento entre a e b seja de pelo menos p.

Os níveis de eficiência e a acurácia são ajustáveis pelos parâmetros γ e $max.iter$. Quanto maiores forem os valores desses parâmetros, maior será a acurácia ao custo de se aumentar também o tempo de processamento necessário para executar a operação. Por outro lado, quanto menores forem seus valores, maior será a eficiência, porém com menor garantia de acurácia. Além disso, o parâmetro γ fornece a probabilidade do casamento entre, digamos a e b, ser avaliado corretamente na etapa de refinamento do algoritmo de junção espacial. Dessa forma, provê-se ao cliente da solução a possibilidade de ajustar o balanço desejado entre tempo de processamento e eficácia da solução ao se fixar probabilisticamente o nível de confiança desejado na avaliação de cada casamento em potencial na junção espacial.

1.5 Organização do trabalho

O capítulo 2 apresenta a fundamentão teórica do trabalho e os principais trabalhos correlatos. O capítulo 3 traz as duas propostas de operações espaciais tolerantes à imprecisão dos dados: pSkyEMC (para a p-skyline) e JEPMCP (para a junção espacial probabilística). São apresentados lemas e teoremas que provam a eficácia dos métodos apresentados. O capítulo 4 apresenta e discute os resultados experimentais obtidos. Por fim, tem-se o capítulo de conclusão do trabalho.

CAPÍTULO 2

Fundamentação teórica e trabalhos relacionados

Neste capítulo, serão apresentados algumas conceitos importantes para o trabalho. Na seção 2.1, são definidos alguns termos comumente usados equivocadamente e de maneira intercambíavel como precisão, acurácia, vagueza e ambiguidade. Na seção 2.2, a consulta *skyline* é definida e um exemplo é apresentado. Também é definida sua versão probabilística. A última seção apresenta a Junção Espacial e sua versão probabilística.

2.1 Incerteza nos dados

Os dados espaciais estão sujeitos a várias formas de incerteza. No que diz respeito a dados espaciais, duas formas de incerteza se destacam: a existencial e a posicional. A incerteza existencial está relacionada à confiança que se tem de que o objeto espacial representado nos dados de fato existe. Isso pode ocorrer quando se está extraindo objetos de uma imagem de satélite com baixa resolução ou definição de cores. Nesse cenário, pode-se não estar 100% certo se uma dada configuração de pixels corresponde a um objeto real [8]. Por outro lado, a incerteza posicional se refere à confiança que se tem na posição do objeto. Nesse caso, as reais coordenadas da representação do objeto na base de dados se encontram deslocadas em relação às coordenadas do objeto na realidade. A esse tipo de incerteza também se dá o nome de **erro posicional**, ou simplesmente erro. Define-se o erro para um dado objeto como a distância entre as suas coordenadas nos dados espaciais e suas coordenadas na realidade. A distância empregada pode ser de vários tipos, tais como geodésica, euclidiana, *Manhattan*, dentre outras. A magnitude do erro está associada ao método utilizado na produção dos dados.

Na Figura 2.1 se apresenta a categorização proposta por Fisher et al (1999). Trata-se de uma categorização generalista dos tipos de incerteza e pode ser resumida da seguinte forma:

1. Se os objetos estão bem definidos, então a incerteza é causada por desvios numéricos entre sua representação nos dados e o valor que estes possuem na realidade

(erros). Nesse caso, a incerteza possui uma natureza essencialmente probabilística;

2. Se os objetos são definidos de maneira vaga, então outros tipos de incerteza podem ser considerados.
   (a) Se a incerteza é atribuída à falta de clareza na definição dos objetos, então a determinação da classe ou conjunto dentro do universo é vaga e pode ser tratada de maneira conveniente pela teoria Fuzzy.
   (b) A incerteza pode também surgir devido à ambiguidade, conduzindo a diferentes sistemas de classificação. Podem haver duas formas:
      i. Quando um objeto é definido de maneira precisa, mas se demonstra ser membro de duas ou mais classes;
      ii. Quando o processo de associação de um objeto a uma classe está aberto à interpretação.

O presente trabalho lida especificamente com a incerteza posicional nos objetos espaciais, ou seja, o erro presente nas coordenadas das representações de tais objetos nos dados georreferenciados.

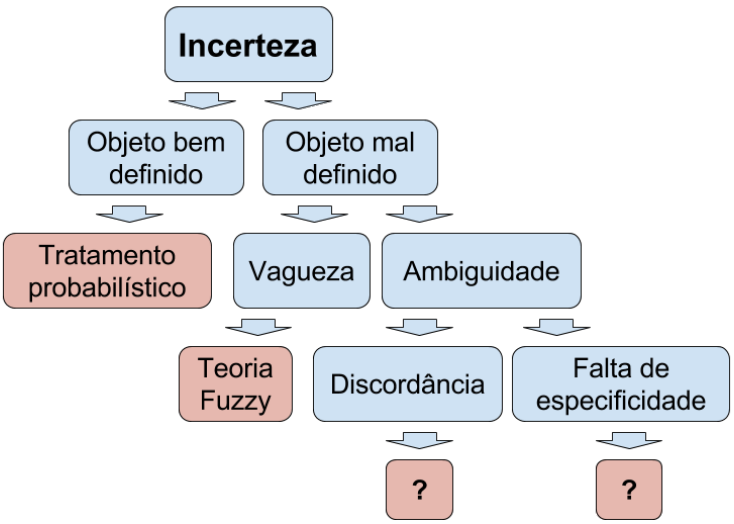

Figura 2.1: *Categorização da incerteza conforme P Fisher et al (1999).*

Alguns termos são comumente e equivocadamente usados de maneira intercambiável, tais como: **precisão**, **acurácia**, **vagueza** e **ambiguidade**. Acurácia é o grau com que uma informação em uma base de dados apresenta valores aceitáveis [12], estando associada à quantidade (proporção, por exemplo) de dados que atendem as especificações de qualidade. Precisão se refere ao nível de variação obtido em uma série de medidas ou estimativas [9], podendo ser expressa pelo desvio-padrão, variância ou outra medida de dispersão.

Para ilustar a diferença entre ambas, usaremos como exemplo um jogo de lançamento de dardos. Nesse jogo, um lançamento pode ser considerado acurado se atingiu o círculo menor ao centro do alvo. Do ponto de vista da precisão, o que importa é a variação em relação ao centro do alvo. Dessa forma, um jogador A pode acertar o círculo menor 8 vezes em 10 lançamentos e apresentar um erro médio de $5cm$, enquanto um jogador B pode acertar tal círculo 7 vezes e apresentar um erro médio de $4cm$. O jogador A apresentou maior acurácia (80% contra 70% do jogador B), porém menor precisão, uma vez que a variação de suas jogadas foi maior. Por sua vez, vagueza se refere à falta de clareza quanto ao significado [9] - estando normalmente associada com a dificuldade de se realizar uma distinção precisa em relação a um objeto no mundo real - e ambiguidade está associada à falta de clareza na definição de pertinência de um objeto a alguma das classes pré-determinadas para um dado fenômeno.

O presente trabalho busca promover e potencializar a aplicação de operações espaciais à dados que apresentam erros aleatórios em suas coordenadas. Os métodos desenvolvidos nesse trabalho são paramétricos, i.e. exigem que determinados parâmetros sejam especificados por seus usuários. Alguns desses parâmetros estão relacionados à distribuição utilizada para modelar os erros. A saber, os parâmetros: média (μ), desvio-padrão (σ) e FDP (a função de densidade de probabilidade). Tais informações podem constar em metadados associados à fonte dos dados, ser provido por especialistas que conheçam o método empregado na produção dos dados ou mesmo estimado caso haja uma amostra para a qual se tenha tanto as posições produzidas pelo método utilizado como as posições reais dos objetos. A partir dessa amostra, pode-se calcular o erro x_i para cada objeto i. De maneira geral, estimativas $\hat{\mu}$ da média μ e S do desvio-padrão σ podem ser obtidas da coleção $x_1, x_2, ..., x_n$, respectivamente por $\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n-1}}$.

Encontrar a distribuição de probabilidade que modela os dados é uma tarefa menos trivial. Leung e Yan ([21] e [22]) recomendam modelar erros posicionais com a distribuição Normal. Entretanto, os erros gerados por processos de aquisição dos dados podem seguir distribuições diferentes. Todavia, desde que haja uma amostra dos erros obtida para alguns objetos, a distribuição do erro pode ser inferida com o uso de alguns métodos estatísticos. Dentre esses métodos, ressalta-se histogramas, gráficos de quantil, teste Kolmogorov-Smirnov e alguns procedimentos de seleção de modelos como Akaike (AIC), Schwarz's Bayesian (BIC) e o critério de informação Hannan-Quinn (HQIC).

2.2 Consulta Skyline

As consultas *skyline* são uma maneira de obter respostas preferenciais de uma base de dados provendo apenas o sentido de ordenação dos valores dos atributos [5]. Esse tipo de consulta retorna as tuplas Pareto eficientes de um conjunto de dados de acordo com

um número d de atributos e seu sentido de ordenação (maximização ou minimização). Pareto eficiência é a propriedade dos objetos que não são dominados por quaisquer outros em um certo conjunto de dados. Sobre a condição de minimização, diz-se que um ponto p_i domina o ponto p_j se e somente se as coordenadas de p_i em qualquer dimensão não são maiores que as coordenadas correspondentes em p_j [30].

Para exemplificar, será usado um problema clássico: "encontrar os hotéis que sejam baratos e próximos à praia". Assim, há dois objetivos a serem buscados, os quais podem ser mutuamente excludentes, i.e, os hotéis mais próximos da praia podem tender a ser os mais caros. Naturalmente, caso haja um hotel A que seja ao mesmo tempo mais barato e mais próximo da praia que um hotel B, não há motivos para que B seja preferido em relação a A. Nesse caso, A domina B.

As Figuras 2.2 e 2.3 mostram um conjunto de hotéis em relação às duas variáveis de interesse: **eixo-x**: *"distância à praia em metros"*; **eixo-y**: *"preço da diária* $(R\$)$*"*. A linha tracejada representa a *skyline* S para o conjunto de hotéis apresentados. No cenário 1, nota-se que A é o hotel mais próximo da praia e L o mais barato. Logo, esses hotéis pertencem a S, já que, por definiçao, não podem ser dominados por quaisquer outros. Além desses, o hotel C também pertence a S, pois não é dominado por qualquer outro neste conjunto de dados. Todos os demais são dominados por algum desses três pontos. Portanto, conforme destacado na Figura 2.2, $S = \{A, C, L\}$. Contudo, uma eventual imprecisão nos dados, pode impactar a consulta. O cenário 2 - Figura 2.3 - mostra como um erro de posição de 150 metros do hotel E o incluiria em S (o retorno da consulta). Nesse caso, faz sentido se perguntar qual seria a probabilidade de E ser Pareto eficiente neste conjunto. Para lidar com situações como essa e incorporar a imprecisão dos atributos na consulta, apresenta-se o conceito de *skyline* ao nível de dominância p ou *p-skyline*.

Dado um conjunto de dados espaciais D (tal como os 13 hotéis apresentados no exemplo anterior), define-se como *p-skyline* em D, o subconjunto $S_p \in D$ formado pelos pontos p_i para os quais a probabilidade de não serem dominados por qualquer dos demais pontos de D é de pelo menos p. De acordo com essa definição e considerando erros posicionais cujos valores modulares podem assumir qualquer valor real positivo, tem-se $S_0 = D$ e $S_1 = \emptyset$, pois qualquer ponto de D possuiria uma probabilidade diferente de zero de pertencer à S, assim como nenhum ponto teria 100% de probabilidade de estar em S.

2.3 Junção Espacial

Antes de definir junção espacial, será definido o conceito de junção. Dados dois conjuntos de dados A e B, chama-se junção em relação a um atributo x comum aos dois conjuntos, o subconjunto do produto cartesiano $(A \text{ x } B)$ denotado por $(A \text{ x } B)_x$, contendo todas as tuplas (a, b) com a pertencente a A e b pertencente a B, tais que a e b apresentam

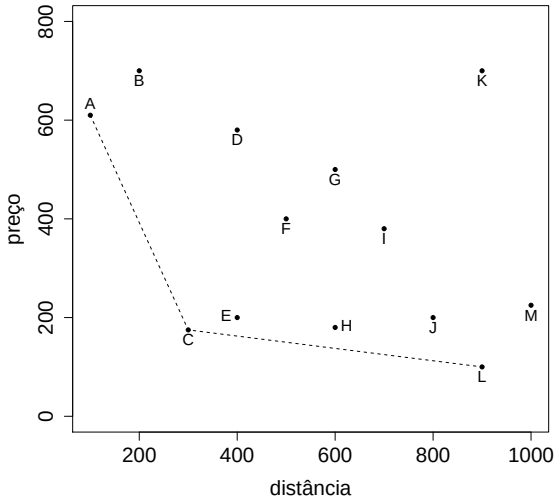

Figura 2.2: *Consulta skyline: cenário 1.*

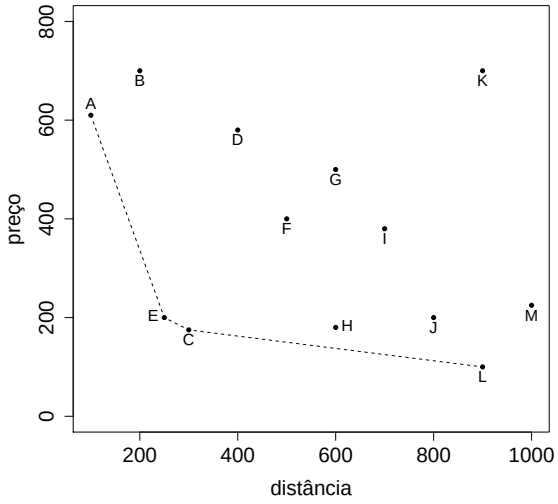

Figura 2.3: *Consulta skyline: cenário 2.*

um casamento em relação a x. A regra utilizada para definir se há o casamento entre as duas tuplas é chamada de *predicado* da junção.

Em muitas aplicações é utilizado o predicado de *igualdade*, ou seja, diz-se que a e b casam se $a = b$. Por exemplo, uma dada consulta a um banco de dados pode ter como objetivo retornar uma tabela com os nomes e telefones dos clientes com cadastro em uma loja, utilizando, para isso, dois conjuntos de dados: $A = \{cpf, telefones\}$ e $B = \{cpf, nomes\_das\_pessoas\}$. Neste caso, uma junção das tabelas pelo atributo *cpf* cumpre com o objetivo.

No caso específico em que há a presença de atributos espaciais, costuma-se empregar predicados diferentes dos utilizados para avaliar o casamento entre atributos textuais ou numéricos. São exemplos de predicados comumente considerados na junção de atributos espaciais: **interseção** e **estar a no máximo *x* metros de distância**. Para junções desse tipo sobre atributos espaciais, dá-se o nome de **junção espacial**. O escopo deste trabalho se restringe ao predicado de interseção. Nos cinco cenários apresentados pela Figura 2.4, há interseção em B, C, D e E. Portanto, diz-se que há casamento em relação ao predicado de interseção nesses cenários.

Quando há incerteza nos dados espaciais, tanto em relação à existência quanto em relação à posição dos objetos, faz-se necessário o conceito de junção espacial probabilística (JEP). Chama-se de junção espacial probabilística com probabilidade de corte p para dois conjuntos de dados A e B e atributo espacial x, o subconjunto do produto cartesiano $(A \text{ x } B)$ contendo todas as tuplas (a, b) com a pertencente a A e b pertencente a B, tais que a probabilidade de haver um casamento entre a e b em relação ao atributo espacial x é no mínimo igual a p.

Ao contrário das versões determinísticas da junção espacial, onde o veredito na avaliação do predicado é booleano, isto é, verdadeiro ou falso, na versão probabilística apenas se pode ter uma **confiança** em tal veredito. Assim, não se pode assegurar que a probabilidade de interseção é maior ou igual a p, mas que existe uma confiança γ que essa afirmação seja verdadeira. Assim, supondo haver erro posicional nas coordenadas dos polígonos da Figura 2.4, não se pode decidir com 100% de confiança se há casamento em qualquer dos cinco cenários apresentados. Contudo, assumindo que tais erros possam ser modelados com a mesma distribuição de probalidade, é razoável esperar que a probabilidade de casamento no cenário B seja maior que no cenário A. Uma JEP deve ser capaz de avaliar tal probabilidade e retornar casamento apenas para os casos em que esta superar o limiar p estabelecido.

A natureza probabilística inerente de uma solução JEP conduz a erros de julgamento, como falsos positivos e falsos negativos. Para exemplificar, tomando $p = 0,90$, haverá um falso negativo quando a **verdadeira** probabilidade de interseção for de pelo menos 90%, mas a JEP não retornar o casamento. Reciprocamente, a JEP incorrerá em falso positivo se retornar casamento para dois objetivos espaciais cuja **verdadeira** probabilidade de interseção seja inferior a 90%. Em uma solução JEP, a eficácia pode ser medida como a proporção de avaliações corretas do predicado da junção para todos os pares de geometrias ou para uma amostra desse conjunto. No capítulo 4, a eficácia dos métodos testados é avaliada para pares de geometrias cuja verdadeira probabilidade de interseção está situada em um raio R de distância de p.

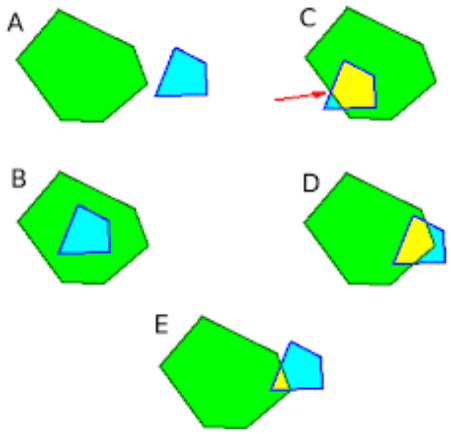

Figura 2.4: *Avaliação do predicado de interseção entre polígonos.*

2.4 Revisão da literatura para a consulta Skyline

Desde de sua introdução em Borzsony et al [2], o processamento de consultas *skylines* tem recebido considerável atenção na área de base de dados multidimensionais. Vários algoritmos para obter *skylines* têm sido propostos. Tan et al [40] usam estruturas

auxiliares para o cálculo progressivo da *skyline*, Kossmann et al [19] apresentam um algoritmo de vizinho mais próximo, Papadias et al [30] introduzem o algoritmo *branch and bound skyline* (BBS) e Chomicki et al ([6] e [7]) propõem o algoritmo *sort-filter-skyline* (SFS), o qual atua alavancando listas pré-ordenadas, bem como uma ordenação por eliminação linear para a *skyline*.

O conceito de consulta *skyline* espacial foi introduzido no trabalho de [36]. Dado um conjunto de pontos P e um conjunto de consultas Q, a cada ponto $p \in P$ se deriva um número de atributos espaciais correspondentes às suas distâncias para os ponto de consulta. [42] propôs o algoritmo *branch and bound farthest search* (BBFS), comparando e demonstrando sua superioridade em relação ao *threshold farthest spatial skyline* (TFSS). Algoritmos eficientes para a TFSS usando distância euclidiana haviam sido propostos em [39, 20]. [38] desenvolveu um algoritmo usando a distância de Manhattan (também conhecida como distância do taxista), a qual se aproxima da menor distância entre dois pontos medida pelas vias urbanas e rodoviárias em áreas bem conectadas.

Khalefa et al [18] apresentaram uma solução para dados incompletos, i.e, quando há dado faltante em alguma das dimensões consideradas na consulta. Para tanto, os autores generalizaram o critério de dominância como a seguir. Dado quaisquer dois pontos P e Q, que podem ter dimensões incompletas, o ponto P domina Q se as duas condições seguintes forem válidas:

1. existe pelo menos uma dimensão i onde tanto $P[i]$ e $Q[i]$ são conhecidos, e $P[i] < Q[i]$;
2. para todas as demais dimensões j, $P[j]$ ou $Q[j]$ são desconhecidos ou $P[j] \leq Q[j]$. Aqui, os símbolos $<$ e $\leq$ denotam o sentido preferencial de optimalidade, o qual pode ser inclusive o de maximização.

A definição tradicional para dados completos se torna um caso particular da definição proposta por Khalefa et al. Lofi et al [25] aborda o problema dos dados incompletos sobre uma perspectiva diferente. Os autores propõem que as tuplas com maior potencial para degenerar a qualidade dos resultados sejam compartilhadas com colaboradores em tempo de execução para que esses resolvam o problema de dados faltantes.

Huang et al [15] introduzem a *skyline* contínua sobre dados móveis (porém precisos). Zhang et al [44] apresentam técnicas que permitem realizar inferência de maneira eficiente tanto nas posições incertas atuais como nas futuras dos objetos. Um novo modelo *skyline* probabilístico é proposto por Ding et al (2014), onde um objeto incerto pode assumir uma probabilidade de pertinência na *skyline* em um certo ponto no tempo.

Pei et al [32] calculam *skylines* para dados incertos. Em seu contexto, incerteza significa que mais de uma instância está disponível para cada atributo para os vários

objetos em avaliação. Os autores citam, como exemplo, dados relativos à performance de jogadores da NBA. Para cada jogador, foram coletadas estatísticas como número de assistências e número de rebotes. Quanto maior o valor reportado em cada uma dessas estatísticas, melhor é considerado o jogador.

Como o desempenho dos jogadores varia de um jogo para o outro, cada um deles possui valores diferentes para a mesma estatística. Para resolver o problema, uma alternativa, conforme mencionam os autores, seria substituir os vários valores de cada atributo por suas médias para cada jogador. Dessa forma, um jogador *A* apresentaria uma única medida para rebote, a qual seria a média dos seus rebotes nas várias partidas reportadas no conjunto de dados. Portanto, uma solução *skyline* tradicional seria suficiente para resolver o problema de se definir os melhores jogadores do campeonato. Contudo, ao substituir os valores por sua média, perde-se a possibilidade de se calcular a **probabilidade** de que um dado jogador de fato esteja entre os melhores. Ao considerar todas as instâncias de cada atributo, Pei et al derivam tais probabilidades e introduzem à literatura da área o conceito de *skyline* probabilística.

2.5 Revisão da literatura para a junção espacial

Existem vários métodos para executar a junção espacial usando ou não conjuntos de dados indexados. Se nenhum conjunto de dados é indexado, um algoritmo *nested-loop* de junção [27], algoritmo *plane-sweep* [1, 17] ou junção espacial baseada em particionamento [26, 31] podem ser usados. Se apenas um conjunto de dados é indexado, então a junção espacial pode ser realizada utilizando um índice de laço aninhado [10, 24] o qual requer um índice IA para o conjunto de dados A, além de um laço para o conjunto de dados B e consultas IA para cada objeto. Se ambos (A e B) são indexados com R-Trees [14], um percurso síncrono [3, 15] pode ser usado para realizar a operação. Esse processo percorre recursivamente as árvores até o nível das folhas onde os objetos são comparados. No entanto, estas abordagens não lidam com junção espacial probabilítica.

Vários trabalhos [41, 33, 41, 43] têm sido propostos para lidar com junções espaciais em objetos em movimento. No entanto, esse tipo de aplicação comumente impõe mais restrições sobre a eficiência do que sobre a precisão, uma vez que o processamento normalmente deve ocorrer em tempo real. Esse não é o cenário para o qual a nossa solução é projetada. Nosso método requer que a incerteza relacionada à posição possa ser modelada probabilisticamente e que a solução final atenda aos três requisitos alvo deste trabalho: acurárica, eficiência e generalidade. Assim sendo, as aplicações que se beneficiariam da nossa solução são aquelas que precisam lidar com erros nos dados relacionados à técnica de georreferenciamento utilizada.

Junção espacial em dados com incerteza existencial foram exploradas por Dai et al [8] e em Ljosa e Singh [23] se apresenta uma abordagem capaz de lidar tanto com incerteza existencial como posional, fazendo uso de função de pontuação na qual os dois tipos de incerteza são ambos levados em consideração. Openshaw [29] propôs a utilização do Método de Monte Carlo (MMC) para calcular as probabilidades de interseção de geometrias com incerteza posicional. Chamaremos esse método de Junção Espacial Aleatória (JEA) e será testada e discutida no capítulo 4.

Na JEP proposta por Ni et al [28], o erro é modelado usando uma distribuição Circular Normal. Devido a algumas especificidades e propriedades bem conhecidas da distribuição Normal, os autores puderam adaptar as etapas de filtragem e refinamento para atingir tanto precisão e eficiência. Chamaremos essa abordagem de Junção Espacial Circular e Normal (JECN) e mais sobre como esta se compara com nossa abordagem é comentado no capítulo de resultados.

A superioridade de JECN sobre JEA é provada por Ni et al [28] em relação às duas dimensões citadas. No entanto, como JEA pode lidar com erros provenientes de várias FDPs e JECN é projetada para funcionar adequadamente apenas com a distribuição Circular Normal para os erros, tem-se que JEA é mais generalista que JECN. Como estamos propondo uma solução - JEPMCP - em que generalidade é um dos requisitos, JECN não é uma competidora direta, mas JEA sim. Portanto, JEPMCP será testada apenas contra JEA. Uma discussão é conduzida no capítulo 4 sobre os cenários em que JECN é a melhor escolha para se realizar uma JEP.

No caso geral, isto é, quando a distribuição pode ser qualquer outra além da Normal, realizar adaptações como as conduzidas por Ni et al não é uma tarefa trivial. Por outro lado, MMC é uma técnica computacionalmente cara. Assim sendo, nenhum deles é projetado para atender bem simultaneamente aos três critérios exigidos do nosso trabalho. Para atingir esse objetivo, a JEPMCP propõe uma adaptação no MMC apresentado no capítulo 3. Chamamos essa abordagem de Método de Monte Carlo Progressivo (MMCP). Além disso, no capítulo 3 é mostrado como o conceito de menor retângulo envolvente pode ser extendido naturalmente para o caso probabilístico.

2.6 Conclusão

Nesse capítulo foram apresentadas as definições tradicionais de consulta *skyline* e junção espacial bem como as suas versões probabilísticas. Também foi discutido o conceito de incerteza e sua relação com as operações espaciais probabilísticas existentes na literatura. Nesse sentido, o tipo de incerteza que lida nosso trabalho foi estabelecido, a saber a incerteza posicional. No próximo capítulo serão apresentadas as duas propostas de nosso trabalho.

CAPÍTULO 3

Soluções Propostas

Conforme descrito previamente, as soluções apresentadas nesse trabalho foram projetadas para atenderem a três requisitos: Apresentamos a estrutura geral de soluções para operações espaciais robustas a coordenadas imprecisas, bem como as adaptações das principais heurísticas utilizadas em cada uma das operações criadas, especificamente. Na JEPMCP, a etapa de filtragem, comum em algoritmos de junção espacial e importante por reduzir o número de cálculos geométricos exaustivos, é adaptada para a JEP mediante a introdução do conceito de retângulo de confiança. Para o refinamento, propomos uma adaptação do MMC que chamamos de Método de Monte Carlo Progressivo (MMCP). O MMCP é capaz de reduzir o custo computacional do MMC ao mesmo tempo que garante o nível especificado de acurácia - representado pelo parâmetro γ da JEPMCP, contemplando dessa forma os requisitos de acurácia e eficiência. O requisito de generalidade é contemplado naturalmente nas soluções baseadas em simulações de Monte Carlo. Na pSkyEMC, utiliza-se o MMC para se avaliar as probabilidades de dominância para qualquer PDF (generalidade) e a aplicação direta de um lema apresentado por Pei et al [32] para economizar processamento (eficiência).

A Seção 3.1 discorre sobre a estrutura geral de soluções para operações espaciais robustas à imprecisão das coordenadas geográficas. A Seção 3.2 mostra como as simulações do erro posicional são realizadas e 3.3 apresenta a adaptação proposta para o MMC de modo a lidar de maneira eficiente e eficaz na avaliação de condições probabilísticas. Mais especificamente, o método permite decidir o valor de verdade da condição $q \geq p$, em que p é uma probabilidade de corte e q é uma probabilidade de sucesso. Nas Seções 3.4 e 3.5, o algoritmo do MMCP é aplicado no desenvolvimento das soluções JEPMCP e pSkyEMC, que lidam respectivamente com junção espacial probabilística e *p-Skyline*. Na Seção 3.6 também é apresentada outra contribuição do presente trabalho para aumentar a eficiência de soluções espaciais com incerteza nas coordenadas: o conceito de **retângulo de confiança** (RC), bem como uma proposta de construção de um RC que contemple uma ampla gama de distribuições de probabilidade para os erros posicionais.

3.1 Discussão sobre a estrutura geral de soluções espaciais robustas à imprecisão nas coordenadas

A estrutura geral de soluções proposta nesse trabalho consiste da conjugação de: 1) simulações de Monte Carlo e 2) adaptações de heurísticas utilizadas nas versões determinísticas de cada operação espacial. A aplicação conjungada desses mecanismos permite que as soluções criadas atinjam níveis satisfatórios de generalidade, eficácia e eficiência, conforme demonstrado por resultados teóricos e empíricos apresentados nesse e no próximo capítulo.

O MMC é o responsável pela generalidade das soluções, permitindo que essas contemplem qualquer FDP associada aos erros nas coordenadas. Por outro lado, o MMC também eleva consideravelmente o custo computacional dos algoritmos. Por esse motivo, é imprescindível que heurísticas sejam empregadas para evitar que o custo máximo associado ao MMC seja atingido. Nesse sentido, o presente trabalho apresentou duas contribuições: o MMCP e o rêtangulo de confiança. Ambas são responsáveis pela redução do número total de simulações a serem realizadas nas soluções desenvolvidas. Além disso, aproveitamos resultados obtidos por outros autores para prover uma heurística para a consulta *skyline* probabilística apresentada.

Trabalhos futuros que busquem aplicar a presente estrutura de soluções a outras operações espaciais devem considerar a aplicação do MMCP para reduzir o número de simulações para cada avaliação condicional, bem como o emprego do uso do retângulo de confiança para evitar até mesmo que qualquer simulação seja executada desnecessariamente.

3.2 Simulações de Monte Carlo

MMC se refere a qualquer método na inferência estatística ou análise númerica que empregue o uso de simulações [34]. São conhecidos na literatura sua aplicação em diversos contextos, tais como na obtenção de estimativas para integrais definidas, valores médios de uma função, cálculo de probabilidades, dentre outros. Para vários exemplos do uso do método, veja o trabalho de Rizzo [34].

Em nosso trabalho, o MMC é empregado em dois contextos: 1) para estimar as **probabilidades** de interseção entre duas geometrias; 2) para estimar as probabilidades de **dominância** entre dois pontos assumindo conhecida (ou estimada) sua função de densidade de probabilidade (*FDP*).

Para que o MMC possa ser aplicado de maneira adequada, as simulações devem representar bem o fenômento estudado. A seguir, será descrito o procedimento empregado nesse trabalho para simular o erro posicional. Cada simulação consiste em substituir um

ponto $P = (x, y)$ de uma geometria g por outra coordenada Q que **simula** sua posição real. Para tanto, Q é obtida a partir do deslocamento de P na direção de um vetor v gerado com um ângulo θ, tal que $\theta \sim Uniforme(0, 2\pi)$ e com uma **norma** r gerada de acordo com a FDP que modela o erro posicional.

A Figura 3.1 ilustra o procedimento usado na simulação. A escolha da distribuição Uniforme(0, 2π) garante que o erro seja *igualmente distribuído* em todas as direções, i.e. que não haja uma direção mais provável para ele. Para calcular o vetor de deslocamento v são usadas as componentes v_x e v_y obtidas a partir das relações trigonométricas: $v_x = r.cos(\theta)$ e $v_y = r.sen(\theta)$. Finalmente Q é obtida com a equação $Q = (x + v_x, y + v_y)$. Esse procedimento garante que cada ponto é deslocado em uma direção aleatória e para outro ponto cuja distância do original é dada pela FDP escolhida.

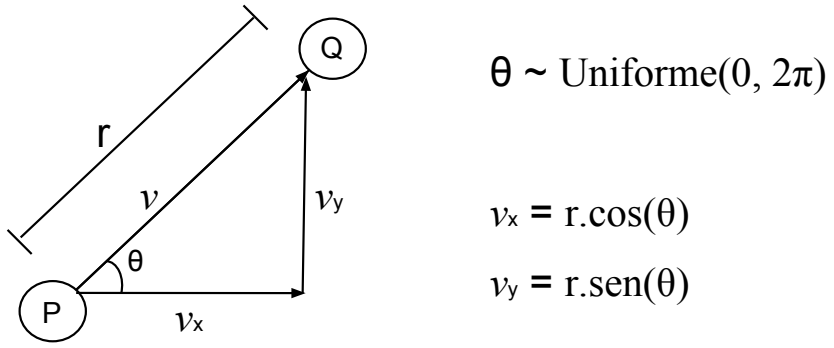

Figura 3.1: *Simulação da real posição do ponto.*

A estimativa de Monte Carlo $\hat{q}$ da probabilidade de interseção q entre duas geometrias - g e h - é obtida pela simulação de n deslocamentos em g e h mediante a fórmula $\hat{q} = \frac{x}{n}$, onde x é o número de simulações nas quais g e h se interceptam. A escolha se justifica pelo fato de que $\hat{q}$ é um estimador não viciado de variância mínima para q. Mais especificamente, tem-se que $E(\hat{q}) = q$ e $Var(\hat{q}) = \frac{q(1-q)}{n}$. Além disso, $Var(\hat{q}) \to 0$ quando $n \to \infty$.

A estimativa obtida é usada para avaliar a condição: $q \geq p$. Caso essa condição seja verdadeira, a junção espacial retorna o casamento entre as geometrias envolvidas. Caso contrário, não há casamento. Contudo, como $\hat{q}$ é uma estimativa, pode-se haver um equívoco na avaliação dessa condição. O Teorema Central do Limite garante que quando n tende a infinito, $\hat{q}$ tende a q. Portanto, a avaliação da condição pode ser tão precisa quanto se queira, bastando aplicar um número suficientemente grande de simulações. Contudo, esse número pode ser computacionalmente impraticável. A próxima seção mostra como o proposto MMCP resolve o problema de se conseguir um número **suficiente** de simulações para que a condição possa ser avaliada com eficácia e de maneira computacionalmente eficiente.

3.3 Método de Monte Carlo Progressivo

O MMCP objetiva estalecer um procedimento que aplica o MMC de uma maneira eficiente e efetiva. Consiste na realização do número de simulações necessárias para se avaliar a condição de interesse, no caso $q \geq p$, com o nível de confiança exigido.

Após realizadas as simulações, pode-se obter a estimativa $\hat{q}$ da probabilidade de interseção q entre duas geometrias quaisquer a e b, $a \in A$, $b \in B$. Fixado um número n de simulações, obtem-se a margem de erro ε para $\hat{q}$, i.e, a distância máxima que $\hat{q}$ se encontra da verdadeira probabilidade de interseção q com nível de confiança γ. Portanto, sabe-se que a probabilidade de q pertencer ao intervalo $(\hat{q}-\varepsilon, \hat{q}+\varepsilon)$ é igual a γ. Logo, baseado nesse intervalo de confiança, pode-se adotar a seguinte regra:

1. Caso $p < \hat{q}-\varepsilon$, decide-se com nível de confiança γ que não há casamento.
2. Caso $p \geq \hat{q}+\varepsilon$, decide-se com nível de confiança γ que há casamento.
3. Senão, não se assegura nível de confiança γ nem a favor, nem contra o casamento. Ainda assim, pode-se tomar uma decisão baseado no valor verdade de $(\hat{q} \geq p)$ com um nível de confiança menor que γ .

Por motivos de eficiência, o número n de simulações não pode ser demasiado grande. Contudo, devido à restrição de acurácia, também não pode ser muito pequeno. Por exemplo, $n = 50$ pode ser suficiente para garantir um tempo de processamento em um patamar aceitável pelo usuário, mas ser insuficiente para que o veredito de casamento seja avaliado com o nível de eficácia desejado. Também pode ocorrer o contrário: $n = 50$ ser **mais do que suficiente** para garantir a acurácia desejada, o que se constituiria como um desperdício de processamento. O MMCP resolve esse problema controlando o número de simulações de Monte Carlo utilizadas para se avaliar se $(q \geq p)$. Para isso, ao invés de executar o número total n simulações para cada par de candidatos, o MMCP executa um número mínimo $m < n$ em cada passo. Assim, caso $n = 1000$, m pode ser 40, por exemplo. Um único lote de $m = 40$ pode ser suficiente para decidir com nível de confiança desejado o valor verdade da condição $(q \geq p)$. Caso isso ocorra, o procedimento estaria economizando 960 simulações sem comprometer o nível de acurácia desejado da solução. Caso não, uma novo lote de $m = 40$ seria executado e as 80 simulações resultantes (40 no primeiro lote mais 40 no segundo) seriam novamente usadas para avaliar $(q \geq p)$. O procedimento se repete até que o valor verdade da expressão possa ser calculado com o nível de confiança especificado. O algoritmo do MMCP é apresentado à seguir.

Algoritmo 3.1: Método de Monte Carlo Progressivo.

Data: a e b: duas geometrias;
p: probabilidade de corte;
m: o tamanho de cada lote de simulações;
n_{max}: o número máximo de simulações a serem executadas
Result: Retorna VERDADEIRO se a proporção de sucessos for superior a p.

1 Inicialize com zero os contadores n do número de realizações de Monte Carlo $(n = 0)$.
2 Desloque m vezes as geometrias a e b.
3 Atualize n em m unidades $(n \leftarrow n + m)$.
4 Calcule a proporção $\hat{q}$ de **sucesso** nas n simulações.
5 Calcule o intervalo de confiança (IC) para q, i.e.,

$$IC(q,\gamma) = \left[\hat{q} - t_c\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}, \hat{q} + t_c\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}\right]$$

onde t_c é o quantil $(1+\gamma)/2$ da distrubição t-Student com $n-1$ graus de liberdade.
6 Se $p \in IC(q,\gamma)$ e n não excedeu o número máximo n_{max} de simulações permitidas, repita os passos de 2 a 5.
7 Se $\hat{q} \geq p$ retorne **VERDADEIRO**, senão retorne **FALSO**.

O MMCP executa lotes de m simulações (linha 2) sempre que o número máximo de iterações não for atingido e a eficácia desejada não for alcançada. No final de cada lote, o intervalo de confiança IC para q com o nível de confiança γ é calculado usando a fórmula para o **IC de uma proporção** fornecida pela teoria da inferência estatística (linha 5). Se $p \in IC(q,\gamma)$ (linha 6), não é possível decidir com uma confiança γ o valor verdade da condição $q \geq p$. Nesse caso, é necessário outro lote de simulações para se reduzir o tamanho de IC (vá para a linha 2). Caso contrário, o IC já é suficiente para decidir o valor verdade da condição com o nível de confiança γ. Dessa forma, a decisão pode ser tomada com a acurácia requerida pelo usuário sem a necessidade de mais simulações, economizando tempo de processamento.

O IC apresentado na linha 5 apresenta margem de erro $E = t_c\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}$, para a proporção estimada de sucesso - em que t_c é o quantil $\frac{1+\gamma}{2}$ da distribuição t-Student. Cada vez que novas simulações são realizadas, o tamanho do IC diminui - pois o valor n no numerado aumenta - até um ponto onde ou será pequeno o suficiente para decidir o valor verdade da condição $q \geq p$ com nível de confiança γ ou atingir o número máximo de iterações definido para o programa. A eficácia do método em avaliar corretamente a condição $(q \geq p)$ está relacionada ao valor de E. Fixado um nível de confiança γ, quanto

menor for E, maior será a chance da probabilidade de corte p não pertencer a IC e, portanto, maior a chance da decisão pelo valor verdade da condição $q \geq p$ - seja qual ele for - ser avaliado de maneira acurada.

O MMCP está alinhado com os três requisitos de projeto do presente trabalho. Por um lado, garante que o predicado seja avaliado com um nível de confiança γ, provendo acurácia à solução. Por outro, realiza um número suficiente de simulações, economizando processamento computacional. Além disso, é aplicável a qualquer distribuição de probabilidade em conformidade com o requisito de generalidade.

O tempo de processamento para a avaliação da condição $q \geq p$ para um dado par de geometrias cresce linearmente em função do número de simulações n, enquanto que a margem de erro decai proporcionalmente à $\sqrt{n}$. Logo, o custo computacional em se avaliar de forma acurada a condição $q \geq p$ cresce quadraticamente em função da quantidade de vezes k que a margem de erro precisa ser reduzida para que a avaliação possa ser determinada com o nível de confiança requerido. Contudo, ressalta-se que a necessidade de um valor k relativamente alto está diretamente associada a uma diferença entre p e q relativamente baixa. Em vários cenários práticos, uma diferença baixa o suficiente para exigir um k demasiado alto pode ser ignorada por parte do usuário. Esse usuário pode prover o relaxamento desejado simplesmente estabelecendo um valor mais baixo para γ.

As próximas seções apresentam aplicações do MMC e do MMCP no aumento da eficiência de técnicas espaciais que lidam com incerteza posional. Sua introdução nessas técnicas permitem que tais soluções avancem em relação às existentes na literatura no que diz respeito ao atendimento dos três requisitos de projeto definidos anteriormente.

3.4 Probabilidades na Skyline

Seja $U = \{A_1, ..., A_r\}$ uma coleção de objetos georreferenciados e $L = \{L_1, ..., L_d\}$ uma coleção de pontos de referência. Deseja-se obter o subconjunto S_p de U, tal que para todo $B_i \in S_p$ a probabilidade de B_i não ser dominado por qualquer dos pontos de U seja no mínimo igual a p. Esta seção mostra como calcular a probabilidade de um ponto A_i pertencer à S_p, formalmente $Pr(A_i \in S_p)$. Tem-se que

$$Pr(A_i \in S_p) = Pr\left[\bigcap_{i=1;i\neq j}^{r} (A_i \succ A_j)^c\right] \tag{3-1}$$

em que

- o símbolo $\succ$ denota dominância, i.e. $A_i \succ A_j$ significa que A_i é demoninado por A_j.
- $(A_i \succ A_j)^c$ é o evento complementar de $(A_i \succ A_j)$, i.e, o evento: "A_i **não** é demoninado por A_j".

Assumindo independência entre os eventos, pode-se escrever

$$Pr(A_i \in S_p) = \bigcap_{i=1;i\neq j}^{r} Pr[(A_i \succ A_j)^c] \tag{3-2}$$

em que $Pr(A_i \succ A_j)$ pode ser calculada por

$$Pr(A_i \succ A_j) = Pr\left[\bigcap_{k=1}^{d} A_{ik} \succ A_{jk}\right] = \bigcap_{k=1}^{d} Pr\left[A_{ik} \succ A_{jk}\right] \tag{3-3}$$

assumindo independência entre as dimensões e com $k = 1,...,d$ percorrendo cada uma das d distâncias para os pontos de referência. Para um dado k em particular, $(A_{ik} \succ A_{jk})$ é equivalente a $(A_{jk} < A_{ik})$, uma vez que a otimização se dá no sentido de **minimização** para os pontos de referência. Finalmente, $Pr(A_{jk} < A_{ik})$ é avaliada com o uso do MMC para um número m de simulações, conforme a equação 3-4.

$$Pr(A_{jk} < A_{ik}) = \#(A_{jk} < A_{ik})/m \tag{3-4}$$

O Lema 1 a seguir, apresentado e provado por Pei et al [32], pode ser usado para evitar o alto custo computacional envolvido na equação 3-2, devido principalmente às simulações requeridas na equação 3-4.

Lema 1. Seja $U = \{A_1,...,A_r\}$ uma coleção de objetos com coordenadas imprecisas e S_p a *p-skyline* para o conjunto U em relação a um conjunto qualquer de atributos espaciais. Se A_i domina A_j, então $Pr(A_i \in S_p) > Pr(A_j \in S_p)$.

3.5 Método pSkyEMC

A *p-Skyline* proposta neste trablaho - pSkyEMC - utiliza em seu algoritmo o Lema 1, as equações (3-1), (3-2) e o Método de Monte Carlo conduzido sobre a FDP utilizada para modelar os erros posicionais. O algoritmo é apresentado a seguir

Algoritmo 3.2: pSkyEMC

Data: S: um conjunto de n objetos georreferenciados;
R: um conjunto de coordenadas de referência;
p: probabilidade de corte;
μ: o erro médio;
σ: o desvio-padrão do erro.
Result: S_p: um subconjunto de S com os objetos cuja probabilidade de serem Pareto eficientes em relação aos atributos de distância derivados dos pontos em R seja de pelo menos p.

1 Calcule a distância de cada ponto para as coordenadas de referência, obtendo uma matriz $n \times d$, com n sendo o número de pontos e d o número de coordenadas de referência.
2 Inicialize os vetores P e Q que, respectivamente, armazenarão os pontos para os quais já se sabe serem pareto eficientes (com probabilidade pelo menos igual a p) e aqueles para os quais já se sabe que não o são.
3 **para** *cada ponto i* **faça**
4 Verifique se i já pertence a um dos conjuntos P ou Q. Caso sim, incremente i e realize a mesma verificação novamente. Caso não, siga os próximos passos.
5 Calcule a probabilidade de i ser dominado por j para cada $j = 1, ..., n$.
6 Estime a probabilidade q de i **não** ser dominado por quaisquer dos demais pontos, i.e, i ser Pareto eficiente (equações 3-1, 3-2).
7 **se** $q \geq p$ **então**
8 inclua i em P e aplique o Lema 1 para também incluir em P aqueles pontos que dominam deterministicamente i, i.e., os pontos j tais que a consulta *skyline* tradicional aponte que j domine i.
9 **fim**
10 **senão**
11 inclua i em Q e aplique o Lema 1 para também incluir em Q todos aqueles que são dominados por i.
12 **fim**
13 **fim**
14 Retorne P.

3.6 Método JEPMCP

Como as principais propostas em junção espacial (determinística), a JEPMCP apresenta as duas etapas clássicas: filtragem e refinamento. Para atingir os três requisitos

para os quais a solução é projetada, essas etapas são adaptadas. A etapa de filtragem é executada em uma versão probabilística e estendida do menor retângulo envolvente MRE, que chamamos de retângulo de confiança (RC), comumente usados para indexar dados espaciais. Na fase de refinamento, aplica-se o algoritmo MMCP apresentado na Seção 3.3 com **sucesso** significando a interseção entre duas geometrias. Na Seção 3.3, discorre-se sobre o ganho de eficiência fornecido por esse método. Contudo, o principal ganho em termos de eficiência é conseguido com o poder do filtro probabilístico fornecido pelo RC proposto. No restante da seção, o RC será definido, terá sua eficácia provada matematicamente e será fornecido um método para sua construção.

Definição 1: Dada uma geometria g e uma probabilidade de corte p, um retângulo de confiança (RC) para g com probabilidade de corte p é aquele que contém o MRE de g e, além disso, a probabilidade de conter o objeto real representado por g seja pelo menos igual a $\sqrt{1-p}$.

Uma primeira consequência da Definição 1 é que um RC não é único, havendo infinitos retângulos satifazendo a definição. No entanto, vamos construir um que seja válido para uma grande família de distribuições de probabilidade e tenha o menor tamanho necessário para garantir a eficácia da etapa de filtragem. Um RC menor fornece maior poder de filtragem, uma vez que impedirá que um número maior de pares seja avaliado na etapa de refinamento com o uso do MMCP, reduzindo custo computacionalmente. A Figura 3.2 exibe a geometria g, seu MRE e o valor do deslocamento d. Outra consequência da definição é expressa no teorema abaixo.

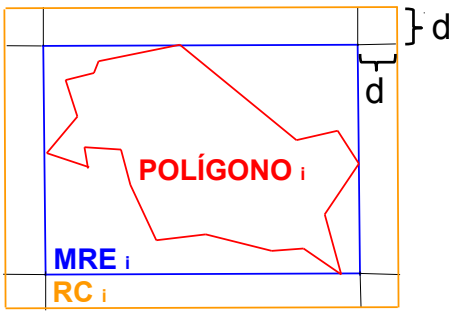

Figura 3.2: *Retângulo de confiança.*

Teorema 1: Sejam G e H retângulos de confiança para as geometrias g e h, respectivamente. Se G e H não se cruzarem, a probabilidade de interseção entre g e h é inferior a p.

Prova: Dado que G e H são RCs para g e h então, segue da definição que $Pr(g \subset G) \geq \sqrt{1-p}$, e $Pr(h \subset H) \geq \sqrt{1-p}$. Logo, assumindo independência na direção

do erro nas coordenadas de g e h,

$$Pr(g \subset G, h \subset H) = Pr(g \subset G).Pr(h \subset H) \geq (\sqrt{1-p})(\sqrt{1-p}) = 1-p$$

portanto,

$$Pr(g \not\subset G \cup h \not\subset H) < p$$

Como G e H não se cruzam, g e h apenas têm uma chance de interseção se pelo menos um deles não estiver contido em seu respectivo RC. No entanto, a última equação mostra que essa probabilidade é inferior a p. Logo, a probabilidade de interseção entre g e h é inferior a p como queríamos demonstrar.

Para construir um RC para uma dada geometria g, seu MRE é expandido por um fator d nas direções vertical e horizontal. Para cumprir com a definição de RC, o valor de d será fornecido pela desigualdade de Chebyshev. Essa desigualdade é válida para qualquer FDP integrável [35], o que vai ao encontro do requisito de generalidade das soluções do presente trabalho.

A desigualdade de Chebyshev declara que: Se X é uma variável aleatória integrável com média finita $\mu = E(X)$ e desvio-padrão σ, então para qualquer k > 0,

$$Pr(|X - E(X)| > k\sigma) < 1/k^2.$$

Uma consequência desta expressão é que, para X positivo, $Pr(X - E(X) > k\sigma) < 1/k^2$, o que implica em

$$Pr(X \leq E(X) + k\sigma) \geq \frac{k^2-1}{k^2}.$$

Na JEPMCP, X é o erro posicional. Consequentemente, a probabilidade do erro ser no máximo d = E(X) + kσ é de pelo menos $\frac{k^2-1}{k^2}$.

Para se criar retângulos de confiança para objetos espaciais de modo a garantir que a probabilidade de interseção destes seja inferior à probabilidade de corte p caso não ocorra a interseção destes retângulos, é suficiente tomar d de modo que para dois objetos A e B quaisquer

$$Pr(X \leq E(X) + k\sigma) > \sqrt{1-p}$$

Portanto, é suficiente igualar $\sqrt{1-p}$ à expressão $\frac{k^2-1}{k^2}$, o que implica em

$$k = \sqrt{\frac{1}{1-\sqrt{1-p}}}.$$

Consequentemente, para construir o RC, o valor d é aplicado em todas as coordenadas do MRE (x_{MRE_min}, y_{MRE_min}, x_{MRE_max}, y_{MRE_max}) nas direções horizontais e

verticais. Portanto, para um determinado p,

$$d = E(X) + \sigma\sqrt{\frac{1}{1-\sqrt{1-p}}}$$

As coordenadas do RC são dadas pelas seguintes coordenadas $P_{min} = (x_{min}, y_{min})$ e $P_{max} = (x_{max}, y_{max})$, com $x_{min} = x_{MRE\_min} - d$, $y_{min} = y_{MRE\_min} - d$, $x_{max} = x_{MRE\_max} + d$ e $y_{max} = y_{MRE\_max} + d$.

Os dois conjuntos de dados são indexados usando uma árvore-R com os RCs assumindo o mesmo papel dos MREs em uma solução de junção espacial tradicional (veja [31] para mais detalhes sobre a indexação dos objetos espaciais). Dados dois pares de geometrias a e b, $a \in A$, $b \in B$, a árvore é percorrida e em sua descida, caso os RCs não se interceptem, o par (a, b) já é descartado da junção. Aqueles pares que ainda mantêm o casamento dos respectivos RCs ao nível das folhas são avaliados com o método de refinamento proposto: MMCP.

3.7 Conclusão

Neste capítulo, apresentamos a estrutura geral de soluções para operações espaciais robutas a imprecisão nas coordenadas geográficas. Para ilustrá-la, propusemos alternativas probabilísticas a duas operações espaciais: *skyline* e junção espaciais. Diferentemente das soluções determinísticas correlatas, os métodos aqui apresentados retornam os objetos com probabilidade de "sucesso" superior a uma determinada probabilidade de corte. Para tanto, os métodos empregaram tanto heurísticas desenvolvidas em trabalhos correlatos como adaptações de procedimentos clássicos na literatura ao caso probabilístico. Como exemplo dessas adaptações, ressalta-se duas contrições do nosso trabalho: o MMCP e uma proposta para criação de um retângulo de confiança, os quais favoreceram o atendimento aos três requisitos de projetos considerados no presente trabalho. O desenvolvimento teórico conduzido foi respaldado por um lema provado em trabalhos correlatos e por um teorema apresentado e provado neste capítulo. A eficácia do retângulo de confiança foi provada matematicamente. O próximo capítulo demonstra experimentalmente a validade, acurácia e eficiência de nossa solução.

CAPÍTULO 4

Avaliação dos Resultados

4.1 p-skyline

A pSkyEMC foi avaliada com probabilidade de corte de 10% para dados de 36 escolas de Goiânia-GO mostradas nas Figuras 4.1 e 4.2. As escolas tiveram seus nomes removidos e são identificadas por números por questão de anonimicidade. A pSkyEMC é executada de modo a encontrar as escolas que sejam mais próximas a três locais de interesse, indicados com asterisco nessas figuras. Esses locais podem corresponder, por exemplo, à residência, local de trabalho e faculdade de uma dada pessoa. Dessa forma, a pSkyEMC retorna escolas que estão tanto quanto possível próximas aos três locais de interesse e descarta aquelas que são dominadas nesse conjunto de dados. Esse cenário ilustra como a pSkyEMC pode ser aplicada a serviços de recomendação de localidade baseados em múltiplos critérios. O erro considerado no experimento segue uma distribuição Normal com média 100 metros e desvio-padrão de 1000 metros. Esses valores são comumente encontrados em dados reais obtidos mediante um procedimento de geocodificação. Os três pontos considerados se encontram nas coordenadas: $P_1 = (-16,68; -49,21)$, $P_2 = (-16,71; -49,275)$ e $P_3 = (-16,695; -49,35)$.

A Figura 4.1 mostra o resultado obtido pela consulta *skyline* determinística e a Figura 4.2 o obtido pela pSkyEMC com $p = 0,10$ para as 36 escolas. Os resultados obtidos foram $S = \{1,3,9,10,14,15,17,18,20,26,27,28,30,31,35,36\}$ para a *skyline* determinística e $S_p = \{1,3,4,5,7,8,9,10,14,15,16,17,18,20,24,26,27,28,30,31,35,36\}$ para a pSkyEMC. Nota-se que S_p trouxe, além de todos os elementos presentes em S, também os pontos 4, 5, 7, 8, 16 e 24. Isso implica em um aumento de $37,5\%$ no número de opções a serem disponibilizadas pelo usuário, o que beneficiaria um serviço de recomendação de localidade desenvolvido com a pSkyEMC.

A pSkyEMC recuperou, em sua consulta, 6 escolas que não seriam retornadas como opções atendendo o critério de estarem próximas aos três pontos de referência citados. Essas 6 escolas podem ou não serem dominadas na realidade por alguma das demais escolas. No entanto, foram recuperadas pela consulta da pSkyEMC por essa ter identificado que a probabilidade de não serem dominadas estar acima da probabilidade de corte.

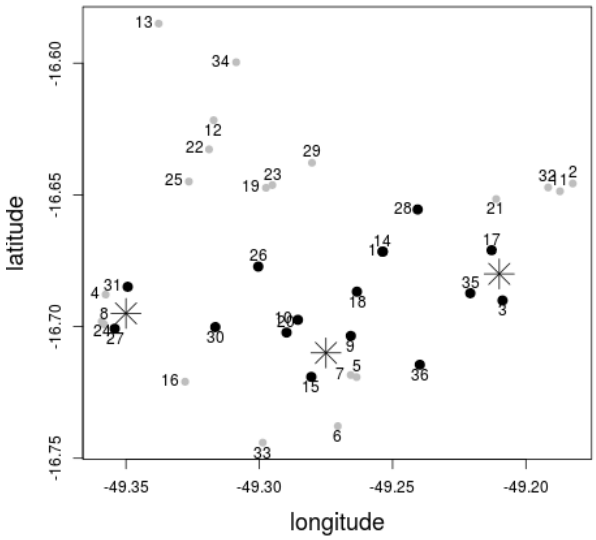

Figura 4.1: *Consulta skyline determinística.*

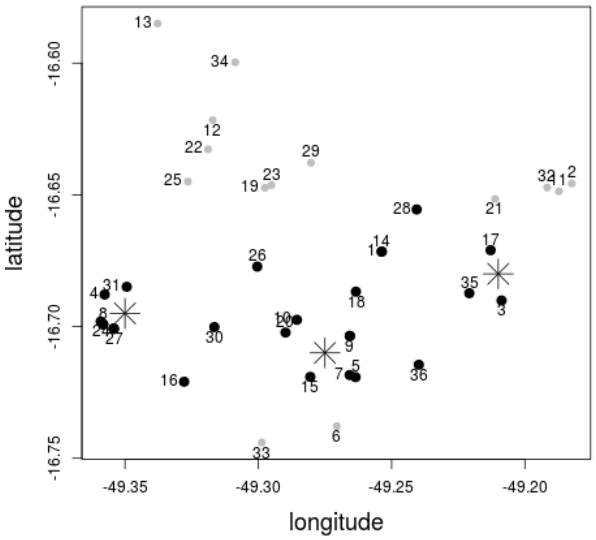

Figura 4.2: *p-skyline com* $p = 0,10$.

Em outras palavras, dado que $p = 0,10$, tem-se que não há uma probabilidade superior a 90% dessas escolas serem dominadas. Além disso, caso as 22 escolas retornadas em S_p possuam a mesma chance de serem a de maior utilidade para o usuário, então, S, ao excluir as 6 escolas citadas estaria incorrendo no risco de $\frac{6}{22} \simeq 27,3\%$ de não retornar a ele a melhor opção. A pSkyEMC evita que soluções potencialmente interessantes ao usuário sejam desnecessariamente descartadas, sendo, portanto, indicada para serviços de recomendação de localidade para dados com coordenadas geográficas imprecisas. Os demais métodos para realizar uma *p-skyline*, tais como o proposto por Pei et al [32] não são aplicáveis a esse contexto, pois não modelam o erro posicional dos objetos espaciais.

4.2 Junção espacial probabilística

4.2.1 Conjuntos de dados para teste

Para testar a junção espacial probabilística, serão usados três conjuntos de dados: 1) vegetação, 2) desmatamento e 3) queimada na pastagem, todos cobrindo o território do estado de Goiás. As figuras 4.3, 4.4 e 4.5 mostram as camadas de dados de vegetação, desmatamento e áreas de queimadas florestais em Goiás. Na Figura 4.3, o desmatamento e as geometrias de queimada quase não podem ser percebidas. Isso ilustra a diferença de escala entre os objetos espaciais representados nos conjuntos de dados citados. Nas Figuras 4.4 e 4.5, as áreas de queimada e desmatamento estão mais visíveis, bem como as interseções dessas com as áreas de vegetação. Algumas dessas interseções podem apenas ser uma consequência do erro posicional na camada de dados. Reciprocamente, devido

à imprecisão dos dados, algumas geometrias que não estão se interceptando na figura podem corresponder a objetos espaciais que se interceptam na realidade.

Para julgar os resultados obtidos pelas JEPs, foi construída um conjunto de referência que chamaremos de REF. O conjunto REF é composto pelas estimativas de probabilidade de interseção de cada par de geometrias (a, b), $a \in A$ e $b \in B$, em que A e B são os conjuntos avaliados na junção espacial. Para obter tais estimativas, aplicou-se $N = 500$ simulações de Monte Carlo para avaliação da junção espacial em cada possível par (a, b). A margem de erro E_0 para a probabilidade de interseção estimada q_0 para um dado par de geometrias (a, b) é dado por $E_0 = z_c\sqrt{\frac{p_0(1-p_0)}{500}}$. Assumindo o nível de confiança de 95%, o maior valor que E_0 pode assumir é $0,0447$ quando p_0 for $0,5$. À medida que p_0 se distancia de $0,5$, E_0 decresce. Por exemplo, para $p_0 < 0,1$ ou $p_0 > 0,9$, tem-se que $E_0 < 0,0263$. A subseção 4.2.2 discute em maiores detalhes como REF é utilizada para se avaliar as JEPs.

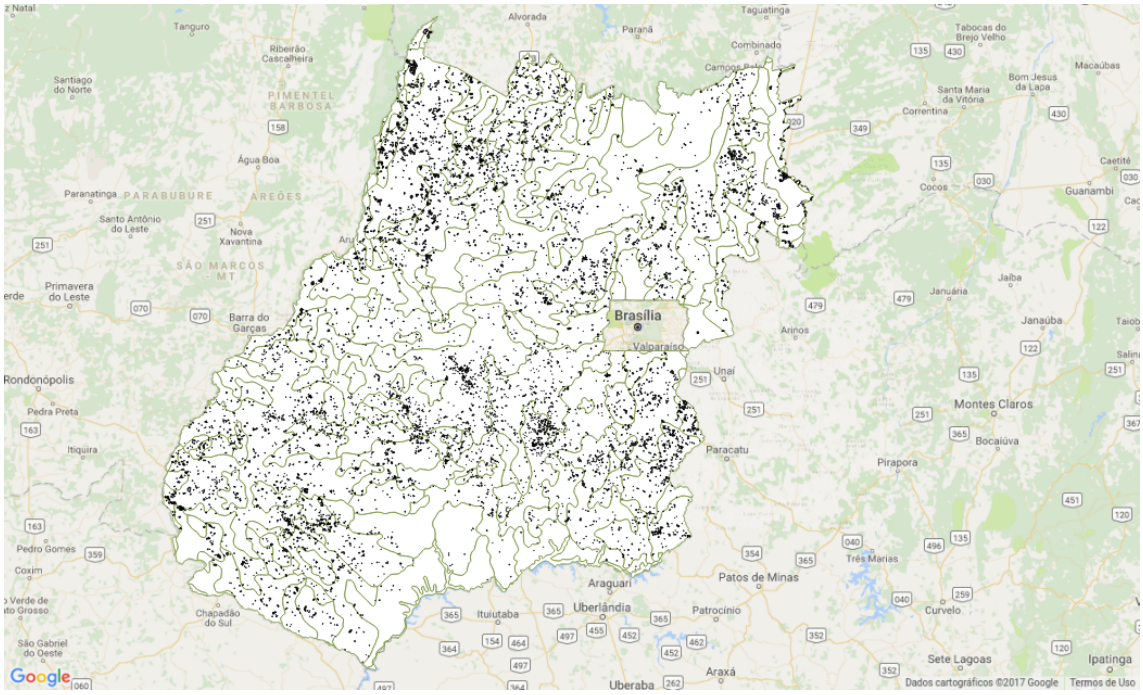

Figura 4.3: *Vegetação (branco, exceto o Distrito Federal), queimada (preto) e desmatamento (cinza) - nível de zoom de estado.*

Os métodos avaliados neste trabalho são: Junção Espacial Randômica (JEA) conforme proposto por [29] com $m = 150$ simulações, e nossa abordagem (JEPMCP) que é executada com duas configurações: 1) com $n_{max} = 150$, $m = 50$ simulações por etapa e nível de confiança $\gamma = 0,99$, que fornece uma confiança de 99% na avaliação do predicado; 2) $n_{max} = 1000$, $m = 50$. No cenário (1), observamos como JEPMCP se compara a JEA para o mesmo número limite de simulações (no caso, 150). No cenário (2), veremos como JEPMCP se compara a JEA quando permitimos que o número de simulações em JEPMCP

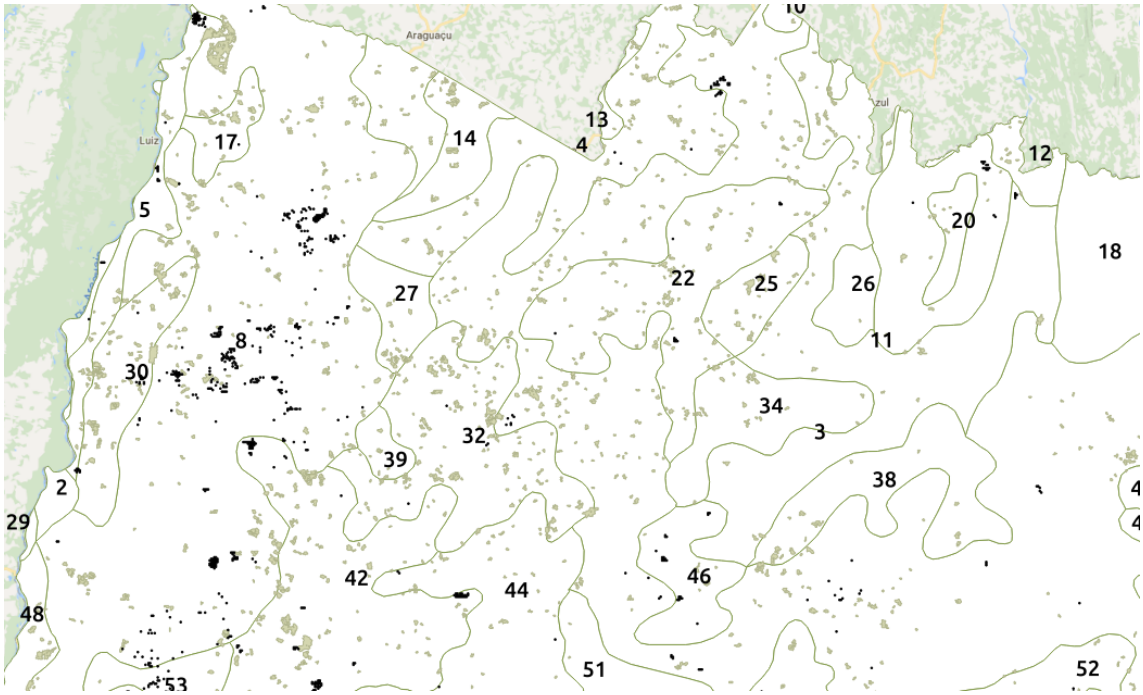

Figura 4.4: *Vegetação (branco), queimada (preto) e desmatamento (cinza) - nível de zoom das cidades.*

possa ser relativamente alto. Além disso, uma discussão é realizada ao final do capítulo sobre a Junção Espacial Circular Normal (JECN) apresentada por [28].

4.2.2 Métricas de desempenho e avaliação

Os métodos serão comparados em um cenário com erros Y seguindo uma distribuição Semi-Normal, definida por $Y = |X|$, com $X \sim N(200, 100^2)$. Os parâmetros de **média** e **desvio padrão** (necessários para a computação do RC) da distribuição Semi-Normal são: $E(X) = \sigma\sqrt{\frac{2}{\pi}}$ e $sd(X) = \sigma\sqrt{1-\frac{2}{\pi}}$.

As probabilidades de corte testadas são: 0,10, 0,20, 0,30, 0,40, 0,50, 0,60, 0,70, 0,80 e 0,90. O parâmetro de precisão é $\gamma = 0,99$, o tamanho dos lotes de simulação é $m = 50$ e o número máximo de simulações é $n_{max} = 150$.

Antes de apresentar as métricas utilizadas na comparação, será necessário definir alguns conceitos. Primeiramente, será chamado de vizinhança de raio R em torno de p e denotado por $V_R(p)$, o intervalo $(p-R, p+R)$. A Figura 4.6 mostra os intervalos usados para comparar os métodos, a saber, para cada valor de p, o intervalo $(p-R,\ p+R)$ sem o intervalo de margem de erro $(p-\varepsilon,\ p+\varepsilon)$ em torno de p. A exclusão do intervalo de margem de erro é aplicada para evitar que os métodos sejam avaliados em um raio de p para o qual não há confiança suficiente no sinal da expressão $(q-p)$, em que q é a probabilidade de interseção entre as duas geometrias avaliadas. Por exemplo, a estimativa $\hat{q}$ da probabilidade de interseção pode ser maior que p quando na realidade q seja

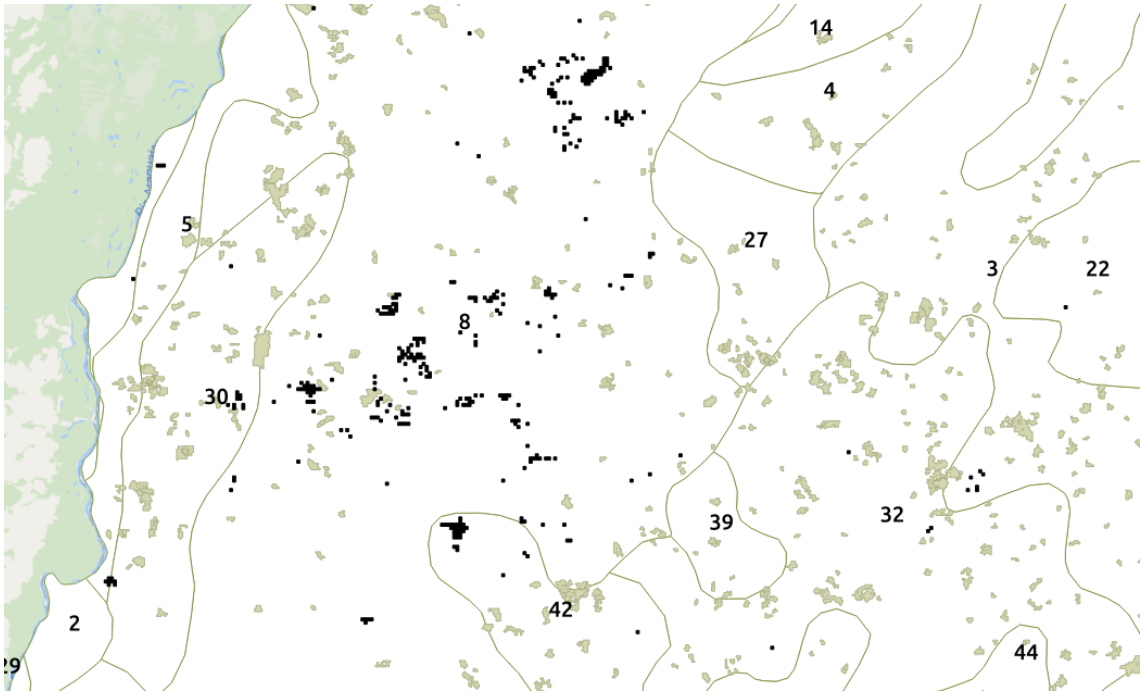

Figura 4.5: *Vegetação (branco), queimada (preto) e desmatamento (cinza) - nível de zoom local.*

menor que p. Nesse cenário um legítimo *verdadeiro positivo* poder ser considerado um *verdadeiro negativo* no conjunto de referência adotado para avaliar os métodos. É definido como vizinhança à esquerda de p com raio R, o intervalo $(p-R,\ p)$. Analogamente, $(p,\ p+R)$ é chamado de vizinhança à direita. A proporção de falsos negativos F_N em $V_R(p)$ é definida como

$$F_N = \frac{\#(-\infty, p)_{TEC}}{\#(p, p+R)_{REF}}$$

e a proporção de falsos positivos F_P em $V_R(p)$ como

$$F_P = \frac{\#(p, \infty)_{TEC}}{\#(p-R, p)_{REF}},$$

onde

- $\#(a,b)_{TEC}$ é o número de pares de geometrias cujas probabilidades de interseção estimadas pela técnica JEP utilizada pertencem ao intervalo (a, b).
- $\#(a,b)_{REF}$ é o número de pares de geometrias cujas probabilidades de interseção estimadas por REF pertencem ao intervalo (a, b).

Como já foi dito, dentro do intervalo de margem de erro em torno de p (veja Figura 4.6), a probabilidade de que a estimativa de REF esteja do lado errado com relação a p, ou seja, à direita quando na realidade está à esquerda e vice-versa - não é desprezível. Por esse motivo, apenas aqueles pares cujas probabilidades estão na **região de vizinhança segura**, mostrada com um traço mais forte na Figura 4.6 são usados para calcular as

proporções de falsos negativos e positivos seguros - S_{FP} e S_{FN}. Assim,

$$S_{FN} = \frac{\#(-\infty, p)_{TEC}}{\#(p+\varepsilon, p+R)_{REF}}.$$

$$S_{FP} = \frac{\#(p, \infty)_{TEC}}{\#(p-R, p-\varepsilon)_{REF}}.$$

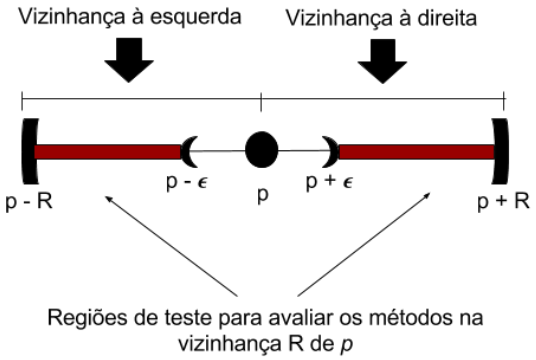

Figura 4.6: *Intervalos de teste para avaliar os métodos no vizinhança R de p.*

4.2.3 Apresentação e discussão dos resultados

A Tabela 4.1 apresenta resultados relativos às métricas S_{FN} e S_{FP}. A coluna *métrica* fornece a probabilidade de corte usada, i.e, se S_{FN} ou F_{FP} e em parêntesis traz a probabilidade de corte considerada. As colunas *JEA-150*, *JEPMCP-150* e *JEPMCP-1000* mostram os valores obtidos em cada métrica por JEA com 150 simulações e JEPMCP com o limite de 150 e 1000 simulações respectivamente.

Quando foi visto no Capítulo 2, JEPMCP se beneficia de não precisar executar todas as simulações estipuladas para conseguir avaliar com acurácia a condição de interesse. Devido a essa capacidade, JEPMCP ecomiza tempo em relação à JEA na avaliação do predicado para cada par de geometrias. Contudo, como o critério de parada para as simulações é probabilístico, pode ocorrer, em alguns casos que o algoritmo julgue equivocadamente que o número de simulações existentes já seja suficiente para avaliar o predicado da junção com precisão (risco que pode ser minimizado ao se aumentar o valor do parâmetro γ). A Tabela 4.1 mostra que para o limite 150, JEA venceu em seis cenários e JEPMCP em três. No entanto, ao ser comparada com JEPMCP com um limite de 1000 simulações, tem-se que JEA perde em cinco de seis cenários.

Para testar se a diferença observada na comparação JEA-150 versus JEPMCP-150 é estatisticamente significante ou se pode ser atribuída ao acaso, um teste de hipóteses é executado - o **teste de sinal** [37]. A coluna *sinal* da Tabela 4.1 apresenta as 18

métrica	JEA-150	JEPMCP-150	JEPMCP-1000	sinal 1	sinal 2
$S_{FN}(0,10)$	0,200	0,000	0,000	+	+
$S_{FN}(0,20)$	0,000	0,056	0,000	-	
$S_{FN}(0,30)$	0,000	0,000	0,000		
$S_{FN}(0,40)$	0,000	0,000	0,000		
$S_{FN}(0,50)$	0,000	0,333	0,167	-	-
$S_{FN}(0,60)$	0,000	0,000	0,000		
$S_{FN}(0,70)$	0,000	0,111	0,000	-	
$S_{FN}(0,80)$	0,133	0,200	0,000	-	+
$S_{FN}(0,90)$	0,000	0,000	0,000		
$S_{FP}(0,10)$	0,000	0,000	0,000		
$S_{FP}(0,20)$	0,000	0,045	0,000	-	
$S_{FP}(0,30)$	0,000	0,077	0,000	-	
$S_{FP}(0,40)$	0,100	0,100	0,000		+
$S_{FP}(0,50)$	0,000	0,000	0,000		
$S_{FP}(0,60)$	0,000	0,000	0,000		
$S_{FP}(0,70)$	0,200	0,000	0,000	+	+
$S_{FP}(0,80)$	0,000	0,000	0,000		
$S_{FP}(0,90)$	0,100	0,000	0,000	+	+

Tabela 4.1: *Comparação da eficácia de JEA e JEPMCP com sinais indicando se JEPMCP foi mais acurada que JEA para $n.max = 150$ e $n.max = 1000$ respectivamente, ("+" se JEPMCP performou melhor que JEA e "-" caso contrário).*

observações emparelhadas com um sinal: "+" se JEPMCP obteve melhor desempenho que JEA e "-" se apresentou desempenho pior. No caso de empate, nenhum sinal é emitido. As duas hipóteses estatísticas - nula (H_0) e alternativa (H_1) - que consideramos no teste são:

- "H_0: JEPMCP é tão eficaz quanto JEA" (a probabilidade de " + " é igual à de " - ";
- " H_1: JEPMCP é menos eficaz que JEA" (a probabilidade de "+" é inferior à de "-").

A estatística de teste **t** é: "número de +". Apenas os casos não empatados são considerados neste tipo de teste - resultando em um total de $n = 9$ casos. Sob a suposição H_0, t segue uma distribuição binomial com $n = 9$ ensaios e probabilidade de sucesso $p = \frac{1}{2}$.

Uma forma de avaliar se o valor observado t_{obs} da estatística t traz evidências a favor ou contra H_0 passa pelo cálculo do p-valor ou valor descritivo associado à t_{obs}. O p-valor associado ao valor observado de uma estatística é a probabilidade de se obter um valor igual ou mais extremo para ela quando H_0 é verdadeira [4]. Em outras palavras, o p-valor fornece a probabilidade de que um valor tão ou mais extremo que o observado pudesse ter sido observado caso H_0 seja verdadeira.

Um p-valor "baixo" indica que H_0 é provavelmente falsa. Normalmente, um p-valor inferior a 0,05 é considero baixo o suficiente para que se rejeite a hipótese H_0 em favor de H_1, pois nesse caso, as chances de que um valor tão extremo pudesse ser simplesmente obra do acaso seria inferior a 1 em 20. Como $t \sim Binomial(9;0,5)$ sob H_0, tem-se que o p-valor é dado por $Pr(t \leq 3|H_0) = 0,254$. Logo, o cenário 6 contra 3 não fornece evidência suficiente para a rejeição de H_0 mesmo com um nível de confiança de 0,80. Na verdade, caso JEPMCP-150 e JEA-150 sejam igualmente eficazes, espera-se que a diferença observada ocorra 1 em cada 4 experimentos. Portanto, não há evidências suficientes - sequer ao nível de 80% de confiança - que apontem para uma diferença estatisticamente significante entre a eficácia de ambos os métodos. Esse resultado está de acordo com a garantia probabilística oferecida pelo Algoritmo 3.1 no Capítulo 3.

A Figura 4.7 mostra a diferença no tempo de processamento para (c) JEA com 150 simulações e JEPMCP com (a) 150 e (b) 1000 simulações. O *hardware* utilizado na execução de ambos os métodos foi um processador Intel Core i5-4200U, CPU de 1.6GHz e com 4 *threads* em paralelo. A JEPMCP-150 gastou uma média de 64 segundos para realizar a junção desmatamento/vegetação contra 73 da JEPMCP-1000 e 257 da JEA-150. Em relação à junção queimada/vegetação, a economia foi de 35 para JEPMCP-150, 40 para a JEPMCP-1000, contra 92 da JEA-150.

Em ambos os cenários, a JEPMCP apresentou desempenho computacional expressivamente abaixo do tempo de processamento demandado pela JEA. Nota-se que o tempo de processamento aumentou apenas 14% ao se aumentar o limite de simulações na JEPMCP de 150 para 1000. Isso ocorre porque a maioria das avaliações do predicado para um dado par de geometrias podem ser revolvidas com o nível de confiança desejado usando poucos lotes de $m = 50$. JEA, por outro, tem seu tempo de processamento impactado linearmente com o aumento do número de simulações. Além disso, como mostra a Tabela 4.1, JEPMCP-1000 apresentou cinco resultados favoráveis em seis possíveis contra JEA-150. Sob a hipótese nula de igualdade de desempenho entre JEPMCP-1000 e JEA-150, tem-se o p-valor de 0,11 relativo a um teste unilateral para testar se JEPMCP-1000 performou melhor que JEA-150. Note que, o p-valor foi mais baixo para este teste que o anterior. Dado o p-valor obtido, tem-se que caso JEPMCP-1000 e JEA-150 tenham a mesma acurácia, a chance de termos observado a vantagem de cinco contra seis em favor de JEPMCP seria de 1 em cada 9 experimentos.

Em síntese, nota-se que, em relação aos resultados obtidos que: 1) pode-se aumentar a eficácia de JEPMCP de modo a deixá-la tão ou mais eficaz que JEA simplesmente aumentando o número máximo de simulações permitido pelo programa; 2) o incremento no número máximo de simulações não impacta drasticamente a performance de JEPMCP, permitindo que esta ainda tenha desempenho significativamente melhor que JEA. Logo, JEPMCP é preferível à JEA como método para executar uma junção espacial

probabilística, a dominando em relação à eficiência e acurácia. Na próxima, fazemos uma discussão final sobre os cenários.

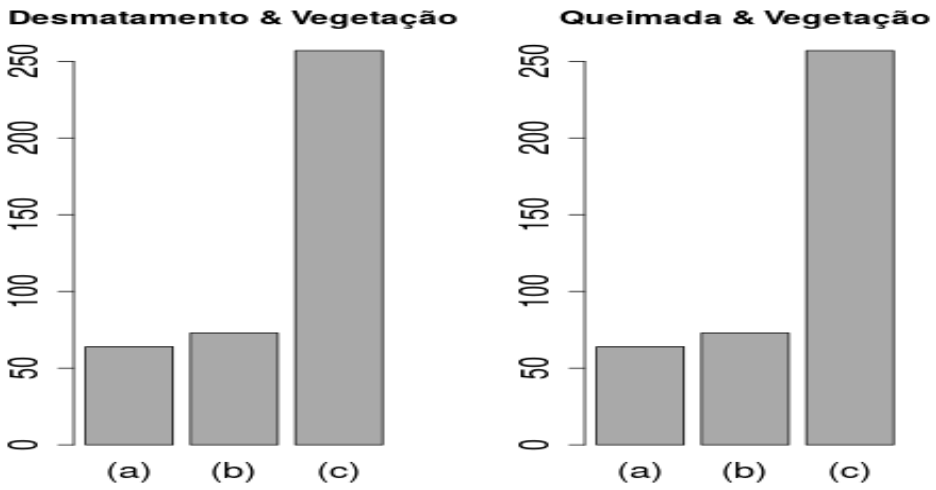

Figura 4.7: *Comparação do tempo de processamento entre JEA e JEPMCP: (a) JEPMCP-150, (b) JEPMCP-1000, (c) JEA-150.*

4.2.4 Conclusão

A JEPMCP demonstrou atender os requisitos para que foi projetada: generalidade, eficácia e eficiência. Além disso, sua comparação com a JEA se mostrou favorável. Como mencionado no Capítulo 1, as soluções presentes na literatura conseguem atender bem a um desses requisitos isoladamente. Por exemplo, JEA é generalista - poder ser executada com qualquer distribuição de probabilidade. Por outro lado, JECN é eficiente, porém por ser uma solução especialista na distribuição Circular Normal. JEPMCP é generalista quanto à FDP que modela os erros posicionais, e apresenta eficácia e eficiência ajustáveis, apresentando eficácia semelhante à JEA com um tempo de processamento melhor.

A decisão pela melhor solução para um dado caso real pode ser conduzida da seguinte maneira. Se os erros exibirem um padrão modelável pela distribuição Circular Normal, escolha JECN para obter a melhor acurácia e eficiência possíveis (solução especialista). Caso contrário, se os requisitos da aplicação em termos de tempo de processamento forem bastante restritivos (talvez devido ao grande volume de dados) **e** à distribuição de erros não for muito distante da Normal, sendo pelo menos simétrica, então o JECN também é uma boa opção para atender a exigência imposta pela restrição de eficiência, sem grandes prejuízos de acurácia. No entanto, se a distribuição definitivamente não é Circular Normal ou pelo menos uma simétrica que não apresente graves desvios da

Normal, então a JEPMCP é o método recomendado (solução generalista). Além disso, se a exigência por desempenho computacional não for rígida, a JEPMCP também é indicada para obter uma solução mais acurada do que JECN mesmo para distribuições próximas da Circular Normal.

CAPÍTULO 5

Conclusão

O presente trabalho demonstrou que operações espaciais, de maneira geral, podem ser adaptadas para funcionarem da maneira eficaz e eficiente em cenários em que os objetos espaciais apresentam imprecisão nas coordenadas. Para tanto, foi proposto uma estrutura geral de soluções e desenvolvidas duas como estudo de caso: *skyline* (pSkyEMC) e junção (JEPMCP) espaciais, respectivamente. Para que essas soluções atingissem os requisitos desejados de desempenho e eficácia, foi necessário que adaptássemos heurísticas determinísiticas para o caso probabilístico, seguindo assim a estrutura geral de soluções. Isso exigiu o desenvolvimento de duas novas ferramentas, as quais são contribuições de nosso trabalho para a área: construção de **retângulos de confiança** e do **Método de Monte Carlo Progressivo**. Esses dispositivos permitiram que a generalidade provida por simulações de Monte Carlo pudesse ser aplicada às operações espaciais de forma computacionalmente eficiente, eliminando muito do reconhecidamente alto custo computacional de simulações dessa natureza. As soluções finais permitem ao seus usuários que ajustem o balanço desejado entre eficicácia e eficiência, de acordo com as exigências específicas que uma dada aplicação que utilize o método possa demandar.

O lema proposto por Pei et al e apresentado no Capítulo 3 foi usado com sucesso pela pSkyEMC para economizar processamento. Os resultados da pSkyEMC mostraram que há um risco significativo de não retornar ao usuário uma solução potencialmente útil caso seja executada uma consulta *skyline* determinística ao invés de uma probabilística. O uso do lema conjugado com as simulações de Monte Carlo aplicados pelo procedimento permitiu a realização de consultas *skyline* de maneira mais eficaz para dados com coordenadas imprecisas.

Os experimentos mostraram que a JEPMCP pode ser: a) generalista em relação á distribuição dos erros posicionais; b) eficaz; e c) eficiente. Trabalhos correlatos existentes dão suporte apenas para uma distribuição de probabilidade ou exigem um tempo de processamento que inviabiliza sua aplicação prática para cenários mais desafiadores. Com a JEPMCP, além do cliente ter a seu dispor uma solução que não é dominada por qualquer uma presente na literatura quanto aos três requisitos mencionados anteriormente, ainda pode configurá-la de modo a aumentar ou sua acurácia ou sua eficiência. Dessa forma, a

JEPMCP se constitui como a única alternativa para JEPs (junções espaciais probabilísticas) que simultanemente: 1) funciona para diferentes distribuições de probabilidade e 2) computa com eficiência uma solução acurada, pondendo ainda o balanço entre esses dois requisitos ser ajustado pelo cliente.

O alto custo de simulações de Monte Carlo é mitigado tanto pela abordagem progressiva proposta como pelos filtragem realizada com os retângulos de confiança. A abordagem progressiva aplica apenas o número suficiente de simulações de Monte Carlo para se atender a exigência de acurácia expressa por γ, desde que não exceda o número máximo de simulações estipulado. O teorema proposto neste trabalho, apresentado e provado no Capítulo 3, garante que caso os retângulos de confiança de duas geometrias não se interceptem, a probabilidade de interseção entre tais geometrias é inferior a p e, portanto, não deve ser retornada na JEP. Esse resultado é o principal responsável pelo ganho de eficiência da JEPMCP.

Como trabalho futuro, vislumbramos a possibilidade de construir uma etapa de filtragem mais poderosa. Como efeito colateral, a acurácia também seria beneficiada, pois o ganho na eficiência permitiria ao cliente da solução usar valores de γ e $n.max$ maiores. Outra possibilidade é tentar adaptar as duas etapas da junção espacial às especificidades de uma família de distribuição - tal como a Família Exponencial, por exemplo. Isso manteria o requisito de generalidade (pois normalmente as famílias englobam um grande número de distribuições) enquanto avançaria em termos de eficiência e acurácia.

Referências Bibliográficas

[1] ARGE, L.; PROCOPIUC, O.; RAMASWAMY, S.; SUEL, T.; VITTER, J. S. **Scalable sweeping-based spatial join**. In: *VLDB*, volume 98, p. 570–581. Citeseer, 1998.

[2] BORZSONY, S.; KOSSMANN, D.; STOCKER, K. **The skyline operator**. In: *Data Engineering, 2001. Proceedings. 17th International Conference on*, p. 421–430. IEEE, 2001.

[3] BRINKHOFF, T.; KRIEGEL, H.-P.; SEEGER, B. **Efficient processing of spatial joins using R-trees**, volume 22. ACM, 1993.

[4] BUSSAB, W. D. O.; MORETTIN, P. A. **Estatística básica**. Saraiva, 2010.

[5] CHOMICKI, J.; CIACCIA, P.; MENEGHETTI, N. **Skyline queries, front and back**. *ACM SIGMOD Record*, 42(3):6–18, 2013.

[6] CHOMICKI, J.; GODFREY, P.; GRYZ, J.; LIANG, D. **Skyline with presorting**. In: *ICDE*, volume 3, p. 717–719, 2003.

[7] CHOMICKI, J.; GODFREY, P.; GRYZ, J.; LIANG, D. **Skyline with presorting: Theory and optimizations**. In: *Intelligent Information Processing and Web Mining*, p. 595–604. Springer, 2005.

[8] DAI, X.; YIU, M. L.; MAMOULIS, N.; TAO, Y.; VAITIS, M. **Probabilistic spatial queries on existentially uncertain data**. In: *International Symposium on Spatial and Temporal Databases*, p. 400–417. Springer, 2005.

[9] DEVENDRAN, A. A.; LAKSHMANAN, G. **A review on accuracy and uncertainty of spatial data and analyses with special reference to urban and hydrological modelling**. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2(8):171, 2014.

[10] ELMASRI, R. **Fundamentals of database systems**. Pearson Education India, 2008.

[11] FAURE, E.; DANJOU, A. M.; CLAVEL-CHAPELON, F.; BOUTRON-RUAULT, M.-C.; DOSSUS, L.; FERVERS, B. **Accuracy of two geocoding methods for geographic**

information system-based exposure assessment in epidemiological studies. *Environmental Health*, 16(1):15, 2017.

[12] FISHER, P. F. **Models of uncertainty in spatial data**. *Geographical information systems*, 1:191–205, 1999.

[13] FOOTE, K. E.; LYNCH, M. **Geographic information systems as an integrating technology: Context, concepts, and definitions**. *The Geographer's Craft Project*, 1995.

[14] GUTTMAN, A. **R-trees: a dynamic index structure for spatial searching**, volume 14. ACM, 1984.

[15] HUANG, Z.; LU, H.; OOI, B. C.; TUNG, A. K. **Continuous skyline queries for moving objects**. *Knowledge and Data Engineering, IEEE Transactions on*, 18(12):1645–1658, 2006.

[16] HUGHES, W. **Wide-area augmentation system performance analysis report**. *Federal Aviation Administration WAAS Test Team, Atlantic City, NJ*, 2002.

[17] JACOX, E. H.; SAMET, H. **Iterative spatial join**. *ACM Transactions on Database Systems (TODS)*, 28(3):230–256, 2003.

[18] KHALEFA, M. E.; MOKBEL, M. F.; LEVANDOSKI, J. J. **Skyline query processing for incomplete data**. In: *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, p. 556–565. IEEE, 2008.

[19] KOSSMANN, D.; RAMSAK, F.; ROST, S. **Shooting stars in the sky: An online algorithm for skyline queries**. In: *Proceedings of the 28th international conference on Very Large Data Bases*, p. 275–286. VLDB Endowment, 2002.

[20] LEE, M.-W.; SON, W.; AHN, H.-K.; HWANG, S.-W. **Spatial skyline queries: exact and approximation algorithms**. *GeoInformatica*, 15(4):665–697, 2011.

[21] LEUNG, Y.; YAN, J. **Point-in-polygon analysis under certainty and uncertainty**. *GeoInformatica*, 1(1):93–114, 1997.

[22] LEUNG, Y.; YAN, J. **A locational error model for spatial features**. *International Journal of Geographical Information Science*, 12(6):607–620, 1998.

[23] LJOSA, V.; SINGH, A. K. **Top-k spatial joins of probabilistic objects**. In: *2008 IEEE 24th International Conference on Data Engineering*, p. 566–575. IEEE, 2008.

[24] LO, M.-L.; RAVISHANKAR, C. V. **Spatial joins using seeded trees**. In: *ACM SIGMOD Record*, volume 23, p. 209–220. ACM, 1994.

[25] LOFI, C.; EL MAARRY, K.; BALKE, W.-T. **Skyline queries in crowd-enabled databases**. In: *Proceedings of the 16th International Conference on Extending Database Technology*, p. 465–476. ACM, 2013.

[26] LUO, G.; NAUGHTON, J. F.; ELLMANN, C. J. **A non-blocking parallel spatial join algorithm**. In: *Data Engineering, 2002. Proceedings. 18th International Conference on*, p. 697–705. IEEE, 2002.

[27] MISHRA, P.; EICH, M. H. **Join processing in relational databases**. *ACM Computing Surveys (CSUR)*, 24(1):63–113, 1992.

[28] NI, J.; RAVISHANKAR, C. V.; BHANU, B. **Probabilistic spatial database operations**. In: *International Symposium on Spatial and Temporal Databases*, p. 140–158. Springer, 2003.

[29] OPENSHAW, S. **Learning to live with errors in spatial databases**. *Accuracy of spatial databases*, p. 263–276, 1989.

[30] PAPADIAS, D.; TAO, Y.; FU, G.; SEEGER, B. **An optimal and progressive algorithm for skyline queries**. In: *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, p. 467–478. ACM, 2003.

[31] PATEL, J. M.; DEWITT, D. J. **Clone join and shadow join: two parallel spatial join algorithms**. In: *Proceedings of the 8th ACM international symposium on Advances in geographic information systems*, p. 54–61. ACM, 2000.

[32] PEI, J.; JIANG, B.; LIN, X.; YUAN, Y. **Probabilistic skylines on uncertain data**. In: *Proceedings of the 33rd international conference on Very large data bases*, p. 15–26. VLDB Endowment, 2007.

[33] PFOSER, D.; JENSEN, C. S. **Capturing the uncertainty of moving-object representations**. In: *International Symposium on Spatial Databases*, p. 111–131. Springer, 1999.

[34] RIZZO, M. L. **Statistical computing with R**. CRC Press, 2007.

[35] ROSS, S. **A first course in probability**. Pearson, 2014.

[36] SHARIFZADEH, M.; SHAHABI, C. **The spatial skyline queries**. In: *Proceedings of the 32nd international conference on Very large data bases*, p. 751–762. VLDB Endowment, 2006.

[37] SIEGEL, S.; CASTELLAN JR, N. J. **Estatística não-paramétrica para ciências do comportamento**. Artmed Editora, 1975.

[38] SON, W.; HWANG, S.-W.; AHN, H.-K. **Mssq: Manhattan spatial skyline queries**. *Information Systems*, 40:67–83, 2014.

[39] SON, W.; LEE, M.-W.; AHN, H.-K.; HWANG, S.-W. **Spatial skyline queries: an efficient geometric algorithm**. In: *Advances in Spatial and Temporal Databases*, p. 247–264. Springer, 2009.

[40] TAN, K.-L.; ENG, P.-K.; OOI, B. C.; OTHERS. **Efficient progressive skyline computation**. In: *VLDB*, volume 1, p. 301–310, 2001.

[41] WOLFSON, O.; SISTLA, A. P.; CHAMBERLAIN, S.; YESHA, Y. **Updating and querying databases that track mobile units**. In: *Mobile Data Management and Applications*, p. 3–33. Springer, 1999.

[42] YOU, G.-W.; LEE, M.-W.; IM, H.; HWANG, S.-W. **The farthest spatial skyline queries**. *Information Systems*, 38(3):286–301, 2013.

[43] YU, X.; MEHROTRA, S. **Capturing uncertainty in spatial queries over imprecise data**. In: *International Conference on Database and Expert Systems Applications*, p. 192–201. Springer, 2003.

[44] ZHANG, M.; CHEN, S.; JENSEN, C. S.; OOI, B. C.; ZHANG, Z. **Effectively indexing uncertain moving objects for predictive queries**. *Proceedings of the VLDB Endowment*, 2(1):1198–1209, 2009.