COPPE
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia
UFRJ

PROCESSAMENTO DE CONSULTAS EM BANCOS DE DADOS GEOGRÁFICOS AMBÍGUOS

Vagner Braga Nunes Coelho

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador: Claudio Esperança

Rio de Janeiro
Dezembro de 2010

PROCESSAMENTO DE CONSULTAS EM BANCOS DE DADOS GEOGRÁFICOS AMBÍGUOS

Vagner Braga Nunes Coelho

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Claudio Esperança, Ph.D.

Prof. Júlia Célia Mercedes Strauch, D.Sc.

Prof. José Luiz Portugal, D.Sc.

Prof. Luiz Felipe Coutinho Ferreira da Silva, D.E.

Prof. Alexandre de Assis Bento Lima, D.Sc.

Prof. Geraldo Zimbrão da Silva, D.Sc.

RIO DE JANEIRO, RJ – BRASIL

DEZEMBRO DE 2010

Coelho, Vagner Braga Nunes

Processamento de consultas em bancos de dados geográficos ambíguos/Vagner Braga Nunes Coelho. – Rio de Janeiro: UFRJ/COPPE, 2010.

XV, 94 p.: il.; 29, 7cm.

Orientador: Claudio Esperança

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2010.

Referências Bibliográficas: p. 90 – 94.

1. Ambiguidade. 2. Similaridade. 3. Representações Múltiplas. I. Esperança, Claudio. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

A Deus

Agradecimentos

A Deus indispensável em todos os momentos da minha vida;

À minha esposa, Delma, pelo apoio constante, pelo incentivo e pela compreensão por tê-la deixado sem a minha companhia em muitos momentos durante o desenvolvimento da tese;

Ao meu filho, Natanael, pelos momentos de imensa alegria que me proporcionou quando as dificuldades tornaram-se demasiadas;

Aos meus pais, Norival e Selma, e meu irmão, Fábio, por tudo o que fizeram ao longo de minha vida;

Ao Exército Brasileiro por acreditar na tese desenvolvida e me liberar de horários de expediente normal para poder realizar o trabalho;

Ao Instituto Militar de Engenharia pelo suporte dispensado à consecução desta tese;

À Seção de Ensino em Engenharia Cartográfica pelo companheirismo nos momentos de maior dificuldade no desenvolvimento da tese;

À Universidade Federal do Rio de Janeiro por ter me permitido conviver com professores espetaculares;

À COPPE por viabilizar os professores que puderam participar na minha formação enquanto aluno do curso;

Ao Programa de Engenharia de Sistemas e Computação pela oportunidade de realizar esta tese;

Ao Laboratório de Computação Gráfica por ter oferecido um ambiente de trabalho fantástico, tanto pelo apoio no uso de equipamentos quanto no ambiente de camaradagem cultivado nele;

Ao Professor Claudio Esperança pela sua orientação segura, sua dedicação e incentivos constantes na condução e conclusão da presente tese;

À Professora Júlia Célia Mercedes Strauch pela sua disponibilidade e ajudas constantes na elaboração do texto e na participação em diversas etapas da pesquisa realizada;

Aos demais membros da banca, Professores Luiz Felipe Coutinho Ferreira da Silva, José Luiz Portugal, Geraldo Zimbrão da Silva e Alexandre de Assis Bento Lima, por terem aceito participar da mesma, cedendo um pouco de seu valioso tempo e conhecimento;

Aos meus colegas do LCG/COPPE pela amizade, pelas brincadeiras e pelo apoio em diversas fases do desenvolvimento desta tese.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

PROCESSAMENTO DE CONSULTAS EM BANCOS DE DADOS GEOGRÁFICOS AMBÍGUOS

Vagner Braga Nunes Coelho

Dezembro/2010

Orientador: Claudio Esperança

Programa: Engenharia de Sistemas e Computação

Esta tese propõe um novo paradigma em Banco de Dados Geográficos (BDG), baseado na integração de respostas a consultas. Este trabalho procura tratar ambiguidades geográficas encontradas no processamento de consultas a diversos BDG. Para isto, são utilizados o conceito de similaridade, cobertura e completude, empregando dois indicadores (índices de similaridade não espacial e espacial) de modo a consolidar a resposta a uma dada consulta.

Para validar esta proposta é apresentada a arquitetura denominada Sistema Avaliador de Respostas Ambíguas (SARA). Esta arquitetura é composta de um catálogo de domínios, um processador de meta-consulta e um classificador de ambiguidades.

Os experimentos realizados atestam que a similaridade dos polígonos representativos das feições proporcionam a integração das respostas. Assim, quando a consulta é realizada sobre representações múltiplas, a similaridade entre eles satisfaz a premissa da não necessidade de se proceder uma integração dos dados originais.

A principal contribuição deste trabalho é a apresentação de uma nova maneira de se obter informações a partir de uma consulta a múltiplas bases de dados que representam um mesmo tema, permitindo uma integração *a posteriori* das respostas ao invés de requerer uma integração *a priori* destas bases.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

QUERY PROCESSING ON AMBIGUOUS GEOGRAPHICAL DATABASES

Vagner Braga Nunes Coelho

December/2010

Advisor: Claudio Esperança

Department: Systems Engineering and Computer Science

This thesis proposes a new paradigm in Geographical Databases (GDB), based on the integration of query answers rather than data integration. This work seeks to address ambiguities found when querying several datasets which represents the same geographical features. For this purpose it is used the concept of similarity, coverage and completeness, using two indicators (nonspatial and spatial similarity indices) in order to consolidate the response to a given query.

To validate this proposal, an architecture named System for Evaluating Ambiguous Answers (SARA) was developped. This architecture consists of a catalog domain, a meta-query and meta-answer processor and an ambiguity classifier.

The experiments show that the similarity of the polygons representing the feature provides integration of responses. Thus, when the query is performed on multiple representations, the similarity between them satisfies the premise of no integrating the original data.

The main contribution of this thesis is to present a new way of getting information from a query to multiple databases, allowing the integration of responses rather the data.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Estado da arte na produção cartográfica

A elaboração da representação gráfica - documento cartográfico - de uma região do globo terrestre é uma atividade antiga, remontando a 2500 AC [1]. Ela requer uma série de cuidados operacionais para que o desenho obtido reflita, com coerência, a realidade física. Assim, a atenção destinada pelos profissionais envolvidos com a precisão nos dados amostrais demanda um tempo considerável na elaboração do documento. A minimização das distorções é um dos principais objetivos dos profissionais.

A cartografia, como ciência e técnica, oferece um conjunto de sistemas de projeção que mapeiam a superfície do globo terrestre em uma representação plana. Neste caso, há uma função f que aplica uma transformação de coordenadas geográficas (φ, λ) resultando em coordenadas planas (X, Y) (Figura 1.1). O conceito sedimentado no âmbito das ciências cartográficas é que o documento cartográfico representa o terreno por intermédio de um mapeamento biunívoco da forma $f_1(\varphi, \lambda) = X$ e $f_2(\varphi, \lambda) = Y$. Este paradigma serve de alicerce para todas as projeções cartográficas desenvolvidas e tem servido à comunidade por séculos.

Entretanto, em virtude de a evolução tecnológica ser uma constante na humanidade, os métodos e os instrumentos utilizados para representar a superfície terrestre têm evoluído. De fato, não apenas os procedimentos e os equipamentos vêm sendo alterados; a forma de exibição do produto obtido também tem sofrido alterações quando comparados à sua apresentação clássica, ou seja, em papel. Neste caso,

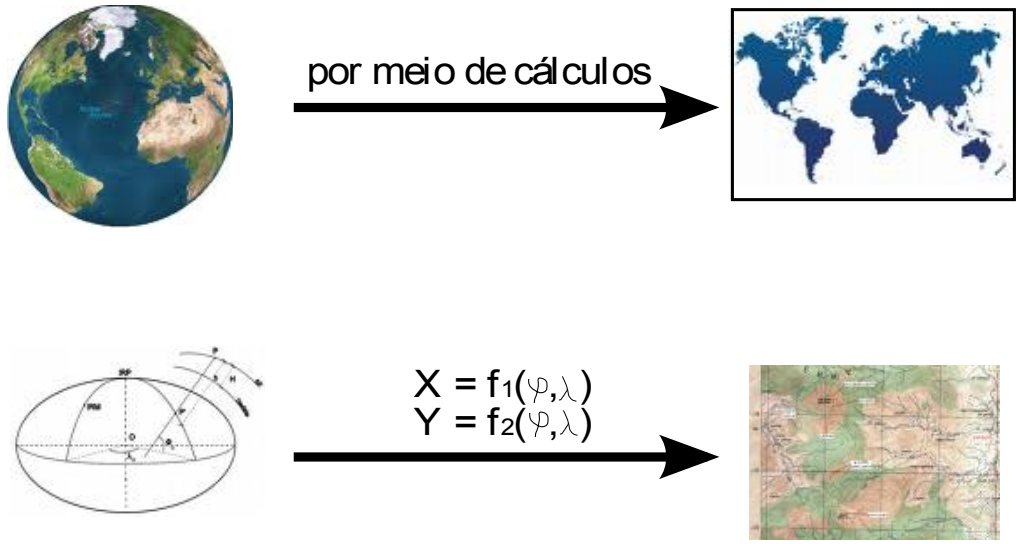

Figura 1.1: Transformação de coordenadas

o intenso uso de recursos computacionais na aquisição, no armazenamento e na manipulação dos dados cartográficos, bem como a crescente demanda pelo uso de computadores e o acesso a *internet*, têm exigido a forma de apresentação visual dos documentos cartográficos em um meio digital. Destarte, o uso de Banco de Dados Geográficos (BDG) tem sido a escolha natural para que os produtores armazenem os dados oriundos do mapeamento.

A base, por sua vez, é gerenciada por um Sistema Gerenciador de Banco de Dados (SGBD) que gerencia, armazena e processa a consulta. Logo, o SGBD é um arcabouço capaz de armazenar e processar dados de forma a permitir que um usuário habilitado consiga extrair informações por meio de consultas realizadas no repositório. Deste modo, o SGBD é o meio mais usual para o armazenamento e gerenciamento das informações geográficas.

Com a atual disponibilidade de dados na *internet*, a facilidade existente para o manuseio dos dados geográficos e para a construção de documentos cartográficos, o usuário passou a dispor de uma abundância de informações sobre uma mesma região. Consequentemente, esta abundância levou à ocorrência de ambiguidades nos dados, o que passou a ser um óbice ao paradigma clássico. Neste caso, a região não é mais representada por uma única função biunívoca. Há múltiplas representações em virtude de haver mais de um produtor de dados. Assim, cada produtor constrói seu próprio modelo, isto é, um documento cuja representação é similar, mas não

igual às dos demais produtores.

Convém ressaltar que, nesta tese, a multirepresentação das feições é tida apenas quando não há a representação temporal da mesma. Assim havendo, haverá dados similares que referem a épocas distintas, logo não podem ser tratados como uma ambiguidade.

1.2 Justificativa

A representação cartográfica é o resultado da modelagem aplicada à feição de forma a permitir a produção de dados que possam ser trabalhados. Nesta tese uma feição é identificada com a entidade real, ou seja, o objeto que existe no mundo real com características próprias. Assim, considerando que o processo de construção de bases cartográficas segue o paradigma da unicidade de representação (Figura 1.2), tem-se que para cada feição do terreno há uma única informação registrada nos diversos BDG. Entretanto, ao se consultar n bases há que se considerar a possibilidade de ocorrência de múltiplas representações para a mesma feição (Figura 1.3).

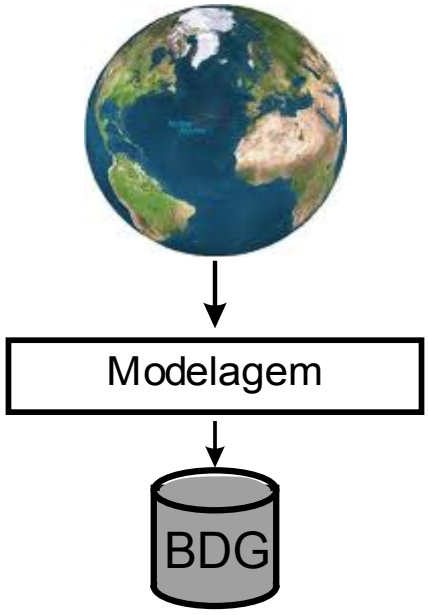

Figura 1.2: Representação unívoca

Isto ocorre devido a vários fatores. Entre eles destacam-se que as representações são modelos diferentes da realidade e que ao serem elaboradas em épocas com metodologias distintas, as representações nos BDG apresentam dados mais atualizados

ou dados com erros menores do que os outros. Ao se processar consultas sobre este BDG distribuído, as respostas encontradas podem apresentar:

- ausência de dados;
- redundância de dados;
- inconsistência de dados.

A ausência de dados se verifica quando não há valores registrados em um determinado BDG e há em outro. Em outras palavras, há falha na cobertura de um *dataset* específico. A redundância é, por assim dizer, o melhor caso, já que há a garantia de que o dado existente em um BDG é análogo ao existente em outro. Deste modo, não há dúvidas quanto à existência da representação, posto que esta foi modelada pelos diversos produtores. A inconsistência de dados ocorre quando os dados são conflitantes. Assim sendo, as representações que deveriam ser similares apresentam-se de tal forma que tornam-se antagônicas.

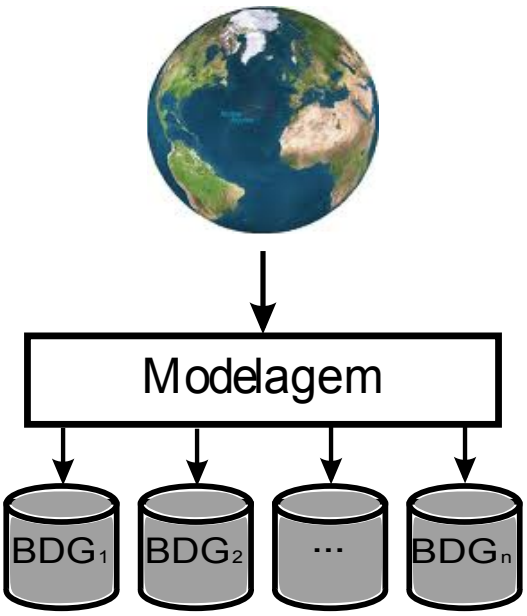

Figura 1.3: Múltiplas representações da mesma feição

No processamento usual de consulta em BDG, os dados geográficos são consolidados em uma representação única. Entretanto, este procedimento restringe o acesso aos diversos dados, posto que ao se integrar dados, alguns serão desprezados – ao se escolher um dado específico em detrimento de outro – ou serão construídos dados derivados ao se proceder uma forma qualquer de ajuste entre estes [2].

Diante da possibilidade de ocorrência de ambiguidades nos dados e da impossibilidade de determinar qual é o melhor modelo, esta tese propõe a mudança no paradigma atual, ao se estabelecer a integração de respostas a consultas e não a integração de dados. Isto porque esta nova forma de conduzir a obtenção de informações propicia uma maior amplitude de respostas, uma redução nos custos produtivos, a preservação de autoria e, evidentemente, serve como um certificador de dados. Esta tese vai ao encontro da proposta de Hessen [3], quando este afirma que "uma representação inadequada, por sua vez, pode ser verdadeira, pois apesar de incompleta pode ser correta, se as características que contém existirem efetivamente no objeto".

1.3 Objetivos

Este trabalho apresenta uma metodologia de processamento de consultas a dados geográficos multirepresentados. Para tal, é proposta uma arquitetura, denominada de Sistema Avaliador de Respostas Ambíguas (**SARA**), capaz de tratar as ambiguidades em dados geográficos. A inovação baseia-se na sumarização dos dados disponíveis construida a partir do uso dos seguintes indicadores de similaridade:

- Índice de Similaridade não espacial ($\mathcal{S}_n$); e,
- Indíce de Similaridade espacial ($\mathcal{S}_g$).

O primeiro avalia o nome da feição empregando o coeficiente de Dice (d_d). Neste caso, há uma análise do conjunto de caracteres encontradas em uma *string* que serve para identificar a feição – *nome*. O segundo avalia a geometria e é um avaliador baseado no Índice de Similaridade Cartográfico (ISC), que possui a finalidade de estabelecer o grau de similaridade entre representações poligonais. Este indicador é comparado com a expansão de um método clássico – Método dos Retângulos Equivalentes [4]– e a sua eficiência é testada em uma aplicação em dois BDG diferentes [5].

São introduzidas, ainda, na arquitetura dois indicadores de modo a facilitar o processamento de integração das respostas às consultas, a saber: Índice de Cobertura ($\mathfrak{Co}I$) e a extensão do Índice de Completude ($\mathfrak{C}I$) ([6] e [7]). O primeiro indicador especifica o quanto uma determinada representação contribui para a determinação do

locus geográfico, enquanto o segundo índice tem por objetivo quantificar a influência de uma região comum frente a uma representação particular.

A utilização destes índices – semânticos e espaciais – permitem a utilização de todos os dados geográficos disponíveis como insumo na geração de uma resposta consolidada com um dado limiar de aceitação.

1.4 Exemplo motivador

Visando facilitar a compreensão e a percepção do problema relativo às ambiguidades espaciais, sejam as representações de uma mesma feição qualquer do terreno – polígono P, por exemplo – oriundas dos fornecedores F_1 (preto) e F_2 (vermelho) (Figura 1.4a) e as representações pontuais – pontos v_i, por exemplo – oriundas dos fornecedores F_3 (verde) e F_4 (azul) (Figura 1.4b). Evidentemente, no caso dos polígonos há um certo grau de similaridade uma vez que eles podem possuir um *locus* geográfico em comum. No caso, os pontos v_1 e v_2 dos fornecedores F_3 e F_4, respectivamente, são similares, enquanto os demais não são.

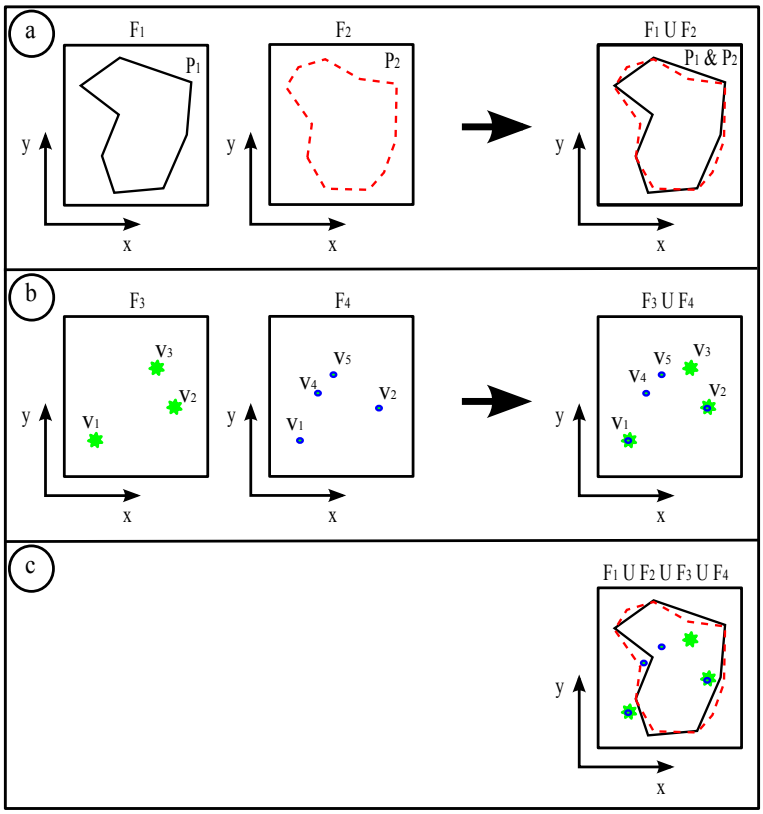

Figura 1.4: Representações ambíguas

O paradigma clássico prescreve a existência de apenas um polígono representativo de cada feição poligonal, bem como um único ponto representativo de cada feição pontual. Entretanto, em virtude da ambiguidade existente, observa-se na Figura 1.4c a multiplicidade de respostas possíveis a uma consulta qualquer. Assim sendo, ao se desejar quantificar quantos *pontos* estão dentro do *polígono*, teremos quatro respostas possíveis à consulta. As opções possíveis podem ser consolidadas observando a Tabela 1.1.

Tabela 1.1: Respostas diversas

opção	quantidade
pontos de F_3 no polígono de F_1	2
pontos de F_4 no polígono de F_1	2
pontos de F_3 no polígono de F_2	2
pontos de F_4 no polígono de F_2	3

Classicamente, um usuário habilitado, e com autoridade para tal, certificaria um dos produtores pelos dados representados por *polígonos* e um dentre os que forneceram os dados dos *pontos*. Neste caso, haveria apenas uma representação poligonal e uma única pontual, eliminando assim o problema de respostas múltiplas.

Diante do espectro de respostas possíveis observado na Tabela 1.1, a arquitetura proposta integra as quatro possibilidades de forma a permitir que o usuário tenha acesso a uma faixa de respostas com graus de confiabilidade associados. Na solução proposta para o caso de uma consulta sobre os pontos e os polígonos a resposta correta deve ser no mínimo o conjunto solução $\{v_2\}$, por ser comum em todas as quatros opções, ou no máximo o conjunto $\{v_2,\ v_3,\ v_4,\ v_5\}$, com algum grau de imprecisão.

1.5 Organização da tese

A presente tese encontra-se estruturada em **8** capítulos.

O capítulo 2 trata do problema de integração de dados geográficos e, para tal, traz uma breve consideração sobre a construção das bases modernas e a publicação das mesmas. Posteriormente, apresenta as arquiteturas mais recentes para integração e

publicação de dados.

No capítulo 3 encontra-se detalhado o procedimento utilizado para analisar e correlacionar as diversas representações disponíveis para cada tema geográfico.

O capítulo 4 descreve as funções de similaridades semântica e geométricas capazes de viabilizar a identificação da identidade entre representações ambíguas.

O capítulo 5 apresenta o detalhamento do processamento de consultas de seleção e de junção.

O capítulo 6 apresenta a arquitetura proposta para a classificação e o tratamento das ambiguidades que porventura exista entre duas ou mais Bases de Dados Geográficos.

O capítulo 7 apresenta os resultados obtidos com a abordagem proposta aplicada a consultas em banco de dados reais, bem como uma análise dos mesmos.

O capítulo 8 apresenta uma breve conclusão da tese.

Capítulo 2

Integração de dados geográficos

2.1 Considerações iniciais

A representação biunívoca do terreno por intermédio de objetos cartográficos, ou seja, a apresentação das diversas feições do terreno por seus respectivos modelos digitais é a essência da construção de bases cartográficas. A feição modelada é, de certa forma, incognoscível, ou seja, não se pode representá-la perfeitamente. Isto porque os processos utilizados para a obtenção das coordenadas referentes à feição são eivados de erros diversos. Na realidade, o que se obtém é apenas uma aproximação da realidade.

Para tal, os métodos de obtenção dos dados são efetuados de maneira a garantir que as coordenadas estejam dentro de uma tolerância aceitável dentro de uma dada escala de representação. As tolerâncias encontradas no país são as prescritas em [8], que trata das diretrizes e bases da cartografia brasileira. Neste caso, encontram-se legisladas um erro gráfico de *0,2 mm* na escala da carta e erro planimétrico de *0,5 mm* na escala da carta para um documento tido como Classe A. Evidentemente, tais tolerâncias são prescritas para os documentos representados em uma determinada escala. Assim sendo, podem ser obtidas múltiplas representações de uma mesma feição em virtude da representação em diversas escalas.

Neste capítulo são apresentados os conceitos necessários para a integração de dados geográficos, o processo de construção de bases geográficas, apresentação de um problema real no Brasil de necessidade de integração de dados e soluções propostas para integração de multirepresentação de dados geográficos.

2.2 Conceitos

Com o intuito de se integrar bases distintas, há que se considerar a existência de três características em relação aos dados, conforme relacionados a seguir:

- ambiguidade;
- cobertura;
- completude.

2.2.1 Ambiguidade

A ambiguidade – nesta tese quantificada por uma função de similaridade $\mathcal{S}$ – é associada à multirepresentação de feições do terreno. Neste caso, há mais de uma informação nos *datasets* disponíveis para uma mesma realidade do terreno. É necessário considerar que a ambiguidade pode ocorrer em dois cenários. O primeiro ocorre quando há em um único *dataset* uma representação ambígua. Na ocorrência desta possibilidade, encontra-se, geralmente, um erro grosseiro na produção. Tal erro pode ser corrigido ao se proceder uma supervisão e inspeção rigorosas sobre a fonte de dados. O segundo caso aparece quando um usuário processa dados de *datasets* diferentes. Assim, confronta-se com representações distintas da mesma realidade física. Este tipo de ambiguidade é comum porque "erros em *datasets* geográficos não podem ser evitados" [6] em função das diferentes amostragens no mundo real.

2.2.2 Cobertura

A cobertura ($\mathfrak{Co}$) pode ser interpretada como uma medida do quanto uma representação específica é recoberta pelo *locus* geográfico da região estimada por todas as representações disponíveis da feição. Assim, é possível estabelecer um grau de cobertura para cada representação da feição individualmente quando avaliada no contexto da região modelada.

2.2.3 Completude

A completude ($\mathfrak{C}$) é uma medida que procura estabelecer o quanto uma dada representação concorda com outra. Neste caso, um índice de completude é capaz de

quantificar o quanto o *locus* geográfico, recoberto concomitantemente por todas as representações disponíveis, encontra-se dentro da região de influência de um modelo em particular.

A análise da completude tem sido aplicada a diversas atividades científicas, geralmente associadas a identificação de base de dados desatualizadas [7]. Neste caso, um avaliador de completude permite quantificar o quanto um *dataset* específico está atualizado em relação a um outro.

2.3 Construção de bases geográficas

Quando se percebe o mundo real, há que se ressaltar que a feição real pode ser apresentada de duas formas, a saber:

- pelo dado cartográfico;
- pelo dado geográfico.

O dado cartográfico é, em verdade, a representação espacial da feição. Desta forma, modela-se a feição por meio de um visão particular, atribuindo-lhe uma geometria de acordo com a escala de representação. Considerando a existência de vários órgãos produtores de cartografia, haverá diferentes dados geográficos para cada produtor, independentemente de representarem a mesma feição, uma vez que para sua obtenção podem ser empregadas diversas técnicas, por exemplo: topografia, sensoriamento remoto, geodésia por satélite, compilação, entre outras. Por sua vez, o dado geográfico é a descrição da feição efetuada de maneira alfanumérica por seus atributos qualitativos e quantitativos que são levantados e associados aos dados cartográficos em um determinado instante de tempo. A obtenção destes dados são os insumos para a construção de uma base cartográfica [9].

A metodologia atual para a construção de uma base cartográfica em meio digital preconiza apenas uma adaptação da forma clássica feita por plástico-gravura. Assim sendo, houve apenas uma migração da produção clássica para um meio apoiado por computadores; isto é, apenas uma mudança na forma de apresentação dos dados e não uma alteração na metodologia. Conceitualmente, a carta ou o mapa continuou sendo produzido dentro das mesmas fases, ou seja, o produtor realiza todas as atividades anteriores, passando, simplesmente, a ser o detentor de um arquivo em meio

digital que contém o subproduto da fase em questão. Estes arquivos são, em sua maioria absoluta, arquivos proprietários com um formato específico que dificulta a troca de informações [10].

O estágio atual de uso da informática pelos produtores viabiliza apenas uma automação do processo clássico de construção dos documentos cartográficos, gerando um *workflow* de aquisição e modelagem baseado em arquivos (Figura 2.1). Isto não tem contribuido para uma integração adequada entre as diversas instituições porque a quantidade de arquivos finais gerados é muito elevada. Acrescenta-se, ainda, que há uma série de arquivos intermediários para cada etapa desenvolvida que podem ser intercambiadas entre as várias fontes de dados.

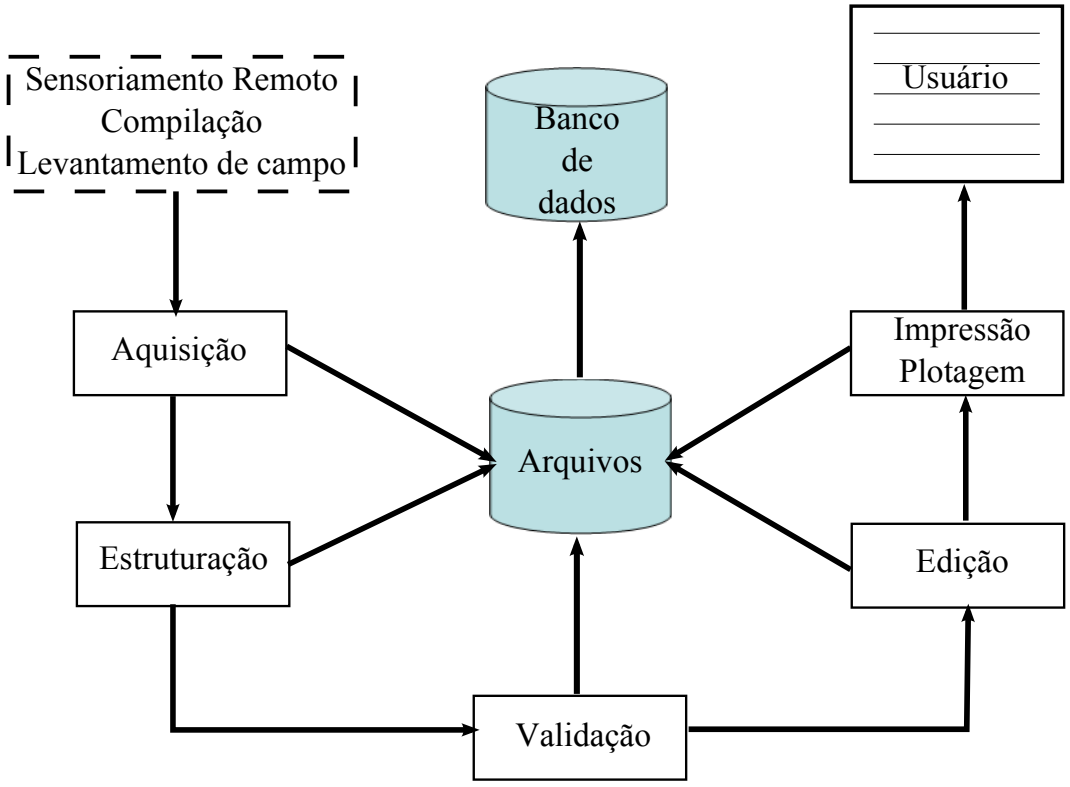

Figura 2.1: *Workflow* atual

Procurando amenizar os problemas com a localização de arquivos, os órgãos produtores da Cartografia passaram a se utilizar de Sistemas Gerenciadores de Bancos de Dados (SGBD). Entretanto, para estas instituições, a utilização dos SGBD tem sido orientada para ser apenas um repositório de arquivos. Neste caso, estes têm-se tornado apenas indexadores da localização dos mesmos em um servidor e facilitador do acompanhamento de responsabilidades pela execução de cada uma das tarefas

envolvidas no processo produtivo.

2.4 Publicação de bases geográficas no Brasil

Com o intuito de otimizar a carga dos dados e facilitar o acesso às informações existentes, observa-se que a necessidade de construção da base cartográfica em um ambiente de banco de dados é imperativa. Para a consecução destes objetivos existe, atualmente, uma série de soluções proprietárias. Dentre as existentes, destacam-se aquelas propostas pelo ArcGIS® da ESRI®[1], o Geomedia® da Intergraph®[2] e PostGIS [3] como proposta livre desenvolvida pela *Refraction Research* com licença GNU [4].

Visando a normalização de dados há, atualmente, uma série de esforços com o intuito de tornar cada vez mais acessível o uso dos SGBD. Aliado a isto, há no Brasil uma política governamental [11] e uma crescente conscientização dos pesquisadores e dos institutos de pesquisas na busca de uma solução baseada em *software* livre e de domínio público. Alguns esforços recentes, no país, estão sendo desenvolvidos visando a implementação de tais soluções já com a tecnologia de banco de dados. Dentre vários projetos, destacam-se o Projeto de Banco de Dados Geográficos - BDGeo [12] – e o Banco de Dados Geográficos do Exército – BDGEx [13].

O BDGeo é, na realidade, um *framework* desenvolvido para uma modelagem conceitual dos dados geográficos baseado no ambiente do Sistema de Informação Geográfica (SIG) denominado *Spring* [5] desenvolvido pelo Instituto Nacional de Pesquisas Espaciais (INPE). É um esforço teórico importante, que visa orientar e permitir a definição de regras para o mapeamento dos esquemas conceituais segundo as regras do formalismo da orientação a objetos. Neste projeto em particular, para cada região geográfica pode-se especificar uma coleção de temas.

O segundo projeto – BDGEx – também é desenvolvido segundo o paradigma da orientação a objetos e a proposta do uso de $software$ livre. Para tal, utiliza-se como

[1] http://www.esri.com/software/arcgis/index.html, capturado em 07 de novembro de 2010

[2] http://www.sisgraph.com.br/geomediasuite/default.asp, capturado em 07 de novembro de 2010

[3] http://postgis.refractions.net/, capturado em 07 de novembro de 2010

[4] *General Public Licence*

[5] http://www.dpi.inpe.br/spring/portugues/index.html, capturado em 07 de novembro de 2010

sistema operacional o Linux e como banco de dados o PostGreSQL [6]. O BDGEx é parte integrante do Sistema de Informações Geográficas do Exército (SIGEx) que está sendo desenvolvido no Centro de Imagens e Informação Geográficas do Exército (CIIGEx), antigo Centro de Cartografia Automatizada do Exército (CCAuEx). Este sistema visa integrar em um banco de dados espaciais todo o espaço geográfico do país, eliminando, assim, as inconsistências relativas às descontinuidades históricas das cartas topográficas. Uma ressalva a este projeto deve-se ao fato de que o mesmo estará disponível apenas aos usuários do Exército Brasileiro (EB). O acesso é negado aos demais usuários da Cartografia Nacional porque o assunto é visto como uma ameaça a Segurança Nacional. Atualmente, as informações espaciais contidas no Banco encontram-se digitalizadas matricialmente e cada carta topográfica é um arquivo. As linhas das tabelas contêm, claramente, entre outras geoinformações, o endereço no servidor do arquivo em trabalho. É um esforço válido como um primeiro passo na direção da integração das informações cartográficas do país em um ambiente digital. De qualquer forma, as iniciativas desenvolvidas ainda são incipientes quando se tem em mente a necessidade de disponibilização do dado a todo e qualquer usuário.

2.5 Integração de multirepresentação de dados geográficos

A multiplicidade de dados geográficos, ou seja, a ambiguidade entre as representações de uma feição específica, tem sido discutida e tratada de muitas formas. Atualmente, ainda nos encontramos sob o paradigma da representação única por meio de um processo de integração [14]. Com o intuito de se desenvolver metodologias para a integração, pode-se encontrar dois tipos de arquitetura que fornecerão um resultado único ao final do processamento. O primeiro é a arquitetura de integração dos dados. Neste caso, desenvolve-se um trabalho *a priori* para que se gere uma base única representativa do mundo real. A segunda opção tem a ver com a publicação da base. Assim, há um processamento no sentido de disponibilizar apenas um dado para cada feição, embora esta não necessariamente corresponda à realidade.

[6] http://www.postgresql.org/, capturado em 07 de novembro de 2010

As características de cada uma das propostas anteriores são relevantes para o contexto da unificação de bases, mas apontam para um esforço produtivo no intuito de se classificar as bases existentes de forma a se permitir uma escolha daquela considerada como a melhor.

2.5.1 Metodologias para a integração de bases

A integração das bases tem sido o esforço maior dos produtores de dados, pois estes desejam um modelo de referência único. Embora haja a possibilidade de se vir a obter várias representações do mesmo objeto no mundo real, os diversos produtores têm optado pela utilização de um profissional responsável pelo desenvolvimento e compilação dos dados para proceder a escolha do modelo a ser usado para representação e visualização das feições [15]. Há várias soluções, como, por exemplo, a integração das bases realizadas por ontologia [16], de modo a propiciar uma integração a partir de bases construídas por instituições diversas em momentos distintos.

Convém ressaltar que existem dois tipos de integração que não são oriundas de ambiguidades. A primeira tem a ver com a perfeita junção de modelos em documentos adjacentes [17]. Neste caso, as bases adjacentes devem ter suas representações digitais contíguas (Figura 2.2). Não deve haver afastamento entre os objetos cartográficos ao se justapor as bases adjacentes.

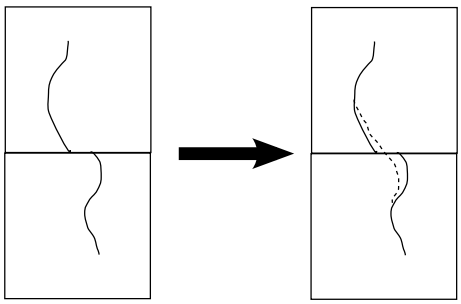

Figura 2.2: Representações do mundo real

No segundo tipo tem-se a integração que é a realizada dentro do próprio documento. Neste caso, o produtor deve fornecer uma base única como resultado de seu trabalho. Entende-se, aqui, como resultado do trabalho, o produto obtido após todo o processo produtivo, incluindo os trabalhos de campo e de gabinete necessários para a construção da base. De maneira genérica tem-se que um produtor não pode,

a partir de um único conjunto de dados, gerar mais de uma representação para uma dada feição em um documento na mesma escala.

Multidatabase

A técnica de integração por *multidatabase* prescreve uma série de operações para viabilizar a unificação da base geográfica que, porventura, esteja distribuída em várias fontes. Trata diferente esquemas nas diversas bases com o intuito de homogeneizar as consultas. Neste caso, o tratamento é feito *a priori*, ou seja, são realizadas as operações de forma que uma consulta simples forneça a mesma resposta independentemente de onde se encontra a informação. O *multidatabase* provê um conjunto de funções de mapeamento para integração de dados que não está presente nas linguagens de manipulação dos dados distribuidos [18].

Neste contexto, o *multidatabase* é, em verdade, uma maneira de se trabalhar os esquemas individuais de forma a gerar um esquema global por meio de uma série de regras para a integração [19]. *A priori* é definida uma metodologia de integração que resolva os conflitos. Tais conflitos, geralmente, ocorrem em função da geometria do dado e em função dos esquemas particulares. Desta forma, uma consulta simples produz tantas informações quantos bancos existam.

Mediadores

Os mediadores são, na verdade, arquiteturas desenvolvidas para viabilizar a coleta de informações esparsas. Nestas arquiteturas são efetuadas pesquisas nos dados disponíveis e após a seleção dos temas que interessam à consulta efetua-se a integração dos mesmos de forma a viabilizar a produção de uma resposta única. Há, dessa forma, a construção *virtual* de uma base única para ser usada como insumo das consultas.

Alguns autores, [20] e [21], estabelecem dicionários com os termos mais usuais de forma a realizar um mapeamento entre os dados textuais com a finalidade de se agilizar o processamento. Atualmente, os mediadores se utilizam de ontologias com o objetivo de integrar os esquemas conceituais. Assim, há uma correlação mais estreita entre os diferentes dados – esparsos – e a construção temporária de uma base única. Na realidade, os mediadores são módulos em sistemas que viabilizam a

junção de múltiplas fontes em sistemas de informação [15].

Comparação entre as arquiteturas

Os objetivos da *multidatabase* e dos mediadores são exatamente os mesmos. Ambas proporcionam a integração dos dados para que as consultas aos dados esparsos sejam analisadas e gerem uma resposta única. A principal diferença reside no fato de que o *multidatabase* gera uma base unificada fisicamente, enquanto nos mediadores é virtual. Entretanto, tanto em um quanto no outro, a base – virtual ou física – não é algo que pertença à cadeia produtiva de um orgão qualquer. Neste caso, é uma base criada apenas e tão somente para a obtenção de respostas; não são processos de desenvolvimento de novas bases.

2.5.2 Metodologias para a publicação

Visando a publicação de bases de dados há um série de formas clássicas, dentre estas destacam-se as seguintes:

- Biblioteca Digital;
- *Clearinghouse*;
- Curadoria Digital.

Biblioteca Digital

A Biblioteca Digital tem sua origem no ano de 1994, quando a Universidade da Califórnia apresentou o projeto denominado de *Alexandria Digital Library* (ADL) que permitia o acesso remoto a dados espaciais, visando a representação de uma imagem [22].

Na realidade, a biblioteca digital para dados cartográficos é um índice localizador de bases. Em outras palavras, é um ponteiro para um repositório, localizado remotamente, que possua os dados de um determinado tema de forma unificada. Neste contexto, o usuário, ao acessar a biblioteca, passa a identificar a instituição detentora do tema de seu interesse.

Este localizador funciona por meio de palavras chave – "tesauros" – que, após a consulta ao conjunto de dados cadastrados, permite a identificação do produtor e do detentor do tema.

Clearinghouse

O *Clearinghouse* é uma evolução da arquitetura de biblioteca digital onde os dados são transferidos para um responsável. Neste caso, um administrador torna-se o responsável pela certificação de dados das diversas instituições produtoras, pelo armazenamento dos dados certificados e pela publicação em geoportais destes dados [23]. Assim sendo, percebe-se que o administrador não altera os dados, apenas os fornece aos usuários após avaliar as possibilidades e inferir sobre quais são os mais adequados.

Curadoria Digital

A técnica mais recente é a curadoria dos dados [24]. Seu emprego tem sido sustentado, principalmente, pelo *Digital Curation Center*[7] (DCC) do Reino Unido. Os primórdios da curadoria de dados remontam o ano de 1998 quando foi criado um repositório para armazenamento de dados digitais [25]. Naquele momento o objetivo foi a preservação de todos os dados digitais, até mesmo aqueles considerados desatualizados. A idéia principal foi a de facilitar a pesquisa e consulta aos dados, primordialmente, em forma de texto.

O cerne da curadoria de dados é a existência de um repositório de informações que pode ser acessado por qualquer usuário, independentemente dele pertencer ou não a uma organização que forneça dados. Este princípio visou à democratização dos dados, bem como a permitir a concentração destes em um local apropriado. O obstáculo atual à implantação da curadoria reside nas diversas legislações nacionais que oferecem uma resistência considerável para publicidade das informações [26].

Para facilitar o empreendimento, a utilização de geoportais tem sido procurada por facilitar a interação produtor-usuário. O acesso aos geoportais permite aos usuários localizar as bases de dados que deseja. Infelizmente, apenas um conjunto de dados são inseridos no geoportal de forma a comporem a base, ou seja, a diversidade

[7]http://www.dcc.ac.uk/, capturado em 14 de agosto de 2009

– ambiguidade – nos dados existente nos diversos mapas e cartas é eliminada após a retirada dos conflitos por um operador autorizado.

Comparação entre as arquiteturas

Diante do exposto, verifica-se que a Biblioteca Digital, o *Clearinghouse* e a Curadoria Digital possuem o mesmo objetivo. O foco das propostas é o fornecimento de dados georreferenciados. Entretanto, a maneira como cada uma destas propõe o acesso aos dados é diferente. Neste caso, faz-se necessária uma abordagem individualizada. De forma simplificada pode-se observar que a Biblioteca Digital oferece a possibilidade de obtenção de múltiplas representações – ambiguidades. O *Clearinghouse*, por sua vez, disponibiliza uma única representação de cada tema após uma certificação e a Curadoria Digital oferece uma representação única após processamento, por parte do órgão certificador, dos dados obtidos junto aos diversos produtores.

2.5.3 Benefício da não integração

Na realidade não há inconsistência técnica ao se possuir mais de uma visão da feição. Inclusive, isto é usual quando da construção da base. Neste caso, o produtor de dados utiliza informações ambíguas com o claro intuito de auxiliá-lo na produção. Entretanto, os profissionais tendem a preferir uma base única a várias potencialmente ambíguas. Porém, a questão primordial no uso dos *datasets* ambíguos reside no fato de que a integração dos dados é custosa e não é possível sem um trabalho intenso de gabinete.

Ao constatar a existência de múltiplas representações, um dado usuário normalmente se vê, geralmente, forçado a optar por uma delas. Entretanto, esta escolha não é fácil até porque não há razões concludentes para a escolha de um *dataset* específico em detrimento de um outro.

Para o construtor, a diversidade de bases introduz uma maior gama de informações representativas do terreno possibilitando, assim, obter outras informações que não apenas aquelas obtidas por uma base única. Quando de posse de apenas uma base, erros grosseiros podem ser mascarados. Quando de posse de diversas bases, tais erros geométricos podem ser evidenciados de forma mais clara (por exemplo, a Figura 2.3), pois haverá várias representações semelhantes contra uma conside-

ravelmente diferente. Caso isto ocorra, está sinalizado que um levantamento foi equivocado e, provavelmente, serve de indicativo sobre que local da região deve ser realizada uma operação de campo para dirimir as dúvidas. Em outras palavras, a multiplicidade de bases permite a identificação de feições com problemas.

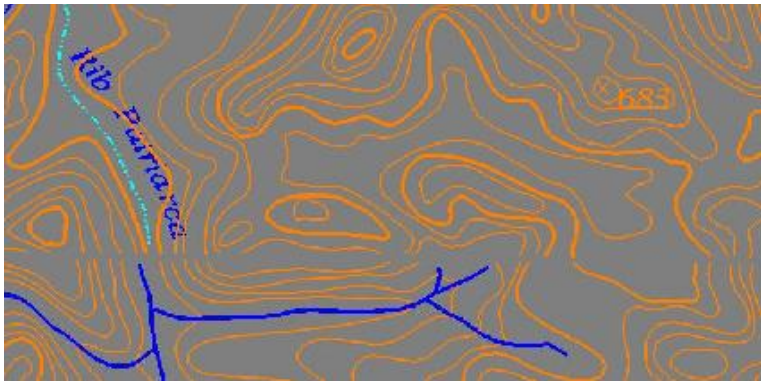

Figura 2.3: Conflitos usuais na ligação de bordas

A integração das bases geográficas tem sido o objeto de contínuas pesquisas devido a ser uma necessidade real nos órgãos produtores de cartografia. Vários são os trabalhos que ressaltam a necessidade ou apresentam como dificuldade de suas respectivas pesquisas a integração [27], [28], [29] e [30] e há, ainda, outros em que a própria integração é o objeto da pesquisa [31] e [32]. O trabalho custoso e volumoso para a obtenção de um *dataset* único que represente uma dada região não é proibitivo, mas pode postergar a obtenção de uma resposta caso haja urgência por parte do usuário.

2.6 Considerações finais

O que se pode observar é que a construção das bases geográficas e a sua publicação é um procedimento que não está estagnado no tempo. Embora o processo tenha mudado muito pouco até o presente momento, verifica-se que a metodologia possui um vasto campo para se desenvolver, possibilitando, além da agilidade na construção das bases, a viabilização de produzir novos produtos e a concessão ao usuário de interagir com o processo e obter ele mesmo o mapa desejado.

A presente tese procura apresentar uma nova maneira de se obter informações a partir de múltiplas bases geográficas. Baseia-se na maior interação produtor-usuário

visando aliar as necessidades de usuários com os dados existentes nos diversos bancos e fornecer possibilidades de consulta para propiciar ao usuário obter a informação com o maior subsídio possível, inclusive sobre a qualidade do dado.

Capítulo 3

Mapeamento de correspondência em BDG ambíguos

3.1 Considerações iniciais

Um banco de dados é um conjunto de registros disposto de forma regular com a finalidade de propiciar acesso a dados específicos. O modelo relacional é o mais comumente usado para a estruturação do banco. Esse modelo contempla a realização de consultas por intermédio da assim chamada álgebra relacional [33]. Dentro dos operadores da álgebra relacional, os mais usuais são os seguintes:

- seleção (σ);
- projeção (π); e,
- junção ($\bowtie$).

A resposta R obtida com uma dada consulta Q é uma relação, isto é, um conjunto de tuplas (registros) que atendem a um predicado específico.

Se um determinado banco de dados admite atributos espaciais [34], tem-se os chamados bancos de dados espaciais. Usualmente, estes bancos são utilizados em aplicações geográficas e, neste trabalho, são denominados por Bancos de Dados Geográficos (BDG). Independentemente de os bancos de dados serem comuns ou espaciais, estes conjuntos de dados (*datasets*) devem atender a restrições de integridade que podem ser classificadas como:

- da relação;
- referencial;
- de domínio;
- da coluna; e,
- definida pelo usuário.

As restrições de integridade permitem a construção de um *dataset* cujas tuplas sejam únicas, ou seja, há uma monorepresentatividade dos dados. Normalmente, pretende-se estabelecer uma correspondência biunívoca entre os dados e o mundo real. O conjunto de dados funciona, neste caso, como um modelo particular da realidade. O modelo é, em verdade, uma descrição de um fenômeno a partir de observações do mesmo.

O princípio da relação biunívoca entre os dados registrados e a realidade pode deixar de existir quando se considera mais de um modelo para uma dada realidade. Deste modo, embora cada *dataset* seja unívoco, a disponibilidade de mais de um *dataset* associado ao mesmo fenômeno cria a percepção de dados multirepresentados, ou seja, o mesmo objeto real encontrado em mais de um BDG. Logo, há uma diferença conceitual entre as consultas aos BDG que são monorepresentativos da realidade e a presente proposta de tratamento de dados potencialmente ambíguos – multirepresentação.

Assim, ao efetuar-se consultas aos diversos BDG monorepresentativos disponíveis obtém-se respostas múltiplas a uma dada consulta quando, classicamente, é esperada apenas uma resposta. Portanto, cumpre que se procure determinar o grau de concordância entre estas múltiplas respostas. Em outras palavras, deve-se tentar estabelecer que dados se referem à mesma realidade e quais se referem a realidades distintas.

Para esse fim, um conceito importante é o de similaridade. Na ausência de uma inspeção de campo, a única forma de determinar se dois dados diferentes são modelos da mesma realidade é por meio da avaliação de sua similaridade. É razoável, por conseguinte, que se investigue técnicas que permitam quantificar o grau de similaridade de dois dados de tal forma que se possa estabelecer um limiar a partir do qual estes sejam julgados como se referindo à mesma realidade.

A identificação da similaridade entre os dados torna possível a construção das classes de correspondência entre feições multirepresentadas. Assim, é possível, por exemplo, inferir a quantidade de feições do mundo real e avaliar a completude e a cobertura de um *dataset* específico em relação a outro. Uma vez obtido um mapeamento destas correspondências, é útil registrá-lo numa estrutura de dados para sua recuperação quando necessário.

Como forma de permitir uma análise da proposta, é apresentada aqui uma metodologia que permite a obtenção dos dados para o caso de uma consulta de seleção e outra de junção. Acrescenta-se, ainda, que a metodologia viabiliza a inferência de índices relativos à similaridade entre representações, a cobertura e a completude de uma representação específica. Os avaliadores propostos são métricas que atendem os aspectos espaciais e não espaciais.

3.2 Premissas

Seja um tema específico τ do mundo real. Tem-se que $\tau = \{F_1, F_2, \ldots, F_n\}$, onde F_i é uma feição particular do mundo real. Um produtor de dados modela τ de tal maneira que as entidades reais F_i sejam instanciadas, no tempo e no espaço, por uma representação $\Phi(\tau)$ ou, simplesmente, Φ. Neste caso, Φ é uma visão particular de τ, ou seja, Φ é uma função de representação do tema τ.

Neste trabalho, o esquema utilizado para a representação será limitado. Assim, a representação Φ está em uma tabela T com dois atributos, um representando o *nome* e outro contendo a *geometria* de cada feição. Tal simplificação é razoável pois estes atributos são aqueles que correlacionam a representação com o mundo real.

Nos casos reais, os dados possuem outros atributos. Entretanto, estes outros apenas registram dados extrínsicos da feição, tais como: a área, a população, o perímetro, a capacidade de carga, dentre outros.

Na realidade, $T = \{\Phi(F_1), \Phi(F_2), \ldots, \Phi(F_n)\}$, onde cada $\Phi(F_i)$ é uma tupla da tabela T. Com o objetivo de simplificar a notação, $\Phi(F_i)$ pode ser apresentada como ϕ_i, ou seja, $\phi_i = \Phi(F_i)$. Convém ressaltar que o índice i funciona como uma chave primária, logo é possível mapear uma representação de um *dataset* específico sobre outras bases de dados. Em outras palavras, assume-se que é possível recuperar uma

dada tupla da tabela com base em seu índice.

A primeira coluna de T – *nome* – identifica o nome da representação pelo qual a feição é instanciada e identificada univocamente. O atributo nome é, nesta tese, uma cadeia de caracteres alfanuméricos. Desta forma, é possível tratar cada caractere, individualmente, com o objetivo de se avaliar a similaridade entre nomes potencialmente ambíguos. Neste caso, há que se identificar os *nomes* que possuem significado similar, mesmo que não estejam associados à mesma feição. Assim, seja $N(T)$ a função de projeção que gera uma tabela com os nomes de T. Neste caso, $N(T) = \pi_{nome}(T)$.

A coluna *geometria*, por sua vez, relaciona os valores das coordenadas instanciadas de cada feição F_i. Assim sendo, $G(T) = \pi_{geometria}(T)$. O atributo geometria pode ser de um dos seguintes sete tipos conforme a prescrição do *Open Geospatial Consortium* (OGC), usualmente adotada [34]:

- *point* para *ponto*;
- *multipoint* para múltiplos pontos – *multiponto*;
- *linestring* para *linha*;
- *multilinestring* para múltiplas linhas – *multilinha*;
- *polygon* para *polígono*;
- *multipolygon* para múltiplos polígonos – *multipolígono*;
- *geometrycollection* para *geometria complexa*.

Acrescenta-se, ainda, que os sete tipos são derivados de três tipos básicos, a saber: o *ponto*, a *linha* e o *polígono*. Neste caso, estes três tipos básicos são definidos da seguinte forma:

1. um *ponto* v é um par ordenado de coordenadas (x_v, y_v);
2. uma *linha* L é um conjunto de pontos ordenados $v_0, v_1, \ldots, v_n$ que subentendem a linha poligonal aberta composta pelos segmentos de reta $\overline{v_0 v_1}$, $\overline{v_1 v_2}$, $\ldots$, $\overline{v_{n-1} v_n}$. A linha existe se, e somente se, a linha não possui auto-interseção;

3. um *polígono* P é uma região do plano delimitada por uma linha poligonal fechada, ou seja, $v_0 = v_{n+1}$.

Os tipos *multiponto*, *multilinha* e *multipolígono* são, na verdade, uma coleção dos tipos básicos. Por sua vez, a *geometria complexa* é o tipo que descreve a combinação entre diferentes tipos.

3.3 Classes de equivalência

Em um BDG, cada feição F_i pode ser representada em um *dataset* de diversas formas. Sua representação ϕ_i, por sua vez, pode ser vazia, ou seja, caso $\phi_i = 0$ há uma indicação falha na cobertura. A representação ϕ_i, pode possuir uma única tupla ou pode ter sido particionada em várias tuplas da tabela. Independentemente do possível particionamento, nesta tese, será admitido que ϕ_i é a única representação de F_i e que a mesma ocupa apenas uma única tupla no *dataset*. Assim, ϕ_i será uma tupla do tipo $(nome_i, geometria_i)$. Logo, haverá apenas um ϕ_i em T para cada F_i do mundo real.

Na prática, o mundo real é modelado por vários produtores de dados. Portanto, há várias funções de representação Φ_j, para $j = 1 \ldots m$, uma para cada produtor. Destarte, cada feição F_i do mundo real pode possuir mais de uma representação ϕ_i. Neste caso, há que se considerar um ϕ_{ij}, onde j representa o índice da função de representação Φ_j e i representa o índice da feição F_i do mundo real na tabela T_j (Figura 3.1).

Para se mapear as correspondências entre as representações é vital a identificação de similaridade entre elas. Em outras palavras, chamamos de correspondência entre duas representações ϕ_{ij} e ϕ_{ik} o fato de serem suficientemente *similares* entre si. O critério de similaridade é abordado na seção 3.4.

Assim, dados dois conjuntos T_1 e T_2 com dados potencialmente ambíguos, é possível verificar a existência de algumas possibilidades de correspondência (Figura 3.2). Pode ser observado que uma possibilidade tem a ver com a identificação unívoca dentre as representações, ou seja, cada ϕ_{ij} em um *dataset* corresponde a apenas um outro na base de dado distinta (Figura 3.2.a). Entretanto, há os casos em que uma representação não possui uma correspondência unívoca com outra

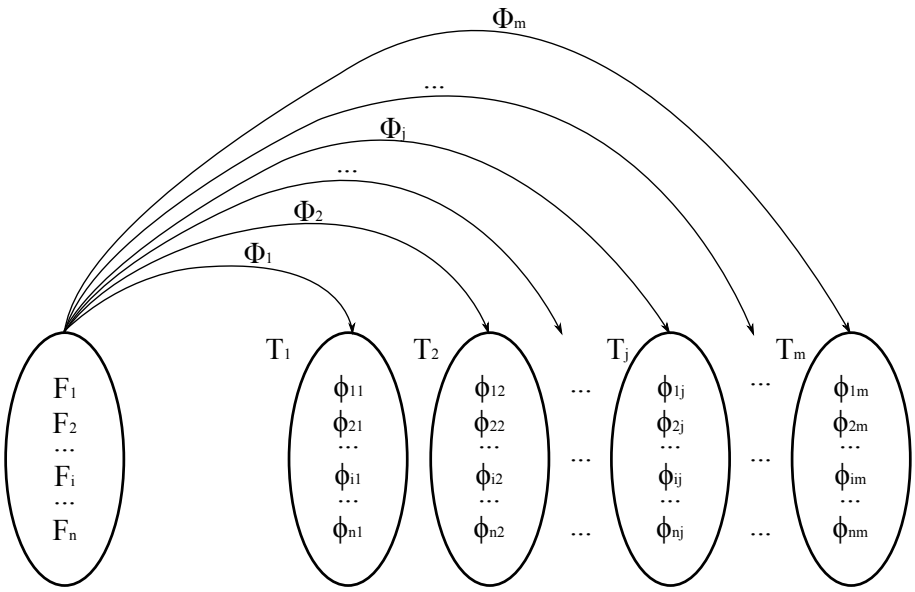

Figura 3.1: Esquema conceitual

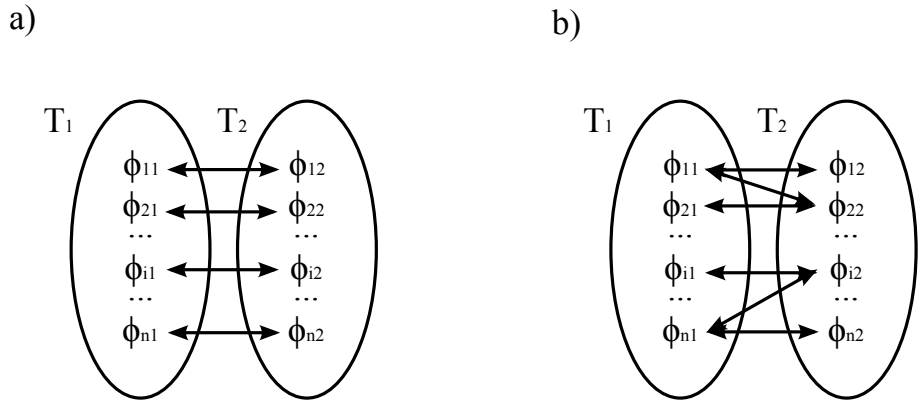

Figura 3.2: Possibilidades de correspondência

(Figura 3.2.b).

Ressalta-se, ainda, que $|T_1|$ não necessariamente é igual a $|T_2|$. Assim, se considerarmos para efeito de álgebra de conjuntos que dois elementos são equivalentes se são suficientemente similares, então pode ocorrer de $T_1 - T_2 \neq \emptyset$ e $T_2 - T_1 \neq \emptyset$. Logo, é possível ter-se os três subconjuntos abaixo a partir dos *datasets* disponíveis:

- $T_1 \cap T_2$;
- $T_1 - T_2$; e,
- $T_2 - T_1$.

Seja cada representação ϕ_{ij} como um nó em um grafo, sendo suas arestas as relações de correspondência. Se admitirmos a existência de n datasets, podemos

tentar inferir as feições do mundo real através de uma análise desse grafo. Em particular, se n representações $\phi_{i1}, \cdots, \phi_{in}$ se referem a uma mesma feição F_i, então ϕ_{ij} e ϕ_{ik} devem ser suficientemente *similares* para quaisquer $1 \leq j, k \leq n$. Em outras palavras, os nós referentes a $\phi_{i1}, \cdots, \phi_{in}$ devem formar uma *clique* do grafo, isto é, um subgrafo totalmente conexo.

Outra característica importante das relações de correspondência é o fato que duas representações ϕ_{ik} e ϕ_{jk} pertencentes a um mesmo dataset T_k *não podem* ser suficientemente similares. Isto significaria que uma mesma feição aparece duas vezes no mesmo *dataset*, revelando um erro de modelagem.

Voltando à analogia com grafos, se considerarmos a existência de dois *datasets* T_1 e T_2 para o mesmo tema, as relações de correspondência devem produzir um grafo bipartite, isto é, ele pode ser dividido em dois conjuntos (T_1 e T_2), de tal forma que arestas $a - b$ só existam se $a \in T_1$ e $b \in T_2$. Analogamente, para n datasets, o grafo correspondente deve ser n-partite.

Considere agora a situação ilustrada na Figura 3.2.b. Neste caso, temos, por exemplo, que a representação ϕ_{11} corresponde simultaneamente a ϕ_{12} e ϕ_{22}. Pelo critério discutido anteriormente, (ϕ_{11}, ϕ_{12}) e (ϕ_{11}, ϕ_{22}) formam duas cliques, ou seja, se referem a duas feições distintas. Esta, obviamente, não é uma situação desejável, indicando um critério de similaridade excessivamente permissivo. No decorrer deste trabalho verificou-se que o projeto das métricas de similaridade espaciais tenderá a eliminar este problema ao requerer uma cobertura do *locus* geográfico comum superior a um limiar proposto para que duas representações sejam julgadas suficientemente similares no aspecto geométrico.

Outra situação indesejável é mostrada na Figura 3.3.b, onde intuitivamente tenderíamos a considerar as representações ϕ_{11}, ϕ_{12} e ϕ_{13} como referentes a uma mesma feição, mas ϕ_{12} e ϕ_{13} não são suficientemente similares. Neste caso, não há necessariamente problemas com a modelagem dos *datasets* ou com o critério de similaridade. O que se propõe então é considerar durante o processamento de consultas que há duas possibilidades de feições do mundo real, representadas respectivamente pelos pares (ϕ_{11},ϕ_{12}) e (ϕ_{11},ϕ_{13}). De maneira análoga, representações sem correspondências serão consideradas referentes a feições não modeladas nos demais *datasets*. Por sua vez, no caso da Figura 3.3.a não há problemas na identificação das correspondências,

pois todos os *datasets* possuem uma representação associada a uma feição do mundo real.

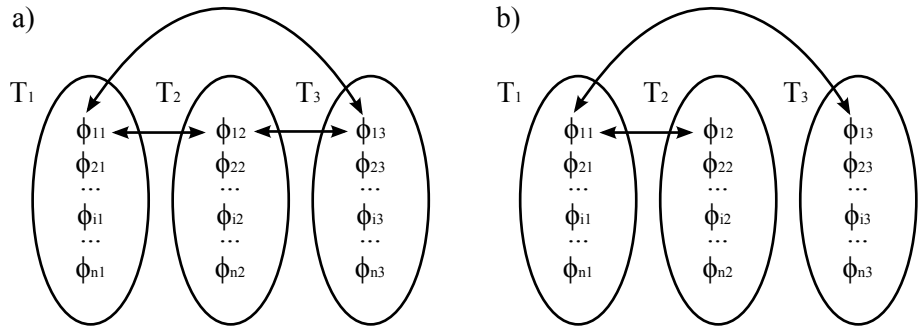

Figura 3.3: Representações ϕ_{11}, ϕ_{12} e ϕ_{13} referem-se a uma única feição em (a), mas a duas feições distintas em (b).

3.4 Mapeamento de correspondência

Diante da existência de representações potencialmente ambíguas, há que se desenvolver um procedimento capaz de identificar os pares das mesmas que são modelos diferentes da feição. Assim, é necessário aplicar um índice que seja capaz de qualificar se, *par-a-par*, as representações referem-se à mesma feição. Desta forma, é viável a aplicação de uma função de similaridade $\mathcal{S}$ para a mesma. Como a tabela possui dois atributos – *nome* e *geometria* –, convém aplicar duas funções de similaridade. Portanto, nesta tese, a função de similaridade para o atributo *nome* é definida, genericamente, pela função $\mathcal{S}_n$ e a função de similaridade para o atributo *geometria* é definida, genericamente, pela função $\mathcal{S}_g$.

Tanto $\mathcal{S}_n$ como $\mathcal{S}_g$ estão compreendidas no intervalo $[0, 1]$. Assim, o valor 0 representa uma total dissimilaridade e o valor 1 uma similaridade perfeita, ou seja, a igualdade (identidade). A igualdade é encontrada quando há igualdade no atributo *nome* e no atributo *geometria* simultaneamente. Pode ocorrer, ainda, a percepção de uma total dissimilaridade, ou seja, *nome* e *geometria* com valores para a função de similaridade iguais a 0.

Assim, é preciso quantificar *limiares mínimos* $\mathcal{L}_n$ e $\mathcal{L}_g$ para as similaridades entre nomes e entre geometrias. Desta forma, definimos a função lógica $similar(\phi_i, \phi_j)$ de tal forma que $similar(\phi_i, \phi_j)$ seja verdadeira se e somente se $\mathcal{S}_n(N(\phi_i), N(\phi_j)) \geq \mathcal{L}_n$

e $\mathcal{S}_g(G(\phi_i), G(\phi_j)) \geq \mathcal{L}_g$.

Observe que a função $\mathcal{S}$ é complementar de uma métrica. Segundo Lima [35], um espaço métrico (X, f), é um conjunto X que possui uma distância (ou métrica) f, onde $f : X \times X \to \mathbb{R}, \forall x, y, z \in X$. Logo, a função $f(x, y) = 1 - \mathcal{S}(x, y)$ deve possuir as seguintes propriedades:

- $f(x, y) \geq 0$: positividade
- $f(x, y) = 0 \Leftrightarrow x = y$: identidade
- $f(x, y) = f(y, x)$: simetria
- $f(x, z) \leq f(x, y) + f(y, z)$: desigualdade triangular

Correspondência não espacial

Os *nomes* são conjuntos de caracteres. Assim, a correspondência entre eles é obtida ao se aplicar uma métrica de *strings* para quantificar a distância ou similaridade entre dois *nomes* quaisquer. Existem diversas formas de se proceder esta análise [36]. Neste tese, será atribuída uma metodologia clássica para a identificação da correspondência, atendendo às propriedades da função associada a um espaço métrico. A função de similaridade não espacial $\mathcal{S}_n$ será descrita no capítulo 4.

Correspondência espacial

Ao se considerar a possibilidade de existência de dados ambíguos, é possível observar que tais dados possuem uma geometria que pode diferir da outra representação da feição. Para tal, é preciso identificar a existência de algumas possibilidades de combinação das representações ambíguas quanto à geometria dos dados, a saber:

- ponto *versus* ponto;
- ponto *versus* linha;
- ponto *versus* polígono;
- linha *versus* linha;
- linha *versus* polígono; e,

- polígono *versus* polígono.

No contexto desta tese, para efeito de prova de conceito, serão consideradas apenas as relações entre geometrias do mesmo tipo. Não há, entretanto, perda de generalidade quanto à tese e a simplificação apenas favorece o entendimento da metodologia subsequente. A função de similaridade espacial $\mathcal{S}_g$ também será apresentada no capítulo 4.

3.5 Estruturas de dados

O mapeamento de correspondências discutido na seção anterior precisa ser registrado em estruturas de dados apropriadas de forma a ser utilizado durante o processamento de consultas ao banco de dados multirepresentados. Deste modo, assume-se que uma feição do mundo real é associada a um conjunto de correspondências. Idealmente, para n *datasets*, uma dada feição é representada n vezes, uma em cada *dataset* e tem-se $n(n-1)/2$ relações de correspondência entre essas representações. Isto significa que o armazenamento explícito de todas as relações de correspondência requer espaço $O(n^2)$.

Uma idéia alternativa é construir uma representação aproximada de cada feição levando em conta todas as representações explícitas nos n *datasets*. A esta estrutura de dados dá-se o nome de *tabela de feições* ou, simplesmente, TF. Desta forma, o processamento de uma consulta envolvendo o tema multirepresentado pode se utilizar da TF como uma espécie de filtro capaz de localizar feições que atendam a algum predicado da consulta.

A TF é uma tabela onde cada linha se refere a uma potencial feição do mundo real descrita por aproximações dos atributos *nome* e *geometria*. A tabela, portanto, contém as seguintes colunas:

- id_F: chave primária de uma feição F_i;
- $NOME_M$: um valor médio para o atributo não espacial de F_i;
- $GEOM_M$: uma estimativa conservadora (*bounding box*) para o atributo espacial de F_i;

Nesta tabela, a coluna id_F é um inteiro sequencial identificando uma feição do mundo real. A segunda coluna, por sua vez, é preenchida por uma *string* média entre aquelas existentes nos *datasets* originais. Neste caso, é utilizada a técnica desenvolvida por Zell [37] para o estabelecimento de uma *string* média a partir de um conjunto destas. A última coluna armazena a caixa envolvente das geometrias correspondentes à feição.

O mapeamento entre cada dataset T_i e a tabela de feições TF é realizado através de uma tabela auxiliar AUX_i com as seguintes colunas:

- id_F – identificador da feição;
- id_{T_i} – identificador na tabela do produtor do dado.

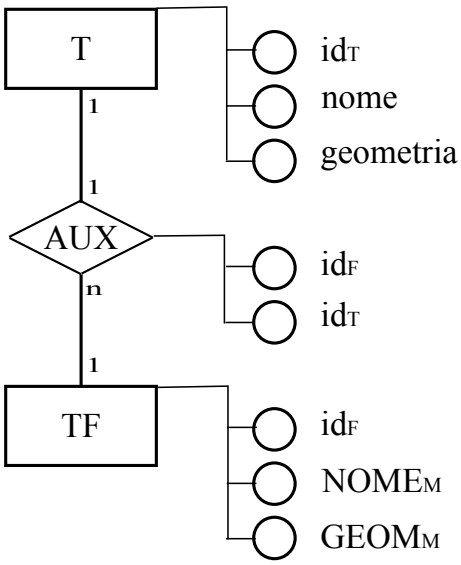

Figura 3.4: Mapeamento entre as tabelas T, AUX e TF

As tabelas AUX_i são construídas à medida que os diferentes *datasets* são incluídos na arquitetura. Assim, para cada *dataset* inserido é criada uma tabela AUX_i e a tabela TF é atualizada. Ao ser inserido o primeiro *dataset* – T_1 –, a arquitetura constrói as duas tabelas – AUX_1 e TF (Algoritmo 1). Neste primeiro momento, é possível constatar que as tabelas AUX_1 e TF possuem a mesma quantidade de elementos que a tabela T_1 original (Figura 3.4).

Ao se acrescentar outro *dataset* – T_2 –, a arquitetura acessa e atualiza a tabela de feições em função da similaridade entre os *datasets* originais e cria outra tabela

entrada: $T_1(id_T, nome, geometria)$

saida : $TF(id_F, nome, geometria)$, $AUX_1(id_F, id_T)$

begin

- `createtable` (TF)
- `createtable` (AUX_1)
- **for** t **in** T_1 **do**
  - `inserir` (TF, [$t.id_T$, $t.nome$, ***bounding_box*** (t.geometria)])
  - `inserir` (AUX_1, [$t.id_T$, $t.id_T$])

Algoritmo 1: Construção da tabela de feições (TF) e primeira tabela auxiliar (AUX_1)

auxiliar (AUX_2). Evidentemente, à medida que se acrescentam *datasets*, a tabela TF é atualizada e as tabelas AUX_i são criadas (Figura 3.5).

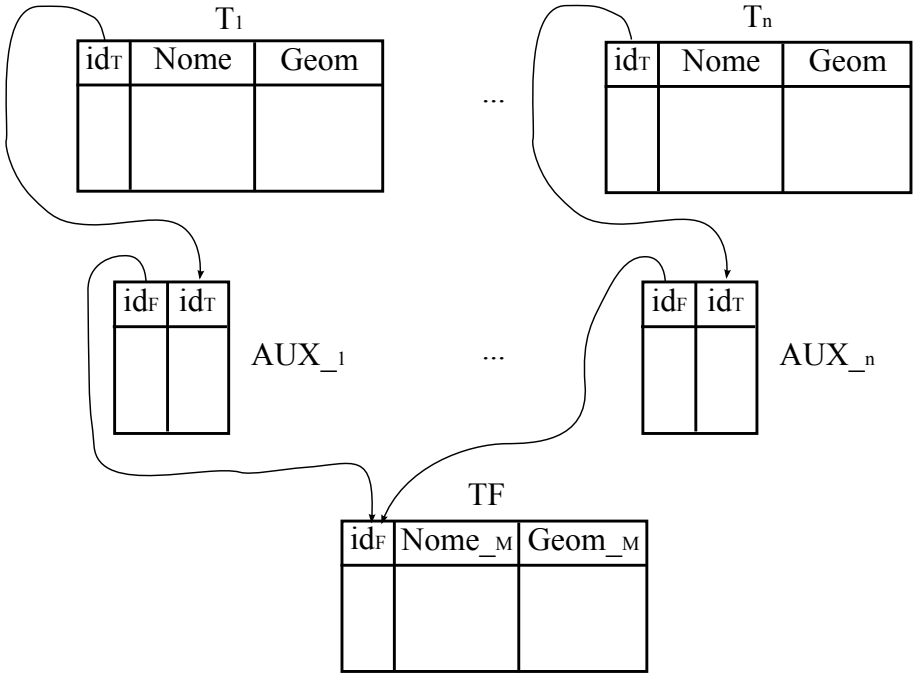

Figura 3.5: Mapeamento geral entre as tabelas

Para a obtenção de uma TF que represente todos os *datasets* disponíveis, é necessário verificar todas as possibilidades de similaridade entre as representações existentes. Assim, à medida que se acrescentam *datasets* T_i é necessário elencar os potenciais candidatos de pares ambíguos. Para tal, realiza-se uma operação de junção entre as tuplas m de T_i e n de TF, considerando a junção a partir da função

similar, e a projeção do resultado, visando criar uma tabela $CAND$ que possua os atributos iguais aos das tabelas AUX_i.

$$CAND \leftarrow \pi_{(id_F, id_T)}(T_i \bowtie_{similar(m,n)} TF)$$

Observe que a função *similar*(m,n) é verdadeira quando o produto das funções $\mathcal{S}_n$(m.nome,n.nome) com a função $\mathcal{S}_g$(m.geometria,n.geometria) é superior a um limiar $\mathcal{L}$.

De posse dos possíveis candidatos a representações ambíguas, é necessário avaliar as possibilidades com o intuito de se identificar aquelas que atenderão a todos os critérios de similaridade. Para tal, são avaliados os atributos *nome* e *geometria* dos pares elencados para que se possa considerar ou não um par de representações ambíguo (Algoritmo 2).

A função `bounding_box` encontrada no Algoritmo 2 é, em verdade, uma ferramenta que permite encontrar a *caixa envolvente* da união de todas as geometrias encontradas no atributo geométrico de uma tabela. Por sua vez, a função `mean_name` retorna uma "*string* média" a partir de um conjunto de *strings*. Neste caso, é gerada uma média entre todos os registros do atributo *nome* encontrado na tabela. Este nome médio é obtido por meio da metodologia desenvolvida por Zell [37], que pode ser sumarizada da seguinte maneira:

1. Sugerir uma *string* aproximada a partir da análise do conjunto disponível (possivelmente, escolher uma *string* qualquer do conjunto);

2. Percorrer o conjunto comparando cada *string* com aquela aproximada usando a métrica sugerida por Levenshtein [38], onde as disisimilaridades correspondem a operações de transformação tais como inserção de um caractere, remoção de um caractere ou transposição de caracteres;

3. Para cada posição da *string* média aplicar a transformação mais frequente gerando assim, uma nova *string* média;

4. Repetir os passos 2 e 3 até que o resultado obtido retorne um número mínimo de transformações.

A função `createtable` cria uma tabela no sistema. Desta forma, é possível integrá-la a arquitetura desenvolvida. Para se eliminar a tabela do sistema, tem-se

entrada: $T_i(id_T, nome, geometria)$, $TF(id_F, NOME_M, GEOM_M)$,
$T_j(id_T, nome, geometria)$, $AUX_j(id_F, id_T), j = 1 \ \ldots \ (n-1)$

saida : $TF(id_F, NOME_M, GEOM_M)$, $AUX_i(id_F, id_T)$

begin

- `createtable` (AUX_i)
- // obter feições similares em TF
- $CAND \leftarrow \pi_{(id_F, id_T)}(T_i \bowtie_{similar(m,n)} TF)$, $m \in T_i$ e $n \in TF$
- **for** t **in** $CAND$ **do**
  - k $\leftarrow 0$
  - `createtable` $(TEMP)$
  - $t_i \leftarrow \sigma_{id_T = t.id_T} T_i$
  - `inserir` $(TEMP, [i, t_i])$
  - // obter representações similares nos demais *datasets*
  - **for** j **in** $1 \ldots n$, $j \neq i$ **do**
    - $h \leftarrow \sigma_{id_F = t.id_F} AUX_j$
    - $t_j \leftarrow \sigma_{id_T = h.id_T} T_j$
    - **if** $similar(t_i, t_j)$ **then**
      - `inserir` $(TEMP, [j, t_j])$
      - k $\leftarrow$ k+1
  - $[NOME_M, GEOM_M] \leftarrow$ [`mean_name`$(TEMP)$, `bounding_box`$(TEMP)$]
  - **if** $k = n - 1$ **then**
    - // t_i forma clique com tuplas dos demais *datasets*
    - `inserir` $(AUX_i, [t.id_F, t.id_T])$
    - `atualizar` $(TF, [id_F, NOME_M, GEOM_M])$
  - **else**
    - // t_i define nova feição
    - **for** $[j, t_j]$ **in** $TEMP$ **do**
      - `inserir` $(AUX_j, [|TF| + 1, t_j.id_T])$
    - `inserir` $(TF, [|TF| + 1, NOME_M, GEOM_M])$
  - `deletetable` $(TEMP)$

Algoritmo 2: Atualização da tabela de feições e criação da tabela AUX_i

a função `deletetable` A função `inserir` povoa a tabela com uma tupla de cada vez, enquanto a função `atualizar` modifica o conteúdo de uma tupla específica.

Após a inserção de n *datasets*, tem-se um total de $2 \cdot n + 1$ tabelas no sistema, a saber, n tabelas T_i, n tabelas AUX_i e uma tabela de feições TF. O total de registros na matriz TF está vinculado ao total de feições do mundo real mapeadas nos *datasets* originais. Desta forma, é possível se avaliar a cobertura [7] de um *dataset* T_i específico ao se proceder o seguinte cálculo (Equação 3.1):

$$\mathfrak{Co}(T_i) = \frac{|T_i|}{|TF|} \tag{3.1}$$

A quantidade de registros na matriz TF, por sua vez, é um indicativo de quantas feições do mundo real foram modeladas nos diversos *datasets* disponíveis. Assim, os casos extremos seriam a ocorrência de um total de registros nos *datasets* igual à quantidade encontrada na tabela de feições, caracterizando a não existência de ambiguidade nos dados disponíveis, e a possibilidade de se encontrar o total de registros individuais dos *datasets* iguais aos da tabela TF, ocorrendo, então, a identificação de similaridades em cliques.

3.6 Considerações finais

O mapeamento das correspondências são o cerne desta tese. Isto porque a sua correta identificação favorece a percepção da similaridade entre as representações que podem ser consideradas semelhantes para o processamento de consultas. Evidentemente, as tuplas na tabela de feições que não possuam relações de correspondência mapeadas podem, ainda assim, ser ambíguas. Entretanto, este caso somente pode ser identificado por uma inspeção de um usuário habilitado.

Em verdade, como o processo de construção e manutenção das tabelas de feições (TF_i) é fortemente influenciado pelos valores estipulados para os limiares $\mathcal{L}_n$ e $\mathcal{L}_g$, sugere-se que o usuário responsável por essas tarefas experimente um intervalo relativamente amplo de valores para tais constantes. Por exemplo, poder-se-ia utilizar valores bastante estritos tais como 0.900, sendo estes progressivamente relaxados até, digamos, 0.500, observando o resultado obtido a cada passo. Espera-se que um usuário habilitado possa detectar através dessa prática os valores mais apropriados

para os dados em questão. É mesmo concebível que um tal usuário possa interferir manualmente no sentido de registrar uma correspondência não obtida pelo sistema.

Capítulo 4

Similaridade

4.1 Considerações iniciais

Uma funcionalidade importante no contexto desta tese consiste em avaliar se representações em *datasets* distintos se referem ou não a mesma feição. No capítulo 3, conforme apresentado, este julgamento está apoiado no conceito de similaridade. Tal similaridade, portanto, deve possuir a característica de permitir a inferência de uma igualdade.

Neste capítulo, discutiremos as funções de similaridade $\mathcal{S}_n$ e $\mathcal{S}_g$ para os atributos *nome* e *geometria*, respectivamente. Para ambas são analisadas várias métricas candidatas, comparadas e escolhidas as mais adequadas com vistas à aplicação no sistema proposto. Concomitantemente, são apresentados os valores para os limiares $\mathcal{L}_n$ e $\mathcal{L}_g$.

4.2 A similaridade

Similaridade é, em verdade, uma qualidade ou um caráter de algo que possui a mesma natureza, a mesma função ou, ainda, o mesmo efeito. Neste caso, este trabalho se apropria do vocábulo similaridade com o intuito de considerar representações de mesma natureza, ou seja, *locus* geográficos e nomes semelhantes.

Pode-se considerar a ambiguidade, no contexto desta tese, como uma similaridade imperfeita. Em outras palavras, quando duas representações são idênticas ou totalmente díspares, não há ambiguidade, pois no primeiro caso é a mesma re-

presentação e no segundo são feições do mundo real diferentes. O problema é a caracterização de situações entre estes dois extremos.

4.3 Parâmetros de avaliação de similaridade

4.3.1 Métodos para a avaliação do *nome*

O *nome* é um atributo identificador da feição. Entretanto, um problema que pode ocorrer é a mesma feição receber, em *datasets* distintos, *nomes* diferentes mesmo que similares em algum grau. Assim sendo, faz-se necessário elaborar um procedimento para avaliar e quantificar o quanto uma *string* é semelhante a outra. Seja $S_{ij} = N(\phi_{ij})$ a *string* correspondente ao nome da representação.

A literatura discute várias opções para a consecução deste objetivo. Dentre estas, destacam-se as seguintes:

- Distância de Damerau-Levenshtein;
- Coeficiente de Dice;
- Distância de Hamming;
- Distância de Jaro-Winkler; e,
- Coeficiente de *Overlap*.

Para facilitar a compreensão do texto, são consideradas as *strings* abaixo, para servirem de exemplo na apresentação dos métodos que se seguem:

- S_1 = "casa"
- S_2 = "casal"
- S_3 = "casa␣"

Distância de Damerau-Levenshtein

Neste método são contadas quantas operações são realizadas para transformar S_1 em S_2 [39]. As operações consideradas consistem em deleção, inserção, substituição de um simples caractere ou, ainda, transposição entre dois caracteres. Esta métrica

é uma generalização da distância de Levenshtein que não prevê a transposição entre caracteres [38].

Assim, dadas S_1 e S_2, tem-se a seguinte operação:

- remoção do "*l*".

Dadas S_1 e S_3, tem-se:

- remoção do ␣".

E, finalmente, dadas S_2 e S_3, tem-se:

- remoção do ␣";
- inserção do "*l*".

Diante do exposto, a distância de Damerau-Levenshtein (d_l) entre as opções serão as seguintes:

- $d_l(S_1, S_1) = 0$;
- $d_l(S_1, S_2) = 1$;
- $d_l(S_1, S_3) = 1$; e,
- $d_l(S_2, S_3) = 2$.

Coeficiente de Dice

O coeficiente de Dice (d_d) mede a similaridade de acordo com o índice de Jaccard [40]. Para tal, o valor calculado é dado por (Equação 4.1)

$$d_d = \frac{2 \cdot n_t}{n_x + n_y}, \tag{4.1}$$

onde n_t é o número de bigramas comuns S_1 e S_2, n_x é a quantidade de bigramas em S_1 e n_y é o número total de bigramas em S_2 [41].

Assim, dadas S_1, S_2 e S_3, tem-se os respectivos bigramas:

- {ca, as, sa}
- {ca, as, sa, al}

- {ca, as, sa, a␣}

Observa-se, que para S_1 e S_2 tem-se: $n_t = 3$, $n_x = 3$ e $n_y = 4$. Para o par S_1 e S_3 tem-se: $n_t = 3$, $n_x = 3$ e $n_y = 4$. Finalmente, para S_2 e S_3 temos: $n_t = 3$, $n_x = 4$ e $n_y = 4$.

Assim, o Coeficiente de Dice assume os seguintes valores:

- $d_d(S_1, S_1) = 1,000$;
- $d_d(S_1, S_2) = 0,857$;
- $d_d(S_1, S_3) = 0,857$; e,
- $d_d(S_2, S_3) = 0,750$.

Distância de Hamming

A distância de Hamming entre duas *strings* é dada pelo número de posições nas quais os conjuntos X e Y são diferentes [42]. Assim sendo, este apenas avalia a distância entre *strings* de igual comprimento, ou seja, $|S_i| = |S_j|$. Neste caso, para os exemplos dados a análise por Hamming somente pode ser realizada para as *strings* S_2 e S_3. O valor da distância é obtida pelo somatório dos valores obtidos pela comparação. No caso, considera-se o valor 0 quando os caracteres forem iguais e 1 quando forem diferentes. Assim, dadas S_2 e S_3, obtém-se distância igual a 1, conforme pode ser observado abaixo:

- S_2 =“c” “a” “s” “a” “l”
- S_3 = “c” “a” “s” “a” “␣”
- “0” “0” “0” “0” “1” (comparação)

Distância de Jaro-Winkler

A distância de Jaro-Winkler [43] é uma medida de similaridade entre duas *strings* onde, quanto mais similar elas sejam, mais próximo do valor 1 será o resultado da distância. Assim sendo, tem-se os valores 0 para uma total dissimilaridade e 1 para uma similaridade perfeita.

É baseada na distância de Jaro (d_j), cuja métrica é dada por (Equação 4.2)

$$d_j(S_i, S_j) = \frac{1}{3}\left(\frac{m}{|S_i|} + \frac{m}{|S_j|} + \frac{m-t}{m}\right), \tag{4.2}$$

onde, m é o número de caracteres iguais dentro da janela de busca, t o número de transposições necessárias e $|S_i|$ é a norma – quantidade de caracteres – da *string*.

Este método busca caracteres similares dentro de uma janela de pesquisa, geralmente de 3 (três). Assim, dadas S_1 e S_2, tem-se os seguintes dados, conforme apresentado da Tabela 4.1:

Tabela 4.1: Exemplo de análise Jaro

	"c"	"a"	"s"	"a"
"c"	1	0	0	–
"a"	0	1	0	1
"s"	0	0	1	0
"a"	–	1	0	1
"l"	–	–	0	0

Tabela semelhante deve ser desenvolvida para os pares de *strings* S_1/S_3 e S_2/S_3. De posse da Tabela 4.1, e das outras, tem-se os seguintes dados (Tabela 4.2):

Tabela 4.2: Valores inferidos para cálculo da Distância Jaro

variável	S_1/S_2	S_1/S_3	S_2/S_3
m	4	4	4
$\|S_i\|$	4	4	5
$\|S_j\|$	5	5	5
t	1	1	1
$d_j(S_i, S_j)$	0,850	0,850	0,783

A distância de Jaro-Winkler (d_w), por sua vez, é dada por (Equação 4.3):

$$d_w(S_i, S_j) = d_j(S_i, S_j) + (\ell \cdot p \cdot (1 - d_j(S_i, S_j))), \tag{4.3}$$

onde, d_j é a distância de Jaro, ℓ é comprimento fixo de caracteres iniciais iguais e p é uma constante de valor 0.1. Assim sendo, dadas as *strings* S_1, S_2 e S_3 tem-se:

- $d_w(S_1, S_1) = 0,981$;

- $d_w(S_1, S_2) = 0.910$;
- $d_w(S_1, S_3) = 0.910$; e,
- $d_w(S_2, S_3) = 0.870$.

Coeficiente de *Overlap*

O coeficiente de *Overlap* (d_o) é uma medida de similaridade baseada no índice de Jaccard [40] que avalia a sobreposição de uma *string* sobre a outra (Equação 4.4). É definida por

$$d_o(S_i, S_j) = \frac{|S_i \cap S_j|}{\min(|S_i|, |S_j|)}. \tag{4.4}$$

Observe que, neste caso, *strings* são consideradas conjuntos de caracteres. Implicando, por exemplo, que anagramas são idênticos entre si.

Caso S_i seja um subconjunto de S_j, o coeficiente assumirá o valor 1 (Equação 4.5), pois

$$d_o(S_i, S_j) = \frac{|S_i \cap S_j|}{\min(|S_i|, |S_j|)} = \frac{|S_i|}{|S_i|} = 1. \tag{4.5}$$

Assim, para as *strings* S_1, S_2 e S_3 tem-se:

- $d_o(S_1, S_1) = 1,0$;
- $d_o(S_1, S_2) = 1,0$;
- $d_o(S_1, S_3) = 1,0$; e,
- $d_o(S_2, S_3) = 0,8$.

Comparação dos métodos de avaliação de nomes

Ao se proceder uma análise nos métodos apresentados, sobretudo comparando-os com as propriedades dos espaços métricos, conforme visto na seção 3.4, tem-se as seguintes observações:

- A Distância de Damerau-Levenshtein atende integralmente as propriedades;
- O Coeficiente de Dice não atende a propriedade da identidade;
- A Distância de Hamming atende apenas para *strings* com o mesmo tamanho;

- A Distância de Jaro-Winkler não atende a propriedade da identidade; e,
- O Coeficiente de Overlap não atende a propriedade da identidade.

Destarte, apenas a Distância de Damerau-Levenshtein atende às propriedade. As demais, não podem ser considerados métricas. Entretanto, para os propósitos desta tese, a distância que atende às propriedades não fornece valores no intervalo $[0, 1]$. Assim, esta não pode ser aceita como a função de similaridade $\mathcal{S}_n$. Dentre as opções, a Distância de Hamming também não serve como função de similaridade porque trata apenas *nomes* de igual quantidade de caracteres.

As três outras opções, não são métricas porque não atendem ao critério da identidade. Porém, é possível adaptar uma função f que seja complementar delas. Assim tem-se as seguintes opções:

- $f_d = 1 - d_d$;
- $f_w = 1 - d_w$; e,
- $f_o = 1 - d_o$.

Ao se proceder a análise destas novas opções tem-se:

- A função complementar da Distância de Dice atende as propriedades;
- O complementar da Distância de Jaro-Winkler não atende a propriedade da identidade; e,
- A função complementar do Coeficiente de Overlap atende as propriedades.

Diante das possibilidades, optou-se por utilizar o Coeficiente de Dice para avaliar pares de nomes. Assim, a função de similaridade não espacial $\mathcal{S}_n$ assume o seguinte valor (Equação 4.6).

$$\mathcal{S}_n(N(\phi_{ij}), N(\phi_{ik})) = d_d(N(\phi_{ij}), N(\phi_{ik})) \tag{4.6}$$

Desta forma, a função $\mathcal{S}_n$ gera valores dentro de um intervalo adequado, no caso, entre 0 e 1 e sua complexidade computacional é baixa.

4.3.2 Métodos para a avaliação da *geometria*

A *geometria* é o atributo espacial da feição. Quando uma mesma feição recebe em *datasets* distintos representações diferentes, mesmo que similares em algum grau, torna necessário a elaboração de um procedimento para avaliar e quantificar o quanto uma dada *geometria* é semelhante a outra.

A literatura discute algumas opções para a avaliação de *geometrias*. Dentre estas, destacam-se as seguintes:

- Método dos Retângulos Equivalentes – MRE;
- Método dos Retângulos Equivalentes Adaptado – MREA; e,
- Índice de Similaridade Cartográfico – ISC.

Método dos Retângulos Equivalentes

Um método que viabiliza a comparação de *geometrias* é o Método dos Retângulos Equivalentes (MRE) [4]. O método foi concebido para ser aplicado a *geometrias* do tipo linha e multilinha. O uso do MRE como avaliador da multirepresentação deste tipo de geometria permite a inferência de um afastamento médio entre elas.

Como o MRE serve como avaliador da discrepância entre as representações lineares, na realidade, ele tenta inferir a distância média entre as representações de uma mesma feição ao quantificar a área e o semi-perímetro de um retângulo equivalente (Figura 4.1).

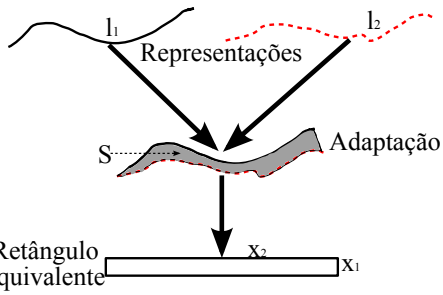

Figura 4.1: Representações lineares usadas para computar o retângulo equivalente

Neste caso, o valor deste afastamento é obtido em unidades métricas e não em percentuais de similaridade. O MRE serve como qualificador das linhas e não como

comparador de similaridade. O avaliador é baseado numa equação quadrática relacionando a área (S) e o semi-perímetro médio (P) de um, assim chamado, retângulo equivalente gerado a partir de duas representações (Equação 4.7).

$$x^2 + S \cdot x + P = 0. \tag{4.7}$$

Para o MRE considera-se como afastamento médio a menor raiz desta equação, ou seja, a solução $x = \frac{-S-\sqrt{S^2-4 \cdot P}}{2}$. Onde, para o MRE, S é a área obtida entre as representações e P é o semi-perímetro da figura resultante da união entre as representações.

Método dos Retângulos Equivalentes Adaptado

No caso particular de as geometrias serem polígonos, ou seja, linhas poligonais fechadas, o MRE pode ser adaptado [5]. Criando, assim, o chamado Método dos Retângulos Equivalentes Adaptado ($MREA$) (Figura 4.2).

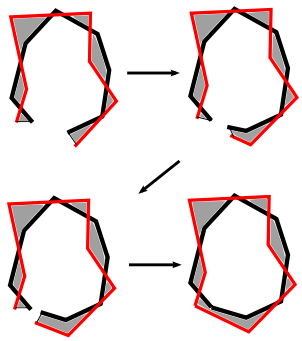

Figura 4.2: Adaptação do MRE para um par de representações poligonais

O $MREA$ utiliza-se dos mesmos princípios que o MRE. Assim, a equação quadrática (Equação 4.7) e a solução pela menor raiz são igualmente válidos. Como o $MREA$ trata polígonos, não há a necessidade de se proceder a adaptação que o MRE exige, ou seja, a junção dos extremos das linhas. Neste caso, o $MREA$ irá tratar a diferença de *locus* geográfico entre as representações poligonais. Por conseguinte, os valores utilizados na equação devem ser ajustados para os valores de S como a área obtida pela diferença da união e da interseção das representações poligonais e P como perímetro das representações.

O $MREA$, assim como o MRE, permite a avaliação da discrepância entre as representações e não um avaliador de similaridade.

Índice de Similaridade Cartográfico

Como forma de se avaliar diretamente a similaridade entre representações poligonais, existe o Índice de Similaridade Cartográfico (ISC) [5]. Este índice parte de uma idéia apresentada por Ali [6] e expandida por Sester [7]. Baseia-se no processamento da união e da interseção das representações. Assim, a avaliação dessas duas regiões permite inferir uma similaridade entre as *geometrias* que lhes deram origem.

A união é fundamental para se identificar a região máxima dentre duas *geometrias* disponíveis ($U = G(\phi_{ij}) \cup G(\phi_{ik})$), enquanto a interseção – $I = G(\phi_{ij}) \cap G(\phi_{ik})$ – serve como identificador da região mínima.

A comparação entre a união e a interseção das representações quantifica a similaridade da seguinte forma (Eq. 4.8):

$$ISC(G(\phi_{ij}), G(\phi_{ik})) = \frac{AREA(I)}{AREA(U)} = \frac{AREA(G(\phi_{ij}) \cap G(\phi_{ik}))}{AREA(G(\phi_{ij}) \cup G(\phi_{ik}))} \tag{4.8}$$

Da análise de 4.8 tem-se que o $ISC \in [0, 1]$. Logo, quando $ISC = 0$ ter-se-á uma total dissimilaridade ($\mathrm{G}(\phi_{ij}) \neq G(\phi_{ik})$) entre as representações e, por sua vez, quando $ISC = 1$ haverá a total similaridade entre as representações ($\mathrm{G}(\phi_{ij}) = G(\phi_{ik})$).

Comparação dos métodos de avaliação de geometrias

Ao se proceder uma análise nos métodos apresentados, sobretudo comparando-os com viabilidade de se realizar uma análise da similaridade, tem-se que apenas o ISC é capaz de avaliar a similaridade entre representações poligonais. Contudo, as *geometrias* disponíveis nos *datasets* não são exclusivamente poligonais. Assim, para a utilização do ISC como uma função de similaridade é preciso fazer adaptações para as demais *geometrias*. Desta forma, propõe-se realizar uma operação de dilatação nas *geometrias* dos tipos ponto, multiponto, linha e multilinha para que as mesmas tornem-se polígonos e o ISC possa ser utilizado.

Considerando que todas as representações possam ser consideradas como poligonais, a função de similaridade $\mathcal{S}_g$ assume o valor do ISC. Logo, tem-se

(Equação 4.9):

$$\mathcal{S}_g(G(\phi_{ij}), G(\phi_{ik})) = ISC(G(\phi_{ij}), G(\phi_{ik})) \tag{4.9}$$

4.3.3 Processo de dilatação

Ponto e multiponto

Um ponto v qualquer possui coordenadas cujas tolerância com a posição real da feição é estabelecida em legislação. A legislação mais recente e que estabelece um valor absoluto para a imprecisão do ponto é a Portaria do Instituto Nacional de Colonização e Reforma Agrária (INCRA) [44] que define um afastamento máximo de $0,50\ m$ para a precisão do ponto obtido por rastreamento de satélites. Note-se que, em virtude deste valor estabelecido, muitos *datasets* sobre temas do território nacional passaram a não atender a este requisito legal.

Outra forma de se avaliar a imprecisão dos pontos reside na quantificação do erro esperado. Assim, considera-se que o ponto no terreno encontra-se afastado, no máximo, $0.5\ mm$ na escala do documento. Nesta tese, será considerada uma adaptação a esta proposta para se estabelecer o valor ϵ como região de mesmo *locus* geográfico.

Assim, é necessário arbitrar um valor ϵ como referencial. Será, na verdade, o raio do círculo cujo centro será o próprio ponto v. Estabelecer o valor de ϵ não é trivial [45]. Neste caso, como os *datasets* possuem um *locus* geográfico variável, é razoável admitir um valor ϵ variável.

Para tal, nesta tese, o valor ϵ será obtido a partir da análise das caixas envolventes (*bounding box*) dos diversos *datasets*. Assim, calcula-se o comprimento das diagonais das caixas envolventes e identifica-se a relação R entre a de maior – ℓ_{max} – e a de menor comprimento – ℓ_{min}. De posse de R, tem-se que $\epsilon = \frac{0.0005\ m}{Escala} \cdot R$, onde $R = \frac{\ell_{max}}{\ell_{min}}$. Evidentemente, haverá casos em que o *dataset* possuirá apenas um ponto. Neste caso, o comprimento da diagonal será $\ell = 0$ e, por definição, R assumirá o valor 1, ou seja, $R = 1$.

Diante do exposto, é possível inferir que o *locus* geográfico do ponto deva estar em uma *buffer zone* cujo afastamento seja o de ϵ. Para tal, basta considerar um *locus* geográfico do ponto em função da sua região de influência. Evidentemente, esta região será um círculo de raio ϵ, para que o ponto possa ser convertido em um

polígono.

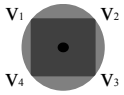

Figura 4.3: Região de influência do ponto

Para facilitar o processamento pode ser considerado como região de influência um quadrado inscrito no círculo de raio ϵ (Figura 4.3). Neste caso, dado um ponto qualquer $v(x_v, y_v)$, tem-se o polígono P representativo da região de influência definido por um quadrado cujos vértices são: $v_1 = (v_x - \epsilon \cdot \frac{\sqrt{2}}{2}, v_y + \epsilon \cdot \frac{\sqrt{2}}{2})$, $v_2 = ((v_x + \epsilon \cdot \frac{\sqrt{2}}{2}, v_y + \epsilon \cdot \frac{\sqrt{2}}{2})$, $v_3 = (v_x + \epsilon \cdot \frac{\sqrt{2}}{2}, v_y - \epsilon \cdot \frac{\sqrt{2}}{2})$ e $v_4 = (v_x - \epsilon \cdot \frac{\sqrt{2}}{2}, v_y - \epsilon \cdot \frac{\sqrt{2}}{2})$.

Desta forma, pode-se usar como função de similaridade $\mathcal{S}_g$ para um ponto as mesmas discutidas no item 4.3.2.

Linha e multilinha

Para se proceder a análise da geometria quando considerada uma linha ou uma multilinha, faz-se necessário identificar a região de influência da mesma. Neste caso, dada uma representação ϕ_{ij} qualquer, o *locus* geográfico é estimado pelo processamento de uma região obtida de forma análoga à que foi sugerida para pontos. Assim sendo, dada uma linha qualquer, esta dará origem a uma região poligonal. Para tanto, considera-se que cada segmento de reta da representação linear possui uma região de influência obtida através do fecho convexo das dilatações dos seus pontos extremos (Figura 4.4). A região poligonal pode então ser computada através da união de todas estes polígonos convexos (Figura 4.5).

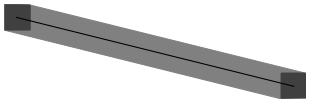

Figura 4.4: Região de influência de um segmento

Vê-se, portanto, que é possível utilizar a função de similaridade $\mathcal{S}_g$ para uma linha da mesma forma como discutida no item 4.3.2.

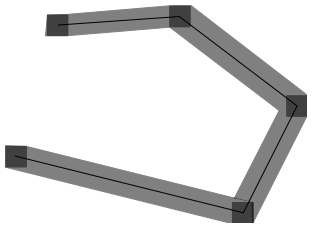

Figura 4.5: Região de influência de uma linha

4.4 Parâmetros de comparação

Há, em alguns casos, a necessidade de se viabilizar a integração das respostas a uma consulta e não dos dados, especialmente quando se está interessado numa análise espacial. Para tanto, é possível o uso do Índice de Completude ($\mathfrak{C}$I) [6], [7] e do Índice de Cobertura ($\mathfrak{Co}$I) [5]. O $\mathfrak{C}I$ permite identificar o quanto uma determinada representação encontra-se dentro de uma região onde não há dúvidas de pertinência. Por sua vez, o $\mathfrak{Co}I$ apresenta um indicativo de quanto uma representação específica participa na região máxima de influência possível em virtude das ambiguidades. Estes dois índices são descritos pelas Equações 4.10 e 4.11.

$$\mathfrak{C}I(\phi_{ij}, \phi_{ik}) = \frac{AREA(\phi_{ij} \cap \phi_{ik})}{AREA(\phi_{ij})} \tag{4.10}$$

$$\mathfrak{Co}I(\phi_{ij}, \phi_{ik}) = \frac{AREA(\phi_{ij})}{AREA(\phi_{ij} \cup \phi_{ik})} \tag{4.11}$$

Ao se observar os índices constata-se de que a representação ϕ_{ij} deve ser, necessariamente, uma representação poligonal. Isto porque há, em todas as possibilidades, no numerador e no denominador a necessidade de se quantificar a área de uma determinada representação. Logo, a dilatação de pontos e de linhas em polígonos viabiliza a aplicação dos índices acima descritos.

4.5 Considerações finais

O estabelecimento das funções $\mathcal{S}_n$ e $\mathcal{S}_g$ permite a avaliação da similaridade entre representações. Como foi discutido no capítulo 3 é preciso estabelecer valores limiares $\mathcal{L}_n$ e $\mathcal{L}_g$ como patamares a partir dos quais um par de representações pode ser considerado como representativo de uma mesma feição do mundo real. No capítulo 7

relata-se experimentos que sugerem um limiar com o valor de 0.700 para ambas as funções.

Observa-se, também, que as funções $\mathcal{S}_n$ e $\mathcal{S}_g$ permitem que se abalize as respostas obtidas ao se realizar uma determinada consulta sobre os múltiplos *data-sets* que representam um determinado tema. Como será mostrado no capítulo 5 a tabela de feições TF serve como um sumário de similaridades entre as diversas representações, permitindo um processamento de consultas conservador que fornece respostas levando em conta todas as possibilidades.

Capítulo 5

Processamento de consulta em BDG ambíguos

5.1 Considerações iniciais

Neste capítulo são apresentadas metodologias para processamento de consultas de seleção (σ) e de junção ($\bowtie$) sobre bancos de dados geográficos multirepresentados, isto é, potencialmente contendo ambiguidades. Estas metodologias empregarão a tabela de feições TF obtida conforme descrito no capítulo 3.

A seleção (σ) é uma operação básica em bancos de dados relacionais que consiste em retornar as tuplas de uma relação (*dataset*) que atendem a um predicado dado. Neste caso, a consulta sobre múltiplos *datasets* resultará em respostas que podem ou não concordar entre si. O que se propõe, então, é apontar nas respostas dadas quais supostas feições do mundo real efetivamente atendem ao predicado. A abordagem proposta consiste em, inicialmente, realizar a seleção sobre TF e, a partir dos resultados obtidos, recuperar representações nos diversos *datasets* que também satisfazem ao predicado.

A junção ($\bowtie$) é uma operação que permite o processamento de relacionamentos entre diferentes temas. Para tal, o sistema dispõe de uma tabela de feições para cada tema. Logo, haverá tantas tabelas de feições quantas sejam os temas disponibilizados. A abordagem proposta consiste em realizar junções sobre as TF's correspondentes e, a partir dos resultados, recuperar representações nos *datasets* originais que possam satisfazer o predicado da junção.

5.2 Consulta de seleção

Ao se efetuar uma consulta de seleção (σ) em uma tabela qualquer é necessário definir um predicado p. A partir deste predicado obtém-se uma relação que serve como resposta à consulta. Evidentemente, o predicado p pode ser de diversos tipos. Nesta tese, em particular, os *datasets* possuem 3(três) atributos, a saber, o identificador da tupla, o atributo não espacial (*nome*) e o atributo espacial (*geometria*).

Como foi descrito na seção 3.5, a arquitetura desenvolvida mantém uma tabela de feições TF para cada tema, a qual busca sumarizar todas as representações disponíveis acerca do tema. O processamento de uma consulta de seleção parte de uma análise de TF na qual se busca obter representações que satisfaçam o predicado p dado. Logo, é necessário ser estabelecido um predicado p' a ser aplicado sobre TF que possa servir de filtro para a obtenção do resultado desejado. Em outras palavras, é preciso estabelecer um mapeamento entre o predicado p proposto por um usuário e um predicado p' que lhe seja equivalente para ser utilizado sobre a TF.

De forma geral a consulta a um *dataset* original ($\sigma_p T = R$) produz uma relação R a partir do predicado p. Assim, um predicado p' deve ser usado para selecionar tuplas de TF ($\sigma_{p'} TF = R'$) produzindo uma relação R'.

Os registros selecionados de TF são indicativos de feições do mundo real que podem atender ao predicado p dado. Na verdade, estas tuplas são aproximações conservadoras de representações destas feições nos múltiplos *datasets*. Vê-se, assim, que TF serve como um filtro capaz de eliminar representações que não podem atender p. Dentre as representações ϕ_{ij} dos diversos *datasets* T_i, associados a uma tupla de TF que satisfaz p', algumas podem não satisfazer o predicado p, demandando assim, um processo de filtragem adicional onde cada ϕ_{ij} é testado em relação a p. Este processo é descrito no Algoritmo 3 que retorna os atributos *nome* e *geometria* de todas as representações ϕ_{ij} que satisfazem o predicado, agrupadas pelo identificador de feição id_F associado à tabela de feições.

Convém, neste ponto, recordar que uma tupla de TF representa, de forma aproximada e conservadora, uma feição do mundo real que se supõe existir a partir de tuplas mutuamente similares em alguns dos *datasets* sobre o tema. Portanto, é razoável que a resposta à consulta de seleção seja avaliada com respeito a essas supostas feições. Desta forma, o resultado do Algoritmo 3 consiste em enumerar, para

cada suposta feição, todas as tuplas que satisfazem o predicado p. Portanto, a relação *Rel* produzida pelo Algoritmo 3, consiste em tuplas da forma $[id_F, nome, geometria]$, onde id_F é um identificador de feição conforme representado em TF e os atributos *nome* e *geometria* são provenientes dos múltiplos *datasets*.

entrada: p, p', $TF(id_F, NOME_M, GEOM_M)$, $T_j(id_T, nome, geometria)$,
$AUX_j(id_F, id_T), j = 1, \ldots, n$
saida : $Rel(id_F, nome, geometria)$

```
begin
    createtable (Rel)
    createtable (TMP)
    TMP ← π_{id_F}(σ_{p'} TF)
    for id_F in TMP do
        for j in 1,...,n do
            h ← σ_{id_F = AUX_j.id_F} AUX_j
            if |h| ≠ 0 then
                t ← σ_{p ∧ h.id_T = T_j.id_T} T_j
                if |t| ≠ 0 then
                    inserir (Rel, [id_F, t_j.nome, t_j.geometria])
    deletetable (TMP)
```

Algoritmo 3: Resposta da consulta de seleção

É importante, neste ponto, observar que os pares de atributos $[nome, geometria]$ poderiam ser igualmente obtidos realizando a consulta de seleção de forma independente sobre os diversos *datasets* representativos do tema. Esta prática, entretanto, não nos dá uma visão unificada das diferentes representações. Ao agrupar os pares selecionados por id_F é possível inferir a existência de feições do mundo real que atendem, em algum grau, o predicado de seleção, mesmo em face de erros de modelagem ou insuficiência de cobertura porventura existentes em um ou mais *datasets*.

5.2.1 Processamento de predicados

O predicado p usado na seleção pode se referir tanto ao atributo não espacial quanto ao espacial. Cumpre, portanto, analisar como os diversos tipos de predicados podem ser transformados em predicados equivalentes p'.

Para um predicado p envolvendo o atributo não espacial, o que se tem é a avaliação de nomes em T_i, o que implica em um predicado equivalente sobre os nomes médios em TF. Os predicados mais comuns sobre nomes consistem em estabelecer a identidade com uma constante (cadeia de caracteres) dada. Tal percepção de identidade, nesta tese, é a aceitação dos nomes que possuam um valor para a função de similaridade não espacial maior do que o limiar proposto ($\mathcal{S}_n \geq \mathcal{L}_n$), conforme visto na seção 4.3. Portanto, seja p dado por $\mathcal{S}_n(T_i.nome, \textit{“nome\_especifico”}) \geq \mathcal{L}_n$. Então, vê-se que o predicado equivalente p' será dado por $\mathcal{S}_n(TF.NOME_M, \textit{“nome\_especifico”}) \geq \mathcal{L}_n$.

Predicados espaciais, por sua vez, podem envolver uma grande variedade de propriedades. Como as representações são poligonais ou dilatadas para criarem um polígono, é conveniente observar que os predicados mais usuais são aqueles relacionados a, por exemplo:

1. Propriedade integrais, tais como área e perímetro;
2. Operações de conjuntos, tais como união, interseção e diferença;
3. Relações topológicas, como “toca”, “cruza” e “dentro”; e,
4. Relações de distância.

Área

Em relação a áreas tem-se que p pode solicitar a comparação com um determinado limiar A. Assim, predicados comuns têm a forma $AREA(geometria) > A$ ou $AREA(geometria) < A$. Ao se analisar TF tem-se que o predicado p' é aplicado sobre o atributo $GEOM_M$, que é uma aproximação conservadora de *geometria*. Em outras palavras:

$$AREA(geometria) > A \Rightarrow AREA(GEOM_M) > A$$

ou

$$AREA(geometria) < A \Leftarrow AREA(GEOM_M) < A.$$

Logo, o predicado equivalente para o caso onde se procura um valor para área superior a um limiar, ou seja, $p \equiv AREA(geometria) > A$ corresponde a $p' \equiv AREA(GEOM_M) > A$. Para o caso contrário, $p \equiv AREA(geometria) < A$, todas as tuplas de TF devem ser relacionadas em R' ($p' \equiv true$).

Perímetro

Predicados envolvendo perímetros podem assumir a forma de $PERIMETRO(geometria) > P$ ou, ainda, a forma de $PERIMETRO(geometria) < P$. Neste caso, o *bounding box* não provê subsídios para distinguir limites para esta propriedade. Assim, para p' faz-se necessário a aceitação de todas as tuplas de TF ($p' \equiv true$) independentemente do predicado p impor um limite superior ou inferior para o valor do perímetro.

Operações de conjunto

A operação de união (UNIAO) entre o atributo *geometria* e alguma geometria G constante dada (Figura 5.1) pode figurar num predicado como argumento de alguma função lógica f, como por exemplo $AREA(UNIAO(geometria, G)) > A$. Pode-se ver, então, que a obtenção de um predicado equivalente p' é dependente da operação UNIAO ser conservadora com relação à propriedade testada. Assim, no exemplo dado pode-se escrever $p \equiv AREA(UNIAO(geometria, G)) > A$ corresponde a $p' \equiv AREA(UNIAO(GEOM_M, G)) > A$. O mesmo não se verifica, por exemplo, em predicados envolvendo perímetro.

Analogamente, a operação de interseção (INTERSECAO) também é conservadora com relação a limites inferiores para área. No caso da operação de diferença (DIFERENCA), sendo esta não comutativa, é necessário analisar separadamente $f(DIFERENCA(geometria, G))$ e $f(DIFERENCA(G, geometria))$. A primeira forma se comporta de modo semelhante às interseções com relação, por exemplo, a predicados relativos a área, enquanto que a segunda forma não é conservadora com relação a estes mesmos predicados.

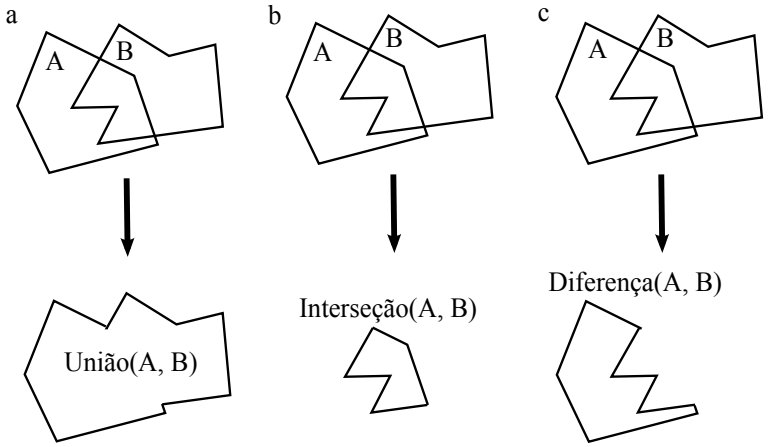

Figura 5.1: Operações entre polígonos

Toca

A função toca (TOCA) avalia se uma representação possui alguma interseção com outra dada. Considerando uma geometria qualquer G, a função é definida, portanto, como:

$$TOCA(geometria, G) \equiv INTERSECAO(geometria, G) \neq \emptyset$$

Neste caso, o predicado $p \equiv TOCA(geometria, G)$ pode ser avaliado, conservadoramente, por $p' \equiv INTERSECAO(GEOM_M, G)$.

Cruza

A função cruza (CRUZA) procura avaliar se uma determinada representação linear atravessa uma outra, em outras palavras, procura identificar se a mesma "entra" e "sai" de outra. Assim, dada uma geometria G, p ($p \equiv CRUZA(geometria, G)$) requer que *geometria* e G tenham interseção não nula. Portanto, um equivalente p' que lhe é conservador ao se analisar o atributo $GEOM_M$ da tabela de feições consiste em $p' \equiv INTERSECAO(GEOM_M, G) \neq \emptyset$.

Dentro

A função dentro (DENTRO) procura inferir se uma geometria G está no interior ou não de uma dada representação. Em outras palavras, procura inferir se uma G está ou não contida na representação. Assim, semelhantemente ao predicado CRUZA, p ($p \equiv DENTRO(geometria, G)$) tem o seu equivalente p' ($p' \equiv INTERSECAO(GEOM_M, G) \neq \emptyset$).

Distância

A função de distância mínima (DISTANCIA) entre dois conjuntos de pontos A e B é definida por:

$$DISTANCIA(A, B) = min_{a \in A, b \in B} d(a, b),$$

onde $d(a, b)$ é a distância euclidiana entre os pontos a e b. Esta distância é obtida a partir da análise dos valores oriundos de cada relação par-a-par das geometrias envolvidas. Por sua vez, é possível definir uma distância máxima ($DISTANCIA_{MAX}$) como:

$$DISTANCIA_{MAX}(A, B) = max_{a \in A, b \in B} d(a, b)$$

Predicados comuns envolvendo funções de distância consiste em avaliar se uma dada geometria G encontra-se mais afastada ou mais próxima que um limiar D. Neste caso, é possível inferir 4 (quatro) possibilidades para p:

- $p \equiv DISTANCIA(geometria, G) > D$;
- $p \equiv DISTANCIA(geometria, G) < D$;
- $p \equiv DISTANCIA_{MAX}(geometria, G) > D$; e,
- $p \equiv DISTANCIA_{MAX}(geometria, G) < D$.

Por sua vez, o predicado equivalente p' a ser aplicado sobre TF toma a seguinte forma, respectivamente:

- $p' \equiv DISTANCIA_{MAX}(GEOM_M, G) > D$;
- $p' \equiv DISTANCIA(GEOM_M, G) < D$;
- $p' \equiv DISTANCIA_{MAX}(GEOM_M, G) > D$; e,
- $p' \equiv DISTANCIA(GEOM_M, G) < D$.

Resumo

A Tabela 5.1 sumariza a relação entre alguns predicados de seleção p e seus predicados equivalentes p'.

Tabela 5.1: Equivalência entre predicados espaciais

p $(\sigma_p T_i)$	p' $(\sigma_{p'} TF)$
$AREA(geometria) > A$	$AREA(GEOM_M) > A$
$AREA(geometria) < A$	$true$
$PERIMETRO(geometria) > P$	true
$PERIMETRO(geometria) < P$	true
$TOCA(geometria, G)$	$INTERSECAO(GEOM_M, G) \neq \emptyset$
$CRUZA(geometria, G)$	$INTERSECAO(GEOM_M, G) \neq \emptyset$
$DENTRO(geometria, G)$	$INTERSECAO(GEOM_M, G) \neq \emptyset$
$DISTANCIA(geometria, G) > D$	$DISTANCIA_{MAX}(GEOM_M, G) > D$
$DISTANCIA(geometria, G) < D$	$DISTANCIA(GEOM_M, G) < D$
$DISTANCIA_{MAX}(geometria, G) > D$	$DISTANCIA_{MAX}(GEOM_M, G) > D$
$DISTANCIA_{MAX}(geometria, G) < D$	$DISTANCIA(GEOM_M, G) < D$

5.3 Consulta de junção

Consultas de junção ($\bowtie$) sobre duas tabelas distintas (T_1 e T_2) requerem que se especifique um predicado p envolvendo relações entre atributos das mesmas. Uma junção retorna todos os pares de tuplas que satisfazem o predicado, isto é, $(T_1 \bowtie_p T_2 = \sigma_p(T_1 \times T_2) = R)$, onde $\times$ denota o produto cartesiano de duas relações.

Ressalta-se que os *datasets* disponíveis estão relacionados com diversos temas. Assim, há tabelas T_{ij} para cada tema. Nesta tese, os diversos temas serão enumerados pela identificação sequencial da tabelas originais da forma $T1_{ij}, \ldots, Tn_{ij}$. Para facilitar a associação das tabelas de feições com os temas nos *datasets*, assume-se que as tabelas de feições possuem os índices associados ao tema da mesma forma, ou seja, $TF_1, \ldots, TF_n$.

Nesta tese, em particular, a junção é realizada sobre as tabelas de feições de cada tema. Logo, um predicado p, a ser usado sobre os *datasets* originais, deve possuir um equivalente p' para que possa ser usado sobre atributos de duas tabelas de feições. Sem perda de generalidade, sejam dois temas distintos τ_1 e τ_2 disponibilizados em múltiplos *datasets*. Para tal, temos que considerar a existência de duas tabelas de

feições TF_1 e TF_2, respectivamente. Assim, temos que a junção em sua forma geral ($TF_1 \bowtie_{p'} TF_2 = R'$) produz uma relação R' com 6(seis) atributos, a saber $TF_1.id_F$, $TF_1.nome$, $TF_1.geometria$, $TF_2.id_F$, $TF_2.nome$ e $TF_2.geometria$. Como existem atributos com o mesmo nome, há a necessidade de renomeá-los (operador ρ da Álgebra Relacional).

A relação R' aponta para pares de representações de feições do mundo real que potencialmente atendem ao predicado p da junção. Para cada um destes pares, representações concretas constantes dos múltiplos *datasets* devem ser testadas com respeito ao predicado p. Para tal, é aplicado o Algoritmo 4 que retorna os atributos *nome* e *geometria* de todas as representações ϕ_{ij} que satisfazem o predicado p agrupados pelos identificadores das feições id_F de cada tema envolvido na junção.

5.3.1 Processamento de predicados

Também no processamento de consulta de junção há a necessidade de se obter predicados conservadores equivalentes para serem usados na consultas às tabelas de feições. Este processo segue as mesmas técnicas descritas na seção 5.2.1, sendo que valores constantes envolvidos nos predicados devem ser substituídos por atributos respectivos do segundo tema. Assim, por exemplo, um predicado para avaliação de distância entre duas feições pertencentes a temas distintos, ou seja, $p \equiv DISTANCIA(geometria_1, geometria_2) > D$ é processado nas tabelas de feições pelo predicado equivalente $p' \equiv DISTANCIA_{MAX}(GEOM_{M_1}, GEOM_{M_2}) > D$, por adaptação dos predicados de seleção vistos na Tabela 5.1.

5.4 Considerações finais

Diante do apresentado neste capítulo, é possível verificar que as consultas realizadas nas tabelas de feições apontam para feições multirepresentadas do mundo real que atendem as restrições dadas. Vemos, portanto, que tais tabelas servem a um duplo propósito. Em primeiro lugar permitem que se realize uma filtragem grosseira das múltiplas representações, eliminando aquelas que não podem satisfazer as restrições da consulta. Adicionalmente, os índices de feições (id_F) presentes nos resultados

entrada: p, p', $TF_1(id_F, NOME_M, GEOM_M)$,
$TF_2(id_F, NOME_M, GEOM_M)$, $T1_i(id_T, nome, geometria)$,
$AUX1_i(id_F, id_T), i = 1, \ \ldots, \ n$, $T2_j(id_T, nome, geometria)$,
$AUX2_j(id_F, id_T), j = 1, \ \ldots, \ m$

saída : $Rel(id_{F_1}, id_{F_2}, nome_1, nome_2, geometria_1, geometria_2)$

begin

- `createtable` (Rel)
- `createtable` (TMP)
- `createtable` (H_1)
- `createtable` (H_2)
- $TMP \leftarrow \pi_{id_{F_1}, id_{F_2}}(\rho_{id_{F_1}/id_F}(TF_1) \ \bowtie_{p'} \ (\rho_{id_{F_2}/id_F}(TF_2))$
- **for** $[id_{F_1}, id_{F_2}\ ]$ **in** TMP **do**
  - **for** i **in** *1,...,n* **do**
    - $h \leftarrow \rho_{nome_1/nome, geometria_1/geometria}\sigma_{id_{F_1}=AUX1_i.id_F}AUX1_i$
    - `inserir` $(H_1,[id_{F_1}, h.nome_1, h.geometria_1])$
  - **for** j **in** *1,...,m* **do**
    - $h \leftarrow \rho_{nome_2/nome, geometria_2/geometria}\sigma_{id_{F_2}=AUX2_j.id_F}AUX2_j$
    - `inserir` $(H_2,[id_{F_2}, h.nome_2, h.geometria_2])$
- $Rel \leftarrow H_1 \ \bowtie_p H_2$
- `deletetable` (TMP)
- `deletetable` (H_1)
- `deletetable` (H_2)

Algoritmo 4: Resposta da consulta de junção

fornecidos pelos Algoritmos 3 e 4 permitem manter o mapeamento entre feições supostamente existentes no mundo real e suas diversas representações, auxiliando, assim, a análise dos resultados das consultas.

Capítulo 6

Sistema Avaliador de Respostas Ambíguas – SARA

6.1 Considerações iniciais

Sistemas de Banco de Dados Geográficos tipicamente pressupõem que cada tema que se deseja representar é modelado por um único *dataset* correspondente. Entretanto, um mesmo tema pode ter sido mapeado por diversos produtores, gerando um conjunto de *datasets*. Esta multiplicidade é benéfica sob certos aspectos, tais como a democratização das informações, a possibilidade de obter diversas versões do mesmo dado, análise temporal, entre outros. Contudo, a multiplicidade pode ser considerada um contratempo à prática produtiva e aos anseios de certos usuários que desejam uma base unificada.

O objetivo da presente tese não é eliminar o problema da multiplicidade, sequer disponibilizar uma única representação da feição por meio de certificação de dados. Esta conduta é comum apenas porque as arquiteturas de BDG correntes não permitem consultar às informações multirepresentadas, possivelmente ambíguas. Tais ambiguidades nas representações podem ser de natureza geométrica, semântica ou topológica.

A ambiguidade geométrica consiste em haver mais de uma figura geométrica representativa da mesma entidade integrante do mundo real. A ambiguidade semântica tem a ver com a multiplicidade de significado – *nomes* – para a mesma feição. Finalmente, a ambiguidade topológica ocorre quando há múltiplas repre-

sentações e a análise espacial entre as diversas possibilidades fornecem relacionamentos topológicos distintos.

Visando dar um tratamento à multipla representação dos dados, esta tese propõe uma arquitetura para processamento de consultas denominada Sistema Avaliador de Respostas Ambíguas (SARA), capaz de tratar ambiguidades e fornecer uma resposta sumarizada a partir dos *datasets* disponíveis.

6.2 Arquitetura SARA

A arquitetura SARA possibilita a consulta a bases cartográficas, ambíguas ou não, e permite a geração de resultados classificados que podem ser recuperados (Figura 6.1). A arquitetura proposta propicia ao usuário obter resultados independentemente das ambiguidades. Assim sendo, pressupõe a existência de uma infraestrutura que realize as consultas geográficas específicas em temas, fornecendo ao usuário um conjunto de respostas, ambíguas ou não, para análise.

As respostas possíveis estão relacionadas com ambiguidades existentes entre os diversos bancos de dados e que fornecem resultados dúbios. Por exemplo, pode-se imaginar a existência de uma consulta espacial solicitando a contagem de um objeto qualquer que esteja afastado a uma determinada distância plana de outra feição. Haverá, neste caso, tantas respostas quanto as combinações possíveis – duas a duas – entre as representações. De modo genérico, tem-se uma única resposta a cada consulta, efetuada sobre cada conjunto de dados individual (Figura 6.2).

O procedimento clássico para a obtenção de uma resposta única propõe a integração dos dados disponíveis para que a consulta seja feita a um único conjunto de dados (Figura 6.3). Desta forma, a ambiguidade estaria eliminada. Entretanto, há que se considerar que o dado gerado e disponibilizado para a consulta é um dado derivado e não original.

A arquitetura proposta permite a obtenção de resultados ambíguos. Para tal, um usuário qualquer realiza o que denominamos de meta-consulta. A **meta-consulta** é uma ação do usuário que tem por objetivo obter conhecimento sobre a realidade, conforme esteja espelhada pelos dados armazenados.

Uma vez idealizada uma meta-consulta, faz-se necessária uma transformação

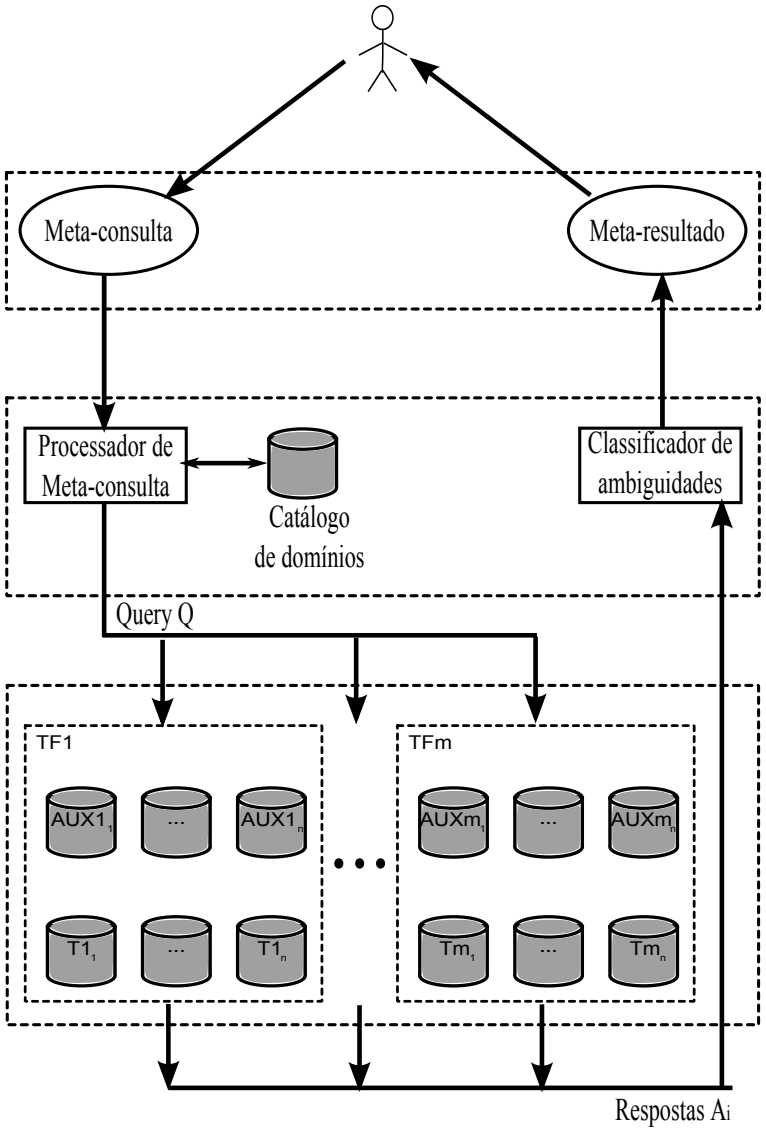

Figura 6.1: Arquitetura proposta

desta para o ambiente digital, particularmente, para a arquitetura. Em outras palavras, a meta-consulta deve ser transformada em algoritmos que permitam o processamento desta nos diversos *datasets* mapeados da arquitetura. Desta forma, a **meta-consulta** é obtida por meio de uma sintaxe SQL adequadamente estendida, onde, ao invés de relações, o usuário emprega o nome de temas ou domínios genéricos. Evidentemente, para se processar tais meta-consultas, o algoritmo deve se apoiar nos metadados.

Na **meta-consulta** os valores utilizados para os limiares $\mathcal{L}_n$ e $\mathcal{L}_g$ são aqueles estabelecidos quando da construção das tabelas de feições. Isto não impede que o usuário utilize valores distintos para a realização de consultas. Tais limiares são

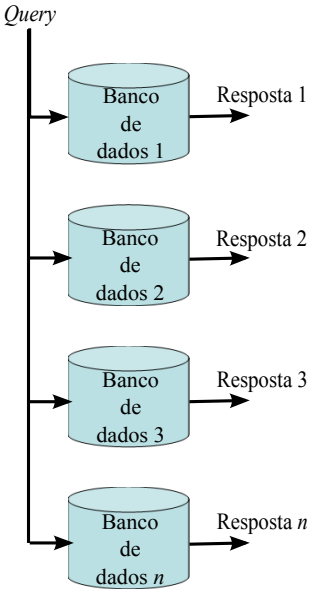

Figura 6.2: Consulta unívoca

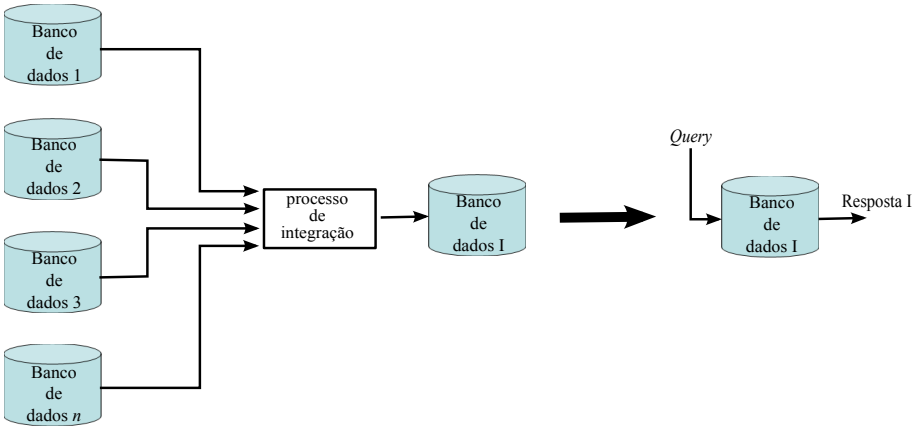

Figura 6.3: Integração de dados

empregados apenas para a avaliação de predicados da consulta, não tendo nenhuma influência no estabelecimento da identidade entre representações. Assim, por exemplo, no Capítulo 7, onde se relata uma série de experimentos realizados com o SARA, em função dos *datasets* disponíveis, foi adotado o valor de 0.700 para ambos limiares.

Os metadados são dados que procuram registrar informações das representações das feições, contextualizando-as. Na realidade, a partir dos metadados das feições é possível identificar as representações e facilitar a consulta ao banco. Os metadados encontram-se registrados no **Catálogo de Domínio**. Este Catálogo, por sua vez, é o local onde ficam registradas as informações referentes aos diversos *datasets* que fornecem os dados para que o $SARA$ proceda a análise. Contém, ainda, os esquemas

conceituais referentes às tabelas necessárias a avaliação de similaridade, no caso, as TF's, as tabelas auxiliares (AUX_i) e os *datasets* originais (T_i).

O **Processador de meta-consultas** gerencia, por meio de algoritmos, o acesso à infra-estrutura e elabora as consultas, distribuindo-as nos diversos *datasets*. Em outras palavras, o processador prepara uma consulta para cada possibilidade de consulta em função das múltiplas representações no banco de dados (T_i's). As informações necessárias para a construção da consulta constam no Catálogo de Domínio.

A arquitetura é responsável pelo acesso aos *datasets* distintos. Cada *dataset* (T_i) é, na realidade, um conjunto de representações das feições do mundo real acerca de um tema. Assim, após a montagem das consultas e o acesso aos bancos individuais, os resultados são apreendidos pelo SARA, sendo uma para cada tema consultado. Tais resultados (Rel, conforme visto na seção 5.2 e 5.3) são os insumos para o Classificador Analítico de Ambiguidades (CAA).

O **Classificador Analítico de Ambiguidades** é um conjunto de algoritmos que permite a classificação dos resultados (Figura 6.4). A classificação é uma forma de se obter respostas consolidadas a partir dos dados potencialmente ambíguos, o que chamamos de **meta-resultado**. Durante este processo é possível que se obtenha uma concordância plena entre as representações, isto é, ausência de ambiguidades. Em outros casos, esta concordância poderá ser parcial, o que sugere problemas de cobertura e/ou modelagem. Assim, o CAA é o responsável por balizar o grau de concordância das representações de forma a possibilitar ao usuário uma tomada de decisão bem informada.

O CAA, de posse da relação R' possui, portanto, uma relação com todas as representações que atendem o predicado da consulta agrupadas pelo identificador de feição (id_F). Assim, o CAA é capaz de qualificar a resposta, bem como, inferir sobre a qualidade dos *datasets* disponibilizados. Evidentemente, quando se possui n *datasets* sobre um mesmo tema, é razoável inferir de que cada *dataset* individualmente possua uma única representação de uma feição qualquer do terreno. Logo, espera-se que haja n representações agrupadas por um mesmo id_F. A não existência destas n representações caracteriza a falha de cobertura ou problemas na modelagem. Assim, o CAA infere a qualidade da cobertura computando o que denomina-se neste

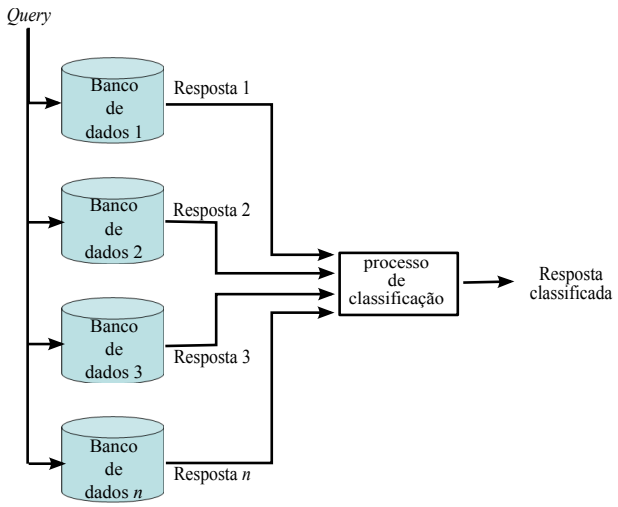

Figura 6.4: Classificação dos resultados

trabalho de *índice de relevância*, que corresponde a uma relação entre a quantidade de tuplas associadas com uma dada feição ou par de feições (F) em Rel e a quantidade máxima de relações – n_{rel} – possíveis entre os *datasets* (Equação 6.1). Para tal, utiliza-se a função CONTAR que serve para quantificar quantas tuplas em Rel possuem, no atributo id_F, um valor dado F.

$$\mathfrak{Q}(F) = \frac{CONTAR(Rel, F)}{n_{rel}} \tag{6.1}$$

No caso, da consulta de seleção, n_{rel} é igual a quantidade de *datasets* disponíveis sobre o tema e F é um valor para o atributo id_F. Assim, o índice de relevância é obtido ao se confrontar o quantitativo de feições identificadas pela quantidade de *datasets* disponibilizados.

Por sua vez, nas consultas de junção, n_{rel} é igual ao produto $n \cdot m$, onde n é o número de *datasets* de um tema e m é a quatidade de *datasets* do outro tema. Aqui, F assume o valor de um par (id_{F_1}, id_{F_2}). Logo, o índice de relevância é calculado a partir da análise do agrupamento dos pares de feições e não de uma feição individual.

O **meta-resultado** consiste numa apresentação legível por humanos dos resultados contidos na tabela Rel, bem como sua qualificação, por exemplo, usando os índices de relevância descritos anteriormente. Neste caso, a tabela pode ser simplesmente apresentada ou podem ser operacionalizadas funções sobre rel de forma a fornecer uma informação mais detalhada. É possível, ainda, apresentar graficamente a resposta em virtude da existência das geometrias em Rel.

De forma a simplificar o entendimento da arquitetura, é possível observar o fluxo de dados a partir do diagrama de atividades correspondente ao SARA (Figura 6.5).

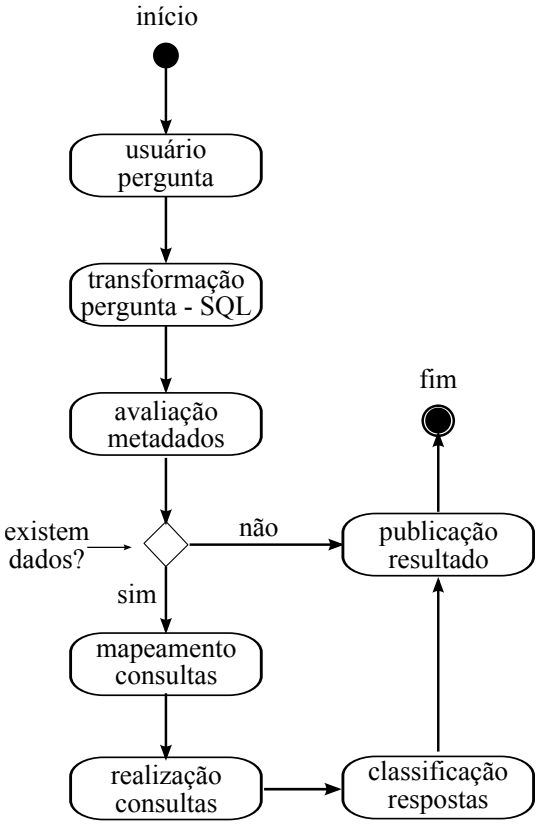

Figura 6.5: Diagrama de atividades

6.3 Exemplo de procedimento

Seja o exemplo motivador encontrado na seção 1.4. Logo, temos as tabelas originais $T1_1$ e $T1_2$ (Tabelas 6.1 e 6.2, respectivamente) para o tema τ_1 e as tabelas $T2_1$ e $T2_2$ (Tabelas 6.3 e 6.4) para o tema τ_2.

Tabela 6.1: Primeiro *dataset* do tema τ_1 ($T1_1$)

id_T	nome	geometria
1	P_1	[[66,54],[84,48],[83,34],[78,21],[65,20],[62,28],[66,39],[56,47]]

O SARA gera, portanto, as tabelas de feições TF_1 e TF_2 (Tabelas 6.5 e 6.6), respectivamente) e as tabelas auxiliares $AUX1_1$, $AUX1_2$, $AUX2_1$ e $AUX2_2$ (Tabelas 6.7, 6.8 e 6.9) a partir das tabelas originais T_i.

Tabela 6.2: Segundo *dataset* do tema τ_1 ($T1_2$)

id_T	nome	geometria
1	P_1	[[66,54],[75,48],[84,48],[84,32],. . . ,[62,28],[63,37],[56,47],[60,52]]

Tabela 6.3: Primeiro *dataset* do tema τ_2 ($T2_1$)

id_T	nome	geometria
1	v_1	[64,25]
2	v_2	[80,34]
3	v_3	[76,44]

O procedimento da arquitetura para a avaliação dos *nomes* segue o procedimento abaixo. Para tal, considerar-se-á os v_i como os nomes da feição:

1. Meta-consulta
   - Quais *pontos* do tema τ_2 estão contidos no retângulo [[0,0],[70,50]]?
2. Catálogo de domínios – mapeia o tema τ_2 nos diversos *datasets*
   - Fontes de dados sobre o tema τ_2: $T2_1$ e $T2_2$; e,
   - Tabela de feições sobre o tema τ_2: TF_2;
3. Processador de meta-consulta – realiza as consultas
   - Invoca o Algoritmo 3 com as entradas TF_2, $T2_1$, $T2_2$, $AUX2_1$, $AUX2_2$, $p \equiv INTERSECAO(geometria, [[0,0],[70,50]]) \neq \emptyset$, $p' \equiv INTERSECAO(GEOM_M, [[0,0],[70,50]]) \neq \emptyset$, obtendo Rel; e,
   - $R' = \sigma_{CONTAR(id_F)} Rel$
4. Respostas
   - Rel dado pela Tabela 6.10; e,
   - R' dado pela Tabela 6.11.
5. Classificador Analítico de Ambiguidades (CAA)
   - $\mathfrak{Q}(1) \equiv \mathfrak{Q}(v_1) = \frac{2}{2} = 1.00$;

Tabela 6.4: Segundo *dataset* do tema τ_2 ($T2_2$)

id_T	nome	geometria
1	v_1	[64,25]
2	v_2	[80,34]
3	v_4	[64,38]
4	v_5	[68,42]

Tabela 6.5: Tema τ_1 (TF_1)

id_F	$NOME_M$	$GEOM_M$
1	P_1	[[56,20],[84,54]]

- $\mathfrak{Q}(4) \equiv \mathfrak{Q}(v_4) = \frac{1}{2} = 0.50$; e,
- $\mathfrak{Q}(5) \equiv \mathfrak{Q}(v_5) = \frac{1}{2} = 0.50$.

6. Meta-resultado

   - O *ponto* v_1 pertence aos diferentes *datasets* (100%). Entretanto, os *pontos* v_4 e v_5 possuem certo grau de incerteza (50%).

6.3.1 Procedimento de consulta de junção

Considerando os mesmos dados contidos na seção 6.3, temos:

1. Meta-consulta

   - Quantos pontos v_i (tema τ_2) encontram-se no interior do polígono P (tema τ_1)

2. Catálogo de domínios – mapeia a meta-consulta

   - Fontes de dados sobre o tema τ_1: $T1_1$ e $T1_2$;
   - Tabela de feições sobre o tema τ_1: TF_1
   - Fontes de dados sobre o tema τ_2: $T2_1$ e $T2_2$; e,
   - Tabela de feições sobre o tema τ_2: TF_2;

3. Processador de meta-consulta – realiza as consultas

Tabela 6.6: Tema τ_2 (TF_2)

id_F	$NOME_M$	$GEOM_M$
1	v_1	[[64,25],[64,25]]
2	v_2	[[80,34],[80,34]]
3	v_3	[[76,44],[76,44]]
4	v_4	[[64,38],[64,38]]
5	v_5	[[68,42],[68,42]]

Tabela 6.7: Auxiliar $AUX1_1 \equiv AUX1_2$

id_F	id_T
1	1

- Invoca o Algoritmo 4 com as entradas TF_1, TF_2, $T1_1$, $T1_2$, $T2_1$, $T2_2$, $AUX1_1$, $AUX1_2$, $AUX2_1$, $AUX2_2$, $p \equiv DENTRO(T1_i.geometria, T2_j.geometria) \neq \emptyset$, $p' \equiv INTERSECAO(TF_1.GEOM_M, TF_2.GEOM_M) \neq \emptyset$, obtendo Rel; e,
- $R' = \sigma_{CONTAR([id_{F_1}, id_{F_2}])} Rel$.

4. Respostas

   - Rel, conforme Tabela 6.12; e,
   - R', conforme Tabela 6.13.

5. Classificador Analítico de Ambiguidades (CAA)

   - $\mathfrak{Q}([2,1]) \equiv \mathfrak{Q}([v_1, P_1]) = \frac{4}{4} = 1.00$;
   - $\mathfrak{Q}([3,1]) \equiv \mathfrak{Q}([v_3, P_1]) = \frac{2}{4} = 0.50$;
   - $\mathfrak{Q}([4,1]) \equiv \mathfrak{Q}([v_4, P_1]) = \frac{1}{4} = 0.25$; e,
   - $\mathfrak{Q}([5,1]) \equiv \mathfrak{Q}([v_5, P_1]) = \frac{2}{4} = 0.50$.

6. Meta-resultado

   - Não há dúvidas de que o ponto v_2 encontra-se dentro do polígono P (100%). Entretanto há dúvidas para os pontos (v_3 e v_5, com 50% de relevância) e o ponto v_4 com 25% de relevância.

Tabela 6.8: Auxiliar $AUX2_1$

id_F	id_T
1	1
2	2
3	3

Tabela 6.9: Auxiliar $AUX2_2$

id_F	id_T
1	1
2	2
4	3
5	4

6.4 Considerações finais

A solução das ambiguidades nem sempre é o desejado pelos cartógrafos. Entretanto, deve-se ter em mente que o dado geográfico consistente tem que ser preservado, independentemente de seu uso. Na realidade, a tese visa a oferecer aos usuários a possibilidade de terem acessos aos dados consistentes, porém ambíguos. Daí a necessidade de se produzir consultas e obter resultados que favoreçam uma análise mais aprofundada.

Os dados ambíguos podem ser irrelevantes para as consultas espaciais, mas podem produzir informações úteis que, atualmente, são desprezadas. Evidentemente, a utilização destes dados podem conduzir a um usuário qualquer responsável por tomar uma decisão a fazê-la da melhor forma.

Tabela 6.10: Resposta *Rel* para a consulta de seleção

id_F	$nome$	$geometria$
1	v_1	[64,25]
1	v_1	[64,25]
4	v_4	[64,38]
5	v_5	[68,42]

Tabela 6.11: Resposta R' para a consulta de seleção

id_F	CONTAR(id_F)
1	2
4	1
5	1

Tabela 6.12: Resposta *Rel* para a consulta de junção

id_{F_1}	id_{F_2}	$nome_1$	$nome_2$	$geometria_1$	$geometria_2$
2	1	v_2	P_1	$T2_1.geometria$	$T1_1.geometria$
2	1	v_2	P_1	$T2_1.geometria$	$T1_2.geometria$
2	1	v_2	P_1	$T2_2.geometria$	$T1_1.geometria$
2	1	v_2	P_1	$T2_2.geometria$	$T1_2.geometria$
3	1	v_3	P_1	$T2_1.geometria$	$T1_1.geometria$
3	1	v_3	P_1	$T2_1.geometria$	$T1_2.geometria$
4	1	v_4	P_1	$T2_2.geometria$	$T1_2.geometria$
5	1	v_5	P_1	$T2_2.geometria$	$T1_1.geometria$
5	1	v_5	P_1	$T2_2.geometria$	$T1_2.geometria$

Tabela 6.13: Resposta R' para a consulta de junção

$[id_{F_1}, id_{F_2}]$	CONTAR($[id_{F_1}, id_{F_2}]$)
$[2, 1]$	4
$[3, 1]$	2
$[4, 1]$	1
$[5, 1]$	2

Capítulo 7

Experimentos

7.1 Considerações iniciais

Para validar o desenvolvimento teórico da tese é necessário gerar um protótipo com a finalidade de permitir a um usuário qualquer a inspeção visual do resultado obtido por meio das consultas SQL e pela aplicação dos índices desenvolvidos.

Assim, nesta tese o protótipo do SARA foi implementado na linguagem Python com o uso do SGBD PostgresSQL e da biblioteca OpenGL para permitir a visualização dos resultados. Neste caso, o aplicativo exibe graficamente as diversas fontes de dados como se encontram e, função do desejado, fornece uma representação na cor verde para o *locus* geográfico onde não há dúvidas quanto a resposta e em amarelo a região do espaço onde as respostas são ambíguas. Desta forma, é possível realizar uma inspeção visual das ambiguidades e obter, via tupla na tabela resposta, o valor númerico dos índices de similaridade não espacial ($\mathcal{S}_g$) e espacial ($\mathcal{S}_g$).

7.2 Dados experimentais

As funcionalidades do protótipo foram testadas por meio de informações obtidas junto a alguns dos órgão produtores de dados existentes no município do Rio de Janeiro. Assim, foram obtidas junto do Instituto Brasileiro de Geografia e Estatística (IBGE) e do Intituto Pereira Passos (IPP) a base cartográfica contendo a malha dos bairros da Cidade do Rio de Janeiro. Com estas duas bases foi avaliada a possibilidade de se quantificar a similaridade quando a geometria se referia a polígonos. No

caso, foram tratados todos os 159 bairros que compõem a Cidade.

Ao se visualizar os *datasets* por tipo de geometria observa-se que há, em todos os casos, uma ambiguidade, no caso dos polígonos (Figura 7.1), no caso da linhas (Figura 7.2) e no caso dos pontos (Figura 7.3). Assim, os dados obtidos para teste são favoráveis para que o protótipo possa ser empregado. Neste caso, o polígono representa o *locus* geográfico do bairro, a linha representa os limites e o ponto representa o centróide dos bairros.

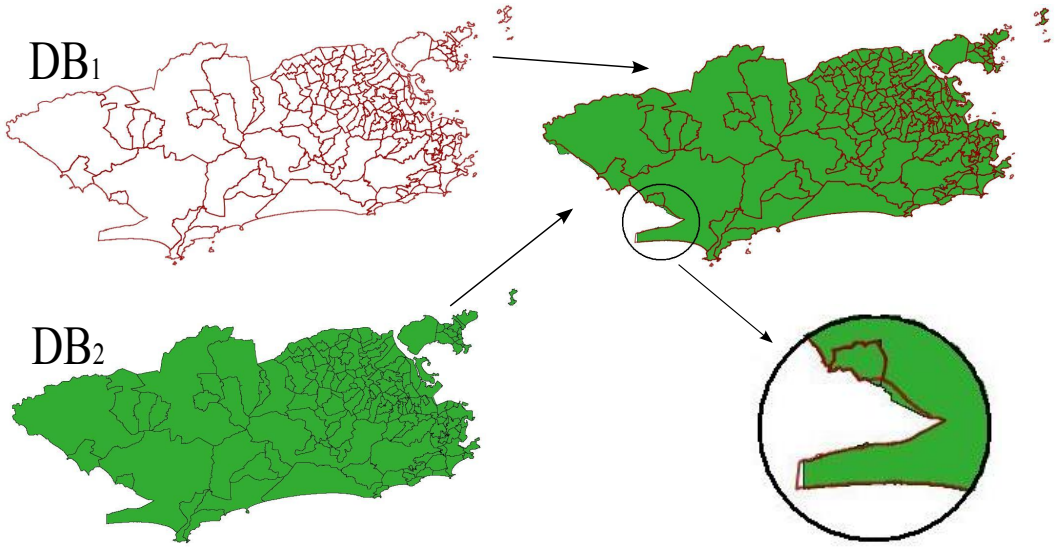

Figura 7.1: Ambiguidade de polígonos – bairros

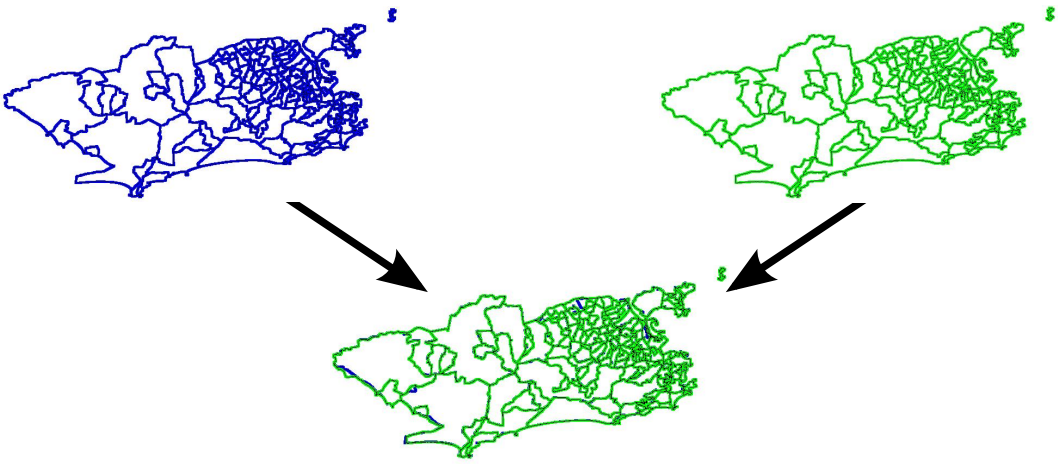

Figura 7.2: Ambiguidade de linhas poligonais – limites dos bairros

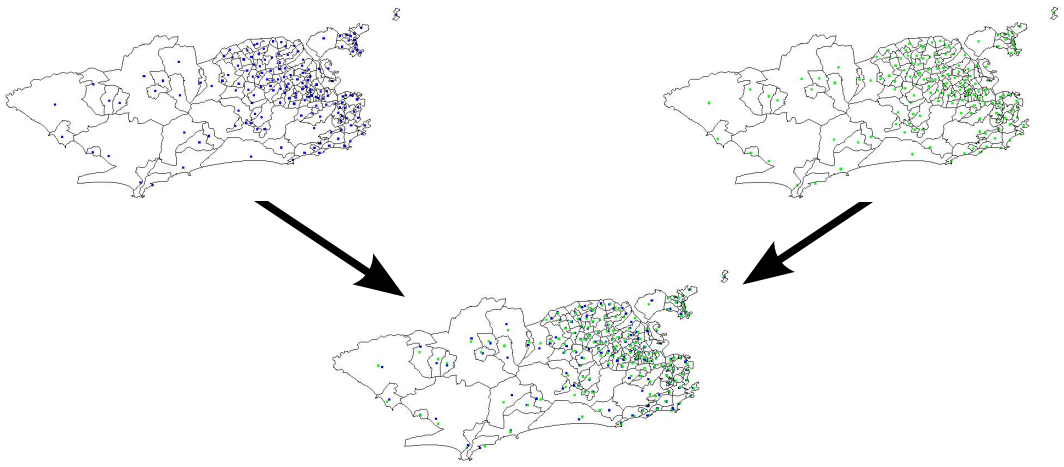

Figura 7.3: Ambiguidade de pontos – centróides dos bairros

7.3 Testes realizados

Numa primeira análise nos *datasets*, constatou-se de que ambos possuíam 159 registros de *nomes* e de *geometrias*. Assim, percebe-se que não há falta de recobrimento nas informações. Desta forma, *a priori* qualquer um dos *dataset* deveria possuir os mesmos dados.

Visando facilitar a compreensão dos testes, estes foram subdivididos em duas partes. A primeira tratou dos *nomes* dos 159 bairros que compõem a Cidade do Rio de Janeiro, enquanto a segunda tratou das *geometrias*. Os testes com a *geometria*, por sua vez, foram realizadas a partir da combinação destas, ou seja, polígono *versus* polígono, linha *versus* linha e ponto *versus* ponto.

7.3.1 Teste do atributo *nome*

Para se avaliar os *nomes* dos bairros nos *datasets*, foi testado cada um destes com os demais do outro *dataset*. Assim, pode ser identificado o par de *nomes* que possuíam o maior Coeficiente de Dice. Dentre os 159 nomes existentes nos dois *datasets*, apenas 9 não identificaram um correspondente com Coeficente de Dice igual a 1. Neste caso, 150 registros de *nomes* em um *dataset* encontram perfeitamente no outro *dataset* com o *nome* equivalente ($d_d = 1$). Os demais casos, apresentados na Tabela 7.1 permitem observar que, embora sejam distintos, há uma correspondência entre eles. Neste caso, o usuário poderia, inclusive, identificar as possibilidades e atribuí-las uma identidade.

Tabela 7.1: Bairros com Coeficiente de Dice diferentes de 1.0

nome no $dataset_1$	*nome* no $dataset_2$	d_d
Alto da Boavista	Alto da Boa Vista	0.9333
Freguesia (Jacarepaguá)	Freguesia Jacarepaguá	0.8947
Turiaçu	Turiaçú	0.8333
Oswaldo Cruz	Osvaldo Cruz	0.8182
Vila Cosmos	Vila Kosmos	0.7778
Quintino Bocaiúva	Quintino	0.5714
Complexo do Alemão	Alemão	0.5000
Freguesia (Ilha do Governador)	Freguesia	0.4571
Complexo da Maré	Maré	0.3333

Os limiares $\mathcal{L}_n$ e $\mathcal{L}_g$ são, nesta tese, arbitrados pelo usuário. Assim, para estabelecer um parâmetro no $SARA$, optou-se por considerar a identidade dos atributos relativos aos nomes quando os mesmos possuíssem um valor de $\mathcal{S}_n$ superior a 70%. Deste modo, pode-se perceber pela Tabela 7.1 que a tabela de feições TF entre os dois *datasets* teria 163 *tuplas*. Neste caso, os bairros que não atendem à tolerância especificada ocupariam, cada um, uma *tupla* em TF.

7.3.2 Teste do atributo *geometria*

Teste nos polígonos – bairros da Cidade do Rio de Janeiro

Visando considerar apenas o uso do atributo geometria, procurou-se realizar o casamento destes ignorando o atributo nome. Assim, identificou-se dentre os pares (P_i, R_k), onde R_k é considerado o candidato para a combinação para um dado polígono P_i, o R_j cujo valor do $\mathcal{S}_g$ obtido fosse máximo. Desta forma, para cada bairro em um $dataset_1$ foi procurado e identificado o seu correspondente no $dataset_2$. De igual modo, para cada bairro do $dataset_2$ foi procurado e identificado o seu correspondente no $dataset_1$. Evidentemente, dado um polígono R_j considerado como o par de P_i não implica que P_i seja um par para R_j.

Com o intuito de permitir a valoração dos resultados obtidos, estabeleceu-se um valor para o índice de forma atribuir qualidade às geometrias. Assim, caso o $\mathcal{S}_g$ fosse maior do que 70% o resultado obtido indicava um casamento correto. Uma análise na

distribuição das quantidades de bairros obtidos por faixa de $\mathcal{S}_g$ (Tabela 7.2) permite inferir que os *datasets* possuem uma grande similaridade, logo as respostas oriundas de uma consulta em qualquer dos dois tem um elevado grau de confiabilidade.

Tabela 7.2: Valores de análise do $\mathcal{S}_g$

$MAXIMO(\mathcal{S}_g)$	quantidade de bairros
$0 \leq 70$	2
$70 \leq 80$	6
$80 \leq 90$	36
$90 \leq 95$	72
$95 \leq 100$	43

Ao se proceder a análise visual nos bairros onde $\mathcal{S}_g < 70\%$ – bairros em vermelho –, aqueles que estão compreendidos entre 70% e 90% – bairros em amarelo – e os demais que possuem $\mathcal{S}_g \geq 90\%$ – bairros em branco –, conforme pode ser verificado na Figura 7.4, tem-se que a percepção de que o *locus* geográfico, cujo $\mathcal{S}_g$ é elevado, é considerável dentro da Cidade do Rio de Janeiro.

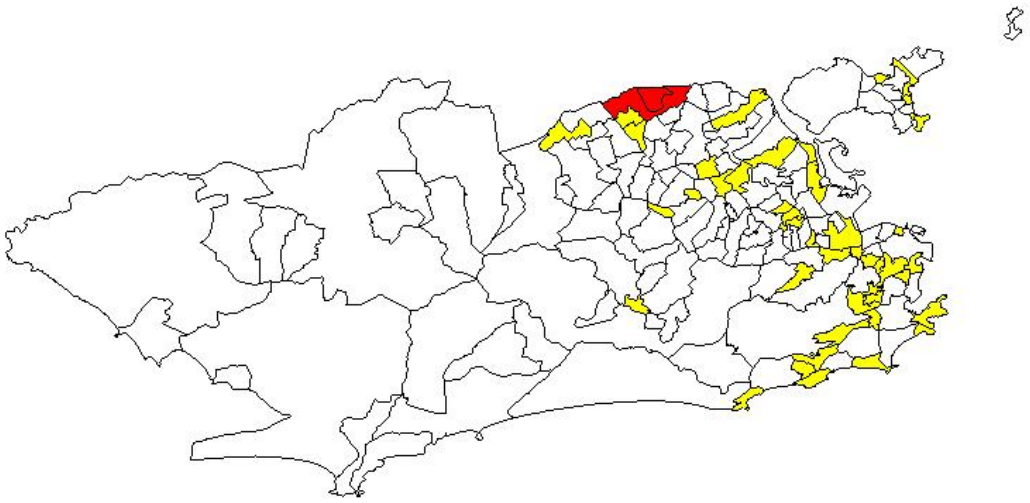

Figura 7.4: Distribuição de similaridade

Ao se processar dentro dos pares (P_i, R_j) o polígono união e o polígono interseção, verifica-se que a região onde não há dúvidas quanto a pertinência do *locus* geográfico dos bairros (Figura 7.5) possui uma área total de $1.187,3334\ km^2$ – área na cor verde. Considerando, ainda, a área total obtida pela união dos *datasets*, obtém-se

o valor de $1.251,4038\ km^2$ – área na cor amarela. Assim, percebe-se que o *locus* geográfico dentro da Cidade do Rio de Janeiro onde não há ambiguidade é de $94,88\%$. Acrescenta-se, ainda, que os *datasets* possuem área total de $1.217,4257\ km^2$ e de $1.221,3115\ km^2$, respectivamente.

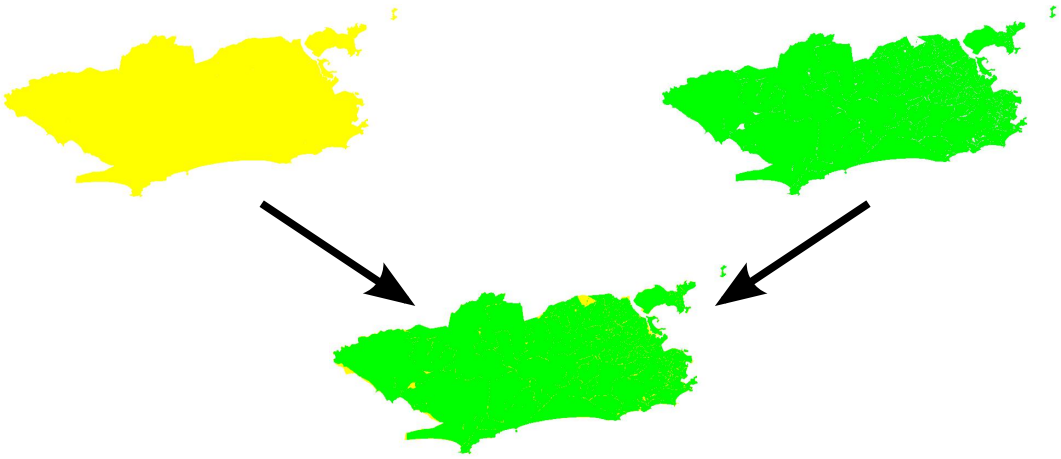

Figura 7.5: Relação entre a interseção e a união dos bairros

Enquanto apenas 2 bairros possuem $\mathcal{S}_g$ inferior a 70.00 % (Tabela 7.3), 7 outros possuem $\mathcal{S}_g$ superiores a 97.50 %, conforme pode ser observado na Tabela 7.4.

Tabela 7.3: Menores valores do $\mathcal{S}_g$

nome do bairro	$\mathcal{S}_g$ (%)
Parque Colúmbia	0.00
Pavuna	54.16

O motivo de o bairro "Parque Colúmbia" possuir um valor de $\mathcal{S}_g = 0$ é em função da existência de erros na construção de um dos dois *datasets* (Figura 7.6). Neste caso, os dados experimentais aqui são um exemplo de que uma certificação *a priori* ou a escolha aleatória de um *dataset* específico pode disponibilizar dados que não representam a realidade.

Diante dos dados obtidos, é possível perceber que a tabela de feições TF correspondente teria 161 pontos. Sendo que 157 atingem a tolerância e os bairros de Pavuna e Parque Colúmbia apareceriam com o atributo nome em duas *tuplas*.

Ao se avaliar o *nome* e a *geometria* em conjunto, obtém-se a tabela de feições TF correspondente, com 153 tuplas onde a similaridade semântica e geométrica atingem o limiar de 0.700 para as funções de similaridade $\mathcal{S}_n$ e $\mathcal{S}_g$ e 12 tuplas idênticas a

Tabela 7.4: Maiores valores do $\mathcal{S}_g$

nome do bairro	$\mathcal{S}_g$ (%)
Bangu	98.32
Campo Grande	98.14
Barra da Tijuca	97.83
Vargem Grande	97.70
Taquara	97.68
Jacarepaguá	97.59
Recreio dos Bandeirantes	97.57

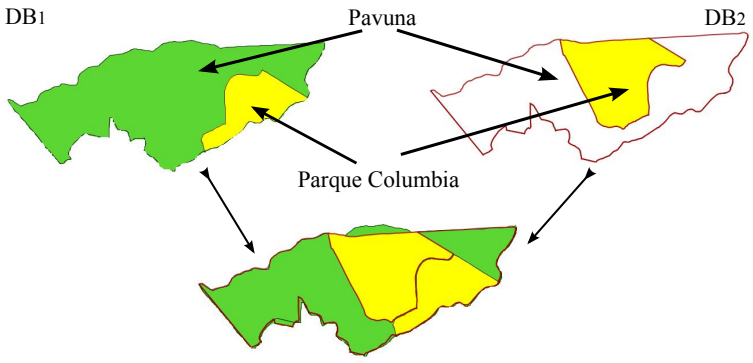

Figura 7.6: Indefinição – "Parque Columbia" *versus* "Pavuna"

alguma *tupla* dos *datasets* originais, perfazendo um total de 165 *tuplas*. Logo, a cobertura dos *datasets*, individualmente, será igual a 96.36%.

Teste em pontos

Com a finalidade de se quantificar o valor de ϵ (seção 4.3.3), identificou-se as coordenadas da caixa envolvente dos dois *datasets*. Assim, foram obtidos os dados da Tabela 7.5 que permitiu inferir para ϵ o valor de 5.00 m, em função de a escala original dos *datasets* serem de $\frac{1}{10000}$.

Tabela 7.5: Coordenadas das caixas envolventes

dataset	pt_{min}	pt_{max}	ℓ
1	(623.567,00; 7.446.647,00)	(695.172,00; 7.483.098,00)	0
2	(623.577,44; 7.447.341,00)	(694.953,00; 7.483.089,00)	0

Inicialmente, foram calculados os centróides de cada um dos polígonos que cons-

tituem os bairros da Cidade do Rio de Janeiro existentes nos *datasets*. De posse do valor para $\epsilon = 5.00\ m$, identificou-se de que o afastamento entre os centróides foram em todos os 159 casos superior a $10.00\ m$. Assim, em nenhum caso foi inferido o valor do $\mathcal{S}_g$ para os pontos que representavam o centróide do bairro.

Percebe-se, então, que os centróides representativos dos bairros não possuem relação com a similiaridade existente entre os polígonos que os representam. Desta forma, identifica-se que não há relação entre as similaridades dos polígonos e seus respectivos centros geométricos.

Embora todos os valores obtidos tenham sido 0, é possível identificar aqueles que se encontram mais próximos uns dos outros (Tabela 7.6) e aqueles que se encontram mais distantes (Tabela 7.7). Logo, percebe-se que embora as áreas dos bairros possuam uma alta similaridade, seus centros não encontram-se dentro do mesmo *locus* geográfico.

Tabela 7.6: Menores distâncias entre os centróides dos bairros

Bairro	Distância entre os centróides (m)
Irajá	18,35
Brás de Pina	33,97
Portuguesa	37,82
Maria da Graça	37,91
Saúde	42,89

Tabela 7.7: Maiores distâncias entre os centróides dos bairros

Bairro	Distância entre os centróides (m)
Senador Camará	2299,03
Vargem Grande	2038,89
Jacarepaguá	1845,10
Itanhangá	1642,16
Alto da Boavista	1513,61

Teste em linhas

Para se processar o $\mathcal{S}_g$ das linhas, fez-se uma transformação destas em polígonos representativos cujos *locus* geográficos permitem a inferência da similaridade. Assim, para cada ponto definidor da linha foi construído um polígono nos moldes do teste realizado nos pontos. Para cada par de polígonos gerados foi processado o fecho convexo. Desta forma, foi construído um polígono único para representar a linha, conforme pode ser observado na Figura 7.2.

Ao se comparar os limites dos bairros existentes nos *datasets* originais, percebe-se que há uma incerteza nos pontos definidores dos limites de forma a gerar um $\mathcal{S}_g$ destas linhas com valores muito aquém daqueles obtidos pelos polígonos que os definem. Assim, os maiores valores (Tabela 7.8) encontrados apresentam um $\mathcal{S}_g$ pouco superior a 15%. Por sua vez, os menores valores (Tabela 7.9) são inferiores a 3%.

Tabela 7.8: Maiores valores de $\mathcal{S}_g$ para os limites

Bairro	$\mathcal{S}_g$ dos limites (%)
Abolição	17,62
Magalhães Bastos	17,40
Oswaldo Cruz	16,54
Higienópolis	15,96
Bento Ribeiro	15,94

Tabela 7.9: Menores valores de $\mathcal{S}_g$ para os limites

Bairro	$\mathcal{S}_g$ dos limites (%)
Parque Colúmbia	0,18
Tomás Coelho	1,74
Moneró	2,31
Olaria	2,52
Vicente de Carvalho	2,56

Assim como no caso dos centróides, percebe-se que o $\mathcal{S}_g$ obtido junto às linhas limitrófes dos bairros não possuem a mesma representatividade que seus polígonos. Embora gerem resultados melhores do que os centróides, os resultados possibilitam

afirmar de que há uma imprecisão considerável nos limites. Entretanto, quando se analisa a área delimitada pelas linhas, as respostas obtidas por meio de uma consulta possuem um alto grau de similaridade.

Atributo $nome \times geometria$

Uma última ánálise foi realizada a partir da *tupla* $(nome, geometria)$. Neste caso, fez-se a correlação entre os dois índices – Índice de Similaridade Não Espacial ($\mathcal{S}_n$) e o Índice de Similaridade Espacial ($\mathcal{S}_g$) – para permitir uma inferência melhor sobre a tupla e não sobre os valores individuais das instâncias das colunas.

Assim, o índice geral (IG) é obtido pelo produtos dos índices anteriores, ou seja, $IG = \mathcal{S}_n \cdot \mathcal{S}_g$. Evidentemente, quanto mais próximo do valor 1 o IG estiver, melhor será o resultado do processamento das ambiguidades. O valor de IG serve como uma probabilidade ou um grau de certeza.

7.4 Análise

Diante dos resultados obtidos junto aos *datasets* oriundos do $IBGE$ e do IPP é possível verificar que a análise das ambiguidades é melhor visualizada quando se tem polígonos representativos das feições do terreno. Evidentemente, não há uma refutação da similaridade das linhas e dos pontos. Entretanto, verifica-se que os valores do $\mathcal{S}_g$ necessários para associar uma linha como similar a uma outra não possuem os mesmos valores caso fossem polígonos. Neste caso, o Método dos Retângulos Equivalentes (MRE) [4] é mais indicado para a análise entre representações lineares. A estimativa de um afastamento médio traz mais informação do que a similaridade entre as linhas. Por sua vez, associar o grau de similaridade permite identificar semelhança entre linhas.

No caso dos pontos, as prescrições legais de tolerância não permitem gerar um polígono representativo de forma a possuir um valor de $\mathcal{S}_g$ adequado. Neste caso, uma solução melhor pode ser obtida ao se identificar pontos homólogos pelo atributo nome. Posteriormente, pode ser calculada a variância entre os possíveis pontos candidatos a representarem a mesma feição do terreno.

7.5 Considerações finais

Os resultados obtidos junto aos dois *datasets* – $IBGE$ e IPP – atestam que o $\mathcal{S}_g$ funciona perfeitamente para o caso dos polígonos. Assim, percebe-se que os valores do $\mathcal{S}_g$ fornecem ao tomador de decisão uma possibilidade maior de utilização de dados. As ambiguidades geram uma superabundância de informações que um usuário pode fazer uso com a finalidade de decidir da melhor forma possível. De igual modo, o valor final do $\mathcal{S}_g$ no caso das *geometrias*, do $\mathcal{S}_n$ no caso dos *nomes* e do IG servem como um valor percentual da qualidade dos dados existentes porventura disponíveis.

O $\mathcal{S}_n$ funciona para qualquer tipo de *nome*. Assim, todo e qualquer dado semântico pode ser avaliado por este Índice. Entretanto, a *geometria* é melhor avaliada quando sua representação é poligonal. O valor estabelecido, nesta tese, como referencial para os limiares $\mathcal{L}_n$ e $\mathcal{L}_g$ (70%) serviu apenas para assegurar que a arquitetura proposta funcionasse. Evidentemente, o valor pode ser alterado a critério do usuário. De modo genérico, os limiares servem com avaliadores da qualidade relativa entre os *datasets*.

Capítulo 8

Conclusões

A construção das bases geográficas envolve uma série de procedimentos que procuram assegurar fidelidade de representação de feições do mundo real. O BDG gerado é uma visão específica do mundo real sujeita a uma série de variáveis que tornam a sua elaboração peculiar. Há o efeito temporal, a variável humana, os métodos empregados, as precisões alcançadas e a escolha dos pontos definidores das representações cartográficas. Diante deste espectro de injunções para a construção de BDG's, dificilmente dois produtores chegarão a resultados idênticos, já que são criadas em momentos distintos, empregando diferentes métodos ou tendo como objetivos finalidades distintas.

Atualmente, o processo de construção dos *datasets* vem sendo continuamente alterado, empregando-se meios computacionais de forma a garantir uma maior velocidade na produção dos dados. Entretanto, isto não garante a unicidade de dados geográficos. Assim, o que se tem nos dias atuais é uma diversidade de *datasets* de uma mesma região geográfica. Esta diversidade é usualmente denominada de ambiguidade. Como a prática da produção e uso destes dados requerem unicidade de representação, diversos meios de se proceder uma integração dos dados existem.

Este paradigma, por sua vez, impede que informações sejam produzidas rapidamente, haja vista que o processo de integração é demorado. A escolha ou a certificação de um produtor não é algo que, geralmente, se decide tecnicamente. Assim, sempre há uma dúvida quanto à qualidade final dos dados disponibilizados. Visando tratar este problema, a presente tese apresenta uma mudança no paradigma de processamento de consulta, efetuando a integração de respostas em dados geográficos

multirepresentados.

Para tal foi apresentado o Sistema Avaliador de Respostas Ambiguas ($SARA$) que serve como $interface$ para que o usuário consiga realizar as tarefas necessárias para obter os respectivos índices de similaridade. Embora as ambiguidades não sejam desejadas pelos produtores, sua existência nos impõe a necessidade de avaliá-las para que tenha a maior confiabilidade possível nas consultas.

As ambiguidades porventura existentes são avaliadas por meio de dois índices de similaridade – um não espacial $\mathcal{S}_n$ e outro espacial $\mathcal{S}_g$ – visando fornecer ao usuário valores que o permitam inferir a qualidade conjunta dos dados existentes. Evidentemente, quanto maior a similaridade entre os dados maiores serão os valores obtidos nos índices. Desta forma, o usuário pode verificar se há dúvidas entre os produtores, bem como pode identificar regiões onde, embora os dados sejam múltiplos, não haja dúvidas. Para tal, este usuário executa as consultas em SQL em uma tabela de feições que sumariza todas as informações sobre um determinado tema. Uma vez identificadas possíveis respostas, a arquitetura viabiliza a identificação das representações originais por meio de uma tabela auxiliar que mapeia os diversos *datasets* com a tabela de feições. Uma vez realizado este procedimento, é obtida uma relação final contendo todas as representações disponíveis, onde uma série de análises podem ser desenvolvidas.

De posse da relação final podem ser inferidos o *locus* geográfico máximo e mínimo abrangidos pela possibilidades: de avaliar a cobertura e a completude de um *dataset* específico; de quantificar as feições mapeadas pelos diversos *datasets*; de informar ao usuário uma resposta qualificada e é possível, ainda, identificar problemas na modelagem dos dados, tal como o ocorrido com as representações do bairros de Parque Colúmbia. A identificação de tais inconsistências é relevante para que se possa reduzir custos, ao se perceber onde a inspeção de campo se faz necessária, e para a eliminação de problemas nas modelagens ao se eliminar representações antagônicas.

Diante do apresentado no corpo da tese, infere-se que os índices são mais adequados para a avaliação dos *nomes* geográficos das feições e das *geometrias* representativas das mesmas quando estas são polígonos. Isto porque, embora o índice $\mathcal{S}_n$ produza resultados, tanto para linhas como para pontos, os valores obtidos não

possuem a mesma relevância daqueles obtidos pelos polígonos. No caso de linhas e pontos, o índice não espacial funciona com a mesma eficácia daquela obtida para o processamento de polígonos. Entretanto, no aspecto espacial, o índice adotado necessita de melhoria.

Ressalta-se, ainda, que os experimentos serviram para atestar de que a tese proposta gera informações a partir dos dados ambíguos. A identificação da similaridade nas representações é crucial para a construção da Tabela de Feições (TF). Assim, como pode ser observado no capítulo 7, os índices propostos servem para inferir as correspondências entre as diversas representações. Uma vez construída a TF, as consultas realizadas podem ser efetuadas sobre as feições cadastradas na tabela, permitindo a obtenção de uma resposta integrada.

8.1 Propostas para trabalhos futuros

A presente tese permite uma série de possibilidades de desenvolvimentos futuros, dentro das quais, destacam-se as seguintes possibilidades:

Certificação de dados: Consiste em avaliar a qualidade de dados por meio de comparação com várias modelagens do mesmo tema. Evidentemente, a tese propõe o acesso indiscriminado a todos os dados existentes. Entretanto, erros na construção do dados podem ser identificados – caso de Parque Colúmbia e Pavuna – e impõe uma verificação de campo. Neste caso, a existência de ambiguidades e a análise dos valores obtidos nos índices podem apontar para um problema insolúvel sob o ponto de vista estritamente computacional. Métodos para se identificar problemas deste tipo são uma vertente da presente tese que servirá para a melhoria dos dados disponíveis e a identificação de produtores comprometidos com a técnica cartográfica.

Similaridade de pontos: Ficou provado que as atuais prescrições técnicas para a obtenção de coordenadas plani-altimétricas de pontos do terreno com o intuito de se produzir uma base, não favorecem a identificação de similaridade entre os pontos. Assim, a repetição de pontos com *nomes* similares e *geometrias* distintas podem gerar informação inadequada. Logo, faz-se necessário criar novas formas de se identificar os pontos por meio de uma análise da similaridade entre eles.

Identificação de atualizações: Comparar *datasets* distintos e se identificar

o quanto um, em especial, encontra-se tão atualizado quanto o outro. Assim, é possível reduzir as atividades de atualização ao se identificar as representações com alta similaridade, aquelas com inconsistências e, principalmente, as que são idênticas – possível fonte de plágio.

Desenvolvimento de índices: Aplicar novos índices aos atributos não espaciais e aos espaciais. É evidente que o índice espacial aplicado é mais sensível nas representações poligonais. Logo, é conveniente o desenvolvimento de novos índices para que sejam aplicados nas representações lineares e nas pontuais, visando identificar melhor a similaridade entre ambos.

Finalmente, cabe-nos reconhecer que a arquitetura SARA ainda é um arcabouço para processamento de consultas não automático. Apesar de termos desenvolvido implementações para os Algoritmos constantes desta tese, estes foram aplicados a *datasets* e consultas específicas. Uma implementação completa do SARA requer o desenvolvimento de estruturas de dados para o Catálogo de Domínio bem como algoritmos para o processamento automatizado de meta-consultas. De forma análoga, as propostas para o Classificador Analítico de Ambiguidades devem ser melhor investigadas, sendo que ao usuário deve ser facultado a escolha dos métodos de classificação e de exibição dos resultados.

Referências Bibliográficas

[1] RAISZ, E., *Cartografia geral*. Científica: Rio de Janeiro, Brasil, 1969.

[2] FILETO, R., "Issues on Interoperability and Integration of Heterogeneous Geographical Data". In: *Proceedings Terceiro Simpósio Brasileiro de Geoinformática*, Rio de Janeiro, Brasil, 2001.

[3] HESSEN, J., *Teoria do conhecimento*. Martins Fontes: Rio de Janeiro, Brasil, 2003.

[4] FERREIRA DA SILVA, L. F. C., *Avaliação e integração de bases cartográficas para cartas eletrônicas de navegação terrestre*, Ph.D. Thesis, Escola Politécnica da Universidade de São Paulo, São Paulo, Brasil, 1998.

[5] COELHO, V. B. N., STRAUCH, J. C. M., ESPERANÇA, C., "Similarity among multiple geographic representations". In: *GeoWeb 2009 Academic Track – Cityscapes*, v. XXXVIII-3-4/C, pp. 16 – 21, International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences – ISPRS: Vancouver, Canada, 2009.

[6] ALI, A. B. H., "Positional and shape quality of areal entities in geographic databases: quality information aggregation versus measures classification". In: *ECSQARU´2001 Workshop on Spatio-Temporal Reasoning and Geographic Information Systems*, Toulouse, França, 2001.

[7] KIELER, B., SESTER, M., WANG, H., et al., "Semantic Data Integration: data of similar and different scales", *Photogrammetrie Fernerkundung Geoinformation*, v. 6, pp. 447 – 457, 2007.

[8] BRASIL, "Decreto Nr 89.817, de 20 de junho de 1984: Instruções reguladoras das normas técnicas da Cartografia Nacional", Diário Oficial da União, 1984.

[9] ABNT, "Rede de Referência Cadastral Municipal - Procedimento", NBR 14166, 1998.

[10] UCHÔA, H. N., DE PAULO, M. C. M., FILHO, L. C. T. C., et al., "Evaluation of Data Conversion of Vectorial Geographic Features in Topographic Maps using Free Software Tools". In: *Workshop de Software Livre*, Porto Alegre, Brasil, 2006.

[11] BRASIL, "Diretriz da Implementação do Software Livre no Governo Federal", Diário Oficial da União, 2003.

[12] FILHO, J. L., COSTA, A. C., IOCHPE, C., "Projeto Banco de Dados Geográficos: mapeando esquemas GeoFrame para o SIG Spring". In: *I Brazilian Workshop on Geoinformatics – GEO-INFO 99*, São José dos Campos, Brasil, 1999.

[13] LUNARDI, O. A., DA SILVA MEYER, W., TRINDADE, C. A., et al., "Banco de Dados Geográficos do Exército (BDGEx)". In: *Anais do XXI Congresso Brasileiro de Cartografia*, Belo Horizonte, Brasil, 2003.

[14] QUAN WU, M., LONG WANG, Z., DING ZHANG, A., et al., "Ontology-driven Heterogeneous Geographic Data Set Integration". In: *Global Congress on Intelligent Systems*, Xiamen, China, 2009.

[15] WIEDERHOLD, G., "Mediators, Concepts and Practice". In: *Handbook of databases*, 2007.

[16] UITERMARK, H. T., VAN OOSTEROM, P. J., MARS, N. J., et al., "Ontology-based integration of topographic data sets", *International Journal of Applied Earth Observation and Geoinformation*, v. 7, n. 2, pp. 97 – 106, 2005.

[17] COELHO, V. B. N., *Algoritmo para edição cartográfica entre bordas de folhas*, Master's Thesis, Instituto Militar de Engenharia, Rio de Janeiro, Brasil, 2001.

[18] SHETH, A., LARSON, J., "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases", *ACM Computing Surveys*, v. 22, n. 3, pp. 183 – 236, 1990.

[19] ALI, M. G., "Object-oriented approach for integration of heterogeneous databases in a multidatabase system and local schemas modifications propagation", *International Journal of Computer Science and Information Security*, v. 6, n. 2, pp. 55 – 60, 2009.

[20] FAHLAND, D., GLABER, T. M., QUILITZ, B., et al., "HUODINI Flexible Information Integration for Disaster Management". In: *Proceedings of the 4th International ISCRAM Conference*, 2007.

[21] NIWATTANAKUL, S., MARTIN, P., EBOUEYA, M., et al., "Learning Object Mediation System Based on an Ontology Model". In: *Proceedings of the Fourth International Conference on eLearning for Knowledge-Based Society*, 2007.

[22] PAZINATTO, E., DE SOUZA BAPTISTA, C., DE ALMEIDA VILAR DE MIRANDA, R., "GeoLocalizador: um Sistema de Referência Espaço-Temporal Indireta utilizando um SGBD Objeto-Relacional". In: *Anais do GeoInfo 2002*, Caxambu, Brasil, 2002.

[23] MILLS, J. W., CURTIS, A., PINE, J. C., et al., "The clearinghouse concept: a model for geospatial data centralization and dissemination in a disaster", *Disasters*, v. 32, n. 3, pp. 467 – 479, 2008.

[24] LORD, P., MACDONALD, A., LYON, L., et al., "From data deluge to data curation". In: *e-Science All Hands Meeting 2004*, Nottingham, Reino Unido, 2004.

[25] BEAGRIE, N., "Digital Curation for Science, Digital Libraries, and Individuals", *The International Journal of Digital Curation*, v. 1, n. 1, pp. 3 – 16, 2006.

[26] CHARLESWORTH, A., "Digital Curation, Copyright, and Academic Research", *The International Journal of Digital Curation*, v. 1, n. 1, pp. 17 – 32, 2006.

[27] AGUILAR, F. J., CARVAJAL, F., AGUILAR, M. A., et al., "Developing digital cartography in rural planning applications", *Computers and Electronics in Agriculture*, v. 55, n. 2, pp. 89 – 106, 2007.

[28] VAIRAVAMOORTHY, K., YAN, J., GALGALE, H. M., et al., "IRA-WDS: A GIS-based risk analysis tool for water distribution systems", *Environmental Modelling & Software*, v. 22, n. 7, pp. 951 – 965, 2007.

[29] ZHAO, H., RAM, S., "Combining schema and instance information for integrating heterogeneous data sources", *Data & Knowledge Engineering*, v. 61, n. 2, pp. 281 – 303, 2007.

[30] KEANE, R. E., ROLLINS, M., ZHU, Z.-L., "Using simulated historical time series to prioritize fuel treatments on landscapes across the United States: The LANDFIRE prototype project", *Ecological Modelling*, v. 204, n. 3-4, pp. 485 – 502, 2007.

[31] BUCCELLA, A., CECHICH, A., "Towards Integration of Geographic Information Systems", *Electronic Notes in Theoretical Computer Science*, v. 168, pp. 45 – 59, 2007.

[32] KOCH, A., HEIPKE, C., "Semantically correct 2.5D GIS data – The integration of a DTM and topographic vector data", *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 61, n. 1, pp. 23 – 32, 2006.

[33] ELMASRI, R., NAVATHE, S., *Sistemas de Banco de Dados*. Pearson Education do Brasil Ltda: São Paulo, Brasil, 2006.

[34] CASANOVA, M. A., CÂMARA, G., DAVIS, C., et al., *Banco de dados geográficos*. Editora Mundogeo: Curitiba, Brasil, 2005.

[35] LIMA, E. L., *Espaços métricos*. Instituto Nacional de Matemática Pura e Aplicada: Rio de Janeiro, Brasil, 2009.

[36] WILLIAM COHEN, P. R., FIENBERG, S., "A comparison of string metrics for matching names and records". In: *Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, 2003.

[37] FISCHER, I., ZELL, A., "String averages and self-organizing maps for strings". In: *Proceedings Second ICSC Symposium on Neural Computation*, 2000.

[38] HAMMING, R., "Binary codes capable of correcting deletions, insertions, and reversals", *Soviet Physics Doklady*, v. 10, pp. 707–710, 1966.

[39] DAMERAU, F., "A technique for computer detection and correction of spelling errors", *Communications of the ACM 7*, v. 3, pp. 171 – 176, 1964.

[40] JACCARD, P., "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", *Bulletin del la Société Vaudoise des Sciences Naturelles*, v. 37, pp. 547–579, 1901.

[41] VAN RIJSBERGEN, C. J., "Retrieval effectiveness", *Progress in Communication Sciences*, v. 1, pp. 91–118, 1979.

[42] HAMMING, R., "Error detecting and error correcting codes", *Bell System Technical Journal*, v. 29, pp. 147–160, 1950.

[43] WINKLER, W., "The state of record linkage and current research problems". In: *Proceedings of the Survey Methods Section*, pp. 73–80, 1999.

[44] BRASIL, "Lei nº 10.267, de 28 de agosto de 2001: Lei de Cadastro de Imóvel Rural", Diário Oficial da União, 2001.

[45] PULLAR, D., "Consequences of using a tolerance paradigm in spatial overlay". In: *Proceedings of the AutoCarto 11*, pp. 288 – 296, Minnesota, Estados Unidos, 1993.