Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

ESCOLA NACIONAL DE SAÚDE PÚBLICA
SERGIO AROUCA
ENSP

"Metodologias de Geocodificação dos Dados da Saúde"

por

Daniel Albert Skaba

Tese apresentada com vistas à obtenção do título de Doutor em Ciências na área de Saúde Pública.

Orientador: Christovam de Barcellos

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

ESCOLA NACIONAL DE SAÚDE PÚBLICA
SERGIO AROUCA
ENSP

Rio de Janeiro, julho de 2009.

Esta tese, intitulada

"Metodologias de Geocodificação dos Dados da Saúde"

apresentada por

Daniel Albert Skaba

foi avaliada pela Banca Examinadora composta pelos seguintes membros:

Profa. Dra. Simone Maria dos Santos

Prof.ª Dr.ª Cláudia Medina Coeli

Prof.ª Dr.ª Evangelina Xavier Gouveia de Oliveira

Prof. Dr. Oswaldo Gonçalves Cruz

Prof. Dr. Christovam Barcellos – Orientador

Profa. Dra. Marilia Sá Carvalho – co-orientadora

Tese defendida e aprovada em 31 de julho de 2009.

Catalogação na fonte
Instituto de Comunicação e Informação Científica e Tecnológica
Biblioteca de Saúde Pública

S237 Skaba, Daniel Albert
Metodologias de geocodificação dos dados da saúde. /
Daniel Albert Skaba. Rio de Janeiro : s.n., 2009.
155 f., il., tab., mapas

Orientador: Barcellos, Christovam de Castro
Carvalho, Marília Sá
Tese (Doutorado) Escola Nacional de Saúde Pública Sergio Arouca

1. Sistemas de Informação Geográfica. 2. Distribuição Espacial da População. 3. Estudos Epidemiológicos. 4. Estudos de Casos. I. Título.

CDD – 22.ed. – 616.959

Aos meus pais, que me ensinaram que querer e saber construiriam meu caminho.

À Rebeca que me acompanha nesta contrução.

Aos meus filhos, Marcelo e Tatiana, que são os melhores resultados.

“Tudo deveria ser tornado tão simples quanto possível.
Mas não mais simples do que isso”

Albert Einstein

BREVE HISTÓRICO E MUITOS AGRADECIMENTOS

Esta história se inicia no início dos anos 1990 com a Marília (Sá Carvalho) e suas brilhantes ideias, batendo à porta do IBGE, em busca de soluções para os estudos utilizando análise espacial na área da saúde. Nesta busca, encontrou o Paulo Cesar (Martins), com seus conhecimentos acumulados sobre o Censo, principalmente no que se refere à Base Territorial. Destes encontros saíram um programa de busca de endereços, que teve a colaboração do Oswaldo (Gonçalves Cruz); e a digitalização dos contornos dos setores censitários 1991 do município do Rio de Janeiro, a partir dos mapas em papel com a marcação dos setores em caneta hidrocor.

Nos anos seguintes foi criado, com participação do Christovam (Barcellos) e Marilia, o Comitê Temático Interinstitucional sobre análise de dados espaciais, da Ripsa (CTI Geo-saúde), com proposta de utilizar o setor censitário como unidade de análise espacial. Isto foi um estímulo a mais para a ideia, já amadurecida, de se criar uma base de dados digital dos setores censitários.

Neste tempo fiz meu mestrado no IME em Cartografia digital e participei, junto com Paulo Cesar e Sonia (Terron), da construção da componente gráfica da base territorial do Censo 2000. Um projeto ousado, desenvolvido inteiramente dentro do IBGE, que tinha como objetivo inicial a construção de mapas digitais de setores urbanos das 1028 maiores cidades do Brasil. O resultado foi bem maior que aqueles objetivos: foram construídos os mapas digitais dos setores de todas as 5507 cidades existentes em 2000, além de um cadastro de segmento de logradouros e da criação de um banco de dados a partir da digitalização das Folhas de Coleta.

Depois disso veio a ideia do doutorado, com o objetivo de somar conhecimentos e de aplicar todo esse conhecimento e material acumulado em algo novo. E, em um novo encontro Marilia e Christovam, apareceu a Saúde Pública na minha vida. Foi como se iniciasse uma nova carreira, depois de mais de 30 anos de trabalho. E com ela, muitos amigos novos e a estatística espacial, que me fez voltar à base territorial com os endereços. Soma-se a tudo isso o Projeto ELSA.

Os meus agradecimentos começam pela minha família, que me apoiou, mesmo com tantas “ausências”. Aos orientadores, pela oportunidade de ingressar no doutorado, mesmo sem ser especialista desta área. À Marília, em especial, e à Evangelina, pela disponibilidade e paciência nas revisões desta tese. Ao IBGE, pela liberação, pelos dados fornecidos e pela bagagem acumulada. Aos amigos Sonia e Paulo Cesar, por construirmos juntos toda esta bagagem. À ENSP, pela estrutura disponibilizada e pelos fantásticos cursos, que me deram uma visão ampla do que é a Saúde Pública. Aos colegas de turma, ou melhor, novos amigos, que dividiram toda essa experiência. Aos amigos que dividiram suas experiências, dando oportunidade aos estudos de casos: Fabíola Nunes (que, inclusive, me aceitou como colaborador em sua dissertação), Wagner Tassinari, Eliana Bender Martins, Dayse Campos e todos os componentes e participantes do Projeto ELSA. Fico por aqui, porque, se eu for citar todos que passaram por este caminho e contribuíram para este “final feliz”, teria que gastar mais folhas que a própria tese. De qualquer forma, muito obrigado a todos.

METODOLOGIAS DE GEOCODIFICAÇÃO DOS DADOS DA SAÚDE

Autor: DANIEL ALBERT SKABA

Orientador: CHRISTOVAM DE CASTRO BARCELLOS

Co-Orientadora: MARILIA SÁ CARVALHO

RESUMO

O objetivo geral desta tese é estudar as necessidades de geocodificação geradas pelos estudos epidemiológicos e propor um método que permita a associação dos endereços dos eventos de Saúde Pública a uma localização geográfica, utilizando como base os cadastros do Censo 2000 do IBGE. Para esta finalidade, são pesquisados os métodos de análise espacial em Epidemiologia e os tipos de endereço no mundo e, particularmente, no Brasil, além dos métodos de tratamento de textos, como as pesquisas fonéticas e os algoritmos de comparação de strings, assim como os métodos de comparação de arquivos. Para investigação dos procedimentos envolvidos no processo, foram feitos 5 estudos de caso, expondo as características e dificuldades encontradas no processo. Como resultado destes estudos e destas experiências, uma proposta de metodologia é apresentada, com definição de uma base de dados e de funções necessárias para o tratamento da entrada de dados e de buscas através de comparações de textos, com vistas ao desenvolvimento de um aplicativo de uso público.

Palavras chaves: Geocodificação, Sistemas de Informação Geográfica, Análise Espacial

GEOCODING METHODOLOGIES OF HEALTH DATA

Author: DANIEL ALBERT SKABA

Advisors: CHRISTOVAM DE CASTRO BARCELLOS

MARILIA SÁ CARVALHO

Abstract

This thesis propose methods that associate health events with geographic locations to provide geocoding needs for epidemiology studies using the Brazilian Census 2000 databases. To achieve the goals, this work assesses: spatial anlysis methods in Epidemiology, types of addresses around the world, data mining methods, Record linkage, phonetic and string matching algorithms. Five case studies investigate the procedures, characteristics and problems of geocoding process. The methodology proposed is presented in results of these research and experiences. To develop an open software, it defines: a database definition, data entry treatment rotines and string matching routines

Keywords: Geocoding, Geographic Information Systems, Spatial Analysis

ÍNDICE

LISTA DE FIGURAS

1. INTRODUÇÃO

"Todas as coisas são parecidas, mas coisas mais próximas se parecem mais que coisas mais distantes" (Waldo Tobler)

O escopo deste trabalho é o georreferenciamento de endereços urbanos, a partir de informações da saúde, obtendo-se como resultado um par de coordenadas ou uma área significativa, ou seja, com atributos socioeconômicos e ambientais associados. Deste modo pode-se associar um conjunto de variáveis geográficas e tabulares correlacionadas, em condições de serem trabalhadas em Sistemas de Informação Geográfica (SIG) e com informações necessárias para se efetuar análises espaciais com aplicações na Epidemiologia e Saúde Pública.

No Brasil há uma grande diferença no nível de produção de bases cartográficas digitais para utilização em SIG (Barcellos et al., 2008) nas diversas regiões do país, fazendo com que se necessite criar alternativas para atender as necessidades de cada projeto.

No Censo 2000, o IBGE produziu para todas as cidades brasileiras as malhas digitais de setores censitários com um padrão único para utilização em SIG (Skaba & Terron, 2003). Estes elementos geográficos possuem a vantagem de estarem associados às informações dos censos brasileiros (IBGE, 2002). Foi produzido também o Cadastro de Segmentos de Logradouros (Cadlog), com os logradouros pertencentes a cada setor censitário e sua numeração inicial e final, para os setores urbanos dos 1028 maiores municípios brasileiros.

Embora haja discussões entre os especialistas sobre alguns critérios de determinação de urbano ou rural, já vivem hoje nas áreas urbanas brasileiras, segundo o IBGE, mais de 80% da população, ou 148 milhões de pessoas. Em 40 anos, de 1960 a 2000, nada menos de 107 milhões de pessoas se somaram à população urbana brasileira, por força das altas taxas de natalidade, principalmente nas décadas de 60/70 e 70/80, e das migrações internas. Nesse processo, a população urbana, que representava 44,67% do total em 1960 (31,3 milhões), passou para mais de 80% no Censo de 2000, enquanto a rural (36,76 milhões ou 55,33%) caiu para menos de 20% no fim do século, com um número de pessoas inferior ao de 40 anos antes.

As ações e propostas de intervenção e planejamento devem se orientar, cada vez mais, a partir de relações entre as partes e o todo urbano (Ramos et al., 2007). Neste

sentido, a construção de territórios digitais urbanos, com a definição de divisões intra-urbanas, é importantíssima na formulação de políticas públicas, para que a distribuição de serviços e benefícios seja consistente com o público alvo.

Na investigação dos problemas de saúde pública, são analisados dados relativos ao meio ambiente, ao perfil de morbidade e de mortalidade, à disponibilidade de equipamentos urbanos, à situação socioeconômica e à utilização dos serviços de saúde. Com base na divisão político administrativa, ou em áreas base de censos e pesquisas, são observados os indicadores disponíveis na área estudada, como etapa do planejamento ou como parte de um processo de identificação de um problema particular (Carvalho & Cruz, 1998).

A utilização de mapas e a preocupação com a distribuição geográfica de diversas doenças é bem antiga. Há cerca de 2400 anos Hipócrates, em seu tratado "Ares, Águas, Lugares", escreveu "*Vocês descobrirão, como uma regra geral, que os estados de saúde e hábitos das pessoas seguem a natureza do local onde vivem*" (Hippocrates et al., 1983). Depois disto, o médico (cirurgião naval) escocês James Lind publicou em 1768 um livro chamado "*An Essay on Diseases Incidental to Europeans in Hot Climates*" no qual procura explicações para a distribuição de doenças, chegando inclusive a atribuir riscos a determinadas áreas geográficas especificas (Barret, 1991). Desde então, diversos trabalhos foram escritos, descrevendo variações geográficas na distribuição das doenças. O mais famoso e marcante é o estudo realizado pelo médico inglês John Snow (1990) que, observando que os casos de cólera ocorriam mais em certas localizações, fez uso dos mapas de Londres e dos registros de óbitos. Ele utilizou os endereços das residências e dos poços de provisão de água existentes na região. Na figura 1.1 observamos um mapa da região de Londres atingida, com a identificação dos casos (●) e das bombas de água (P). Na figura 1.2, um detalhe do mapa preparado por Snow com o registro dos óbitos por local de residência. A partir destas informações, foi observada uma associação entre a maior quantidade de casos e a proximidade com a bomba de água de Broad Street (figura 1.3).

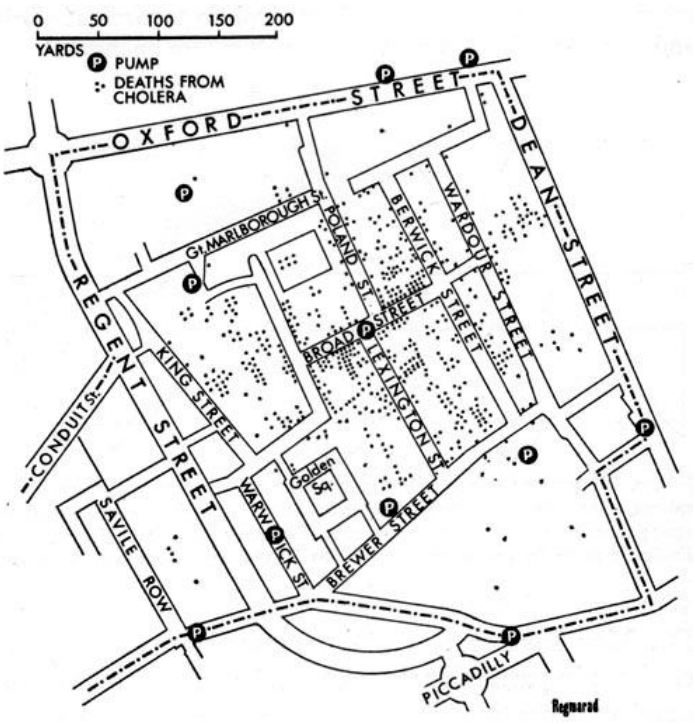


Figura 1.1. Ocorrências de cólera e posição das bombas

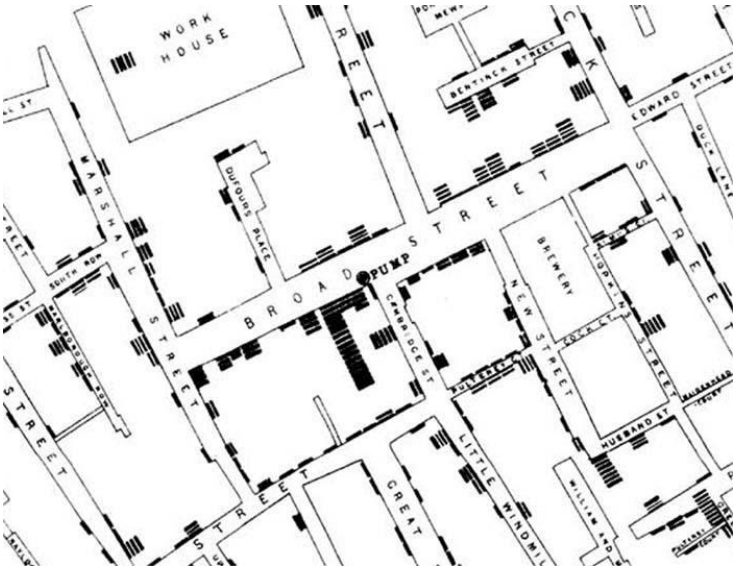


Figura 1.2. Casos de cólera anotados por John Snow

Uma questão fundamental para o planejamento do setor saúde é a distribuição, no espaço geográfico, dos serviços e de seus usuários. Isto facilita a investigação dos

fluxos de pessoas que demandam tais serviços, conectando residência e local de atendimento.

Figura 1.3. Bomba d'água de Broad Street

As análises de trajeto são úteis no planejamento da oferta de serviços de saúde (Francis & Schneider, 1984) e na análise dos deslocamentos populacionais, provocando também o deslocamento de vetores, hospedeiros ou parasitas. Seus fluxos são determinantes na compreensão dos mecanismos de propagação endemo/epidêmica (Smallman-Raynor & Cliff, 1991). Oliveira et al. (2004) visam verificar, através do mapeamento das redes estabelecidas pelo deslocamento das pessoas que buscam atendimento, em que medida a implantação do Sistema Único de Saúde (SUS) se aproxima ou se distancia de seus objetivos de promover a universalização do acesso aos serviços de saúde. Na figura 1.4 observa-se a rede de deslocamento para cirurgias cardíacas.

A delimitação das áreas de abrangência de uma unidade de saúde com base na utilização efetiva pela população permite investigar os níveis de oferta per capita, identificando e orientando a superação de desigualdades na distribuição dos serviços,

bem como analisar os dados de saúde e doença em confronto com as informações socioeconômicas disponíveis.

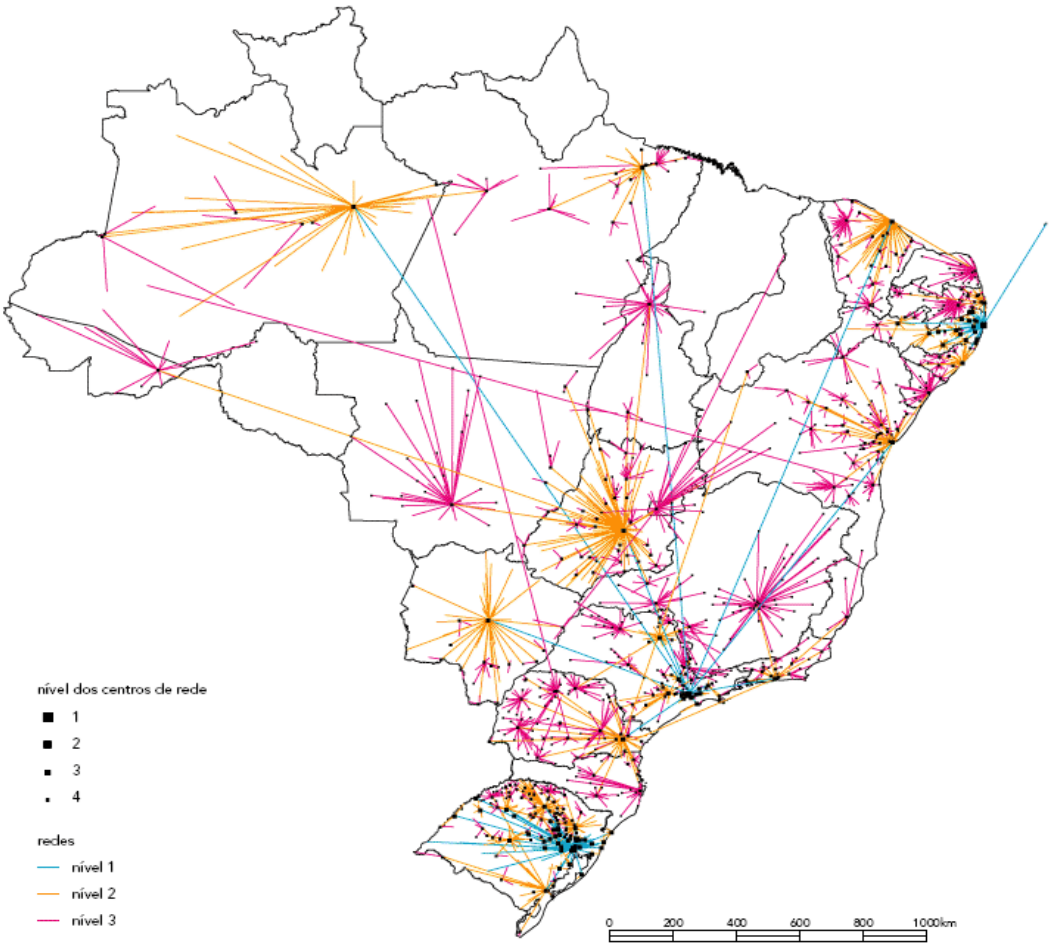


Figura 1.4. Redes de atenção hospitalar básica. População de 5 anos e mais (Oliveira et al., 2004)

Desigualdades no uso dos serviços de saúde, isto é, na atitude de procurá-los, obter acesso e se beneficiar com o atendimento recebido, refletem as desigualdades individuais no risco de adoecer e morrer, assim como as diferenças no comportamento do indivíduo perante a doença, além das características da oferta de serviços que cada sociedade disponibiliza para seus membros. A utilização deste espaço geográfico na investigação das desigualdades socioeconômicas na incidência e mortalidade nas doenças e no acesso ao serviço de saúde é observada em alguns trabalhos (Krieger et al. 2003, 2003-2 e 2005; Curtis, 1998; Pittman, 1986; Almeida-Filho, 2004).

Os desenvolvimentos de tecnologias de mapeamento digital e de análise espacial, principalmente nas duas últimas décadas com a utilização cada vez mais crescente dos Sistemas de Informação Geográfica (SIG), abriram novas possibilidades de entendimento do processo saúde-doença na população e do acesso aos serviços de saúde.

Os SIG são as ferramentas mais utilizadas para o acesso aos dados necessários para as análises espaciais. Na área de saúde, a informação que associa os casos ao território é o endereço do paciente. A forma de se obter os dados para alimentar os SIG é o georreferenciamento dos endereços, chamado de geocodificação. A Figura 1.5 apresenta um esquema de ligação entre os Sistemas de Informação Geográfica, as técnicas de análise espacial e os processos de análise espacial para investigação dos eventos de saúde. Este esquema mostra que o caminho para a inserção das informações destes eventos nos SIG passa pela geocodificação dos endereços contidos neles. O endereço residencial, além de ser a referência de localização das pessoas, é a informação de ligação entre os dados dos Sistemas de Informação em Saúde (SIS) ou outros registros de saúde e os sistemas de informação geográfica, utilizados nas análises e estatísticas espaciais.

O endereço, por ser uma informação textual, precisa de um tratamento para ter condições de ser utilizado em um sistema computacional, incorporado a um SIG e servir de componente de uma estatística espacial para as análises necessárias. Este tratamento é composto por normalização (tratar abreviaturas, espaços, caracteres especiais), separação (em tipo de logradouro, título, nome, número, complemento, além de bairro, cidade e outras referências) e padronização (igualar ao formato do banco de dados base). Após o tratamento, o endereço fica pronto para a comparação com as bases de dados de endereços disponíveis.

As grandes fontes de informação sobre condições de vida e saúde das populações são as pesquisas do IBGE – Censos Demográficos, Contagens de População e Assistência Médica (AMS) –, os dados dos sistemas de informações de mortalidade (SIM), de nascimento (SINASC), de internações hospitalares (SIH), de notificações de doenças (SINAN), de atenção básica (SIAB) entre outros, e os sistemas de cadastro de

unidades assistenciais (CNES) (Datasus, 2008). Exceto as informações completas do censo, que podem ser estimadas para regiões muito pequenas, os setores censitários, todas as demais em geral têm como unidade territorial de referência o município ou a unidade de saúde de atendimento, sem qualquer referência geográfica. Mesmo o SIAB, que define como alvo a atenção à família, entendida a partir do ambiente e espaço geográfico em que vive, não possibilita a análise a partir do território e do domicílio, mas somente consolidados por agente de saúde (Portugal, 2003).

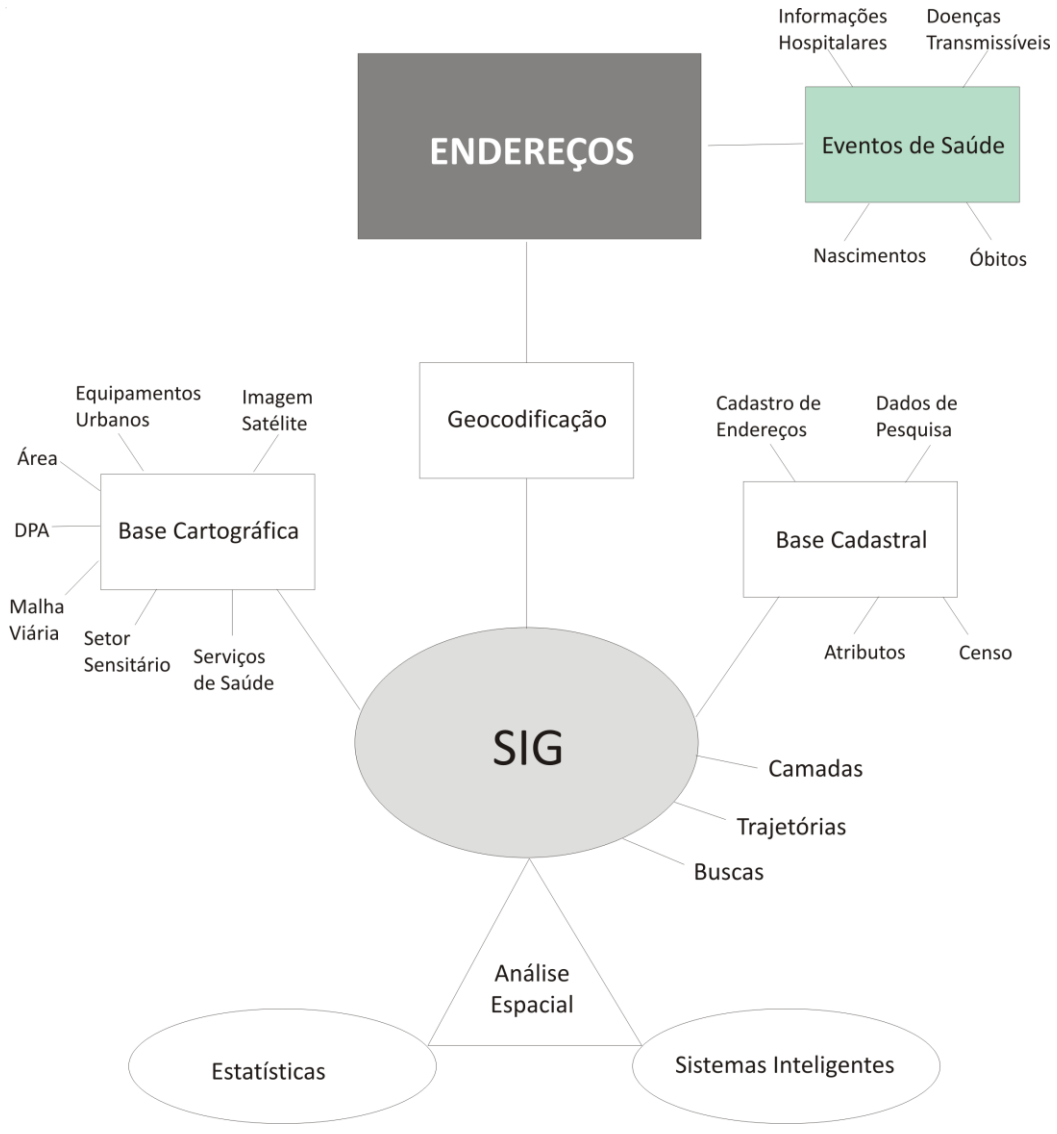


Figura 1.5. Importância dos endereços na incorporação dos eventos de saúde ao SIG

Nesse sentido, a primeira questão metodológica importante é o recorte territorial intramunicipal a ser adotado e, conseqüentemente, a forma de georreferenciamento. Uma opção é dividir o território em áreas. Essa é a forma utilizada pelo IBGE no censo demográfico, cuja delimitação territorial pode ser adquirida, juntamente com as informações do Censo 2000 (IBGE, 2002). O setor censitário, entretanto, ainda que proposto pelo Comitê Temático Interdisciplinar sobre Geoprocessamento e Dados Espaciais em Saúde (CTI-Geo) da Rede Interagencial de Informações para a Saúde - RIPSA como área mínima para georreferenciamento das informações de saúde, é apenas uma definição operacional do censo (RIPSA, 1997). Assim, a unidade territorial adotada pode ser um agregado destes setores, segundo parâmetros pré estabelecidos, dependendo do modelo utilizado. A construção de SIG utilizando unidade geográfica intra-municipal deve ser planejada cuidadosamente, considerando custos, recursos materiais e humanos, com as vantagens de cada possível modelo (Carvalho et al., 2000).

A principal característica do desenho de áreas na análise de dados em saúde é permitir relacionar informações socioeconômicas, demográficas e de saúde. Ressalta-se, portanto, a importância de fazer coincidir os limites dessas áreas com uma agregação de setores censitários, que permita a utilização das informações geradas no censo. Além disso, sempre que trabalhando com áreas, a informação é proveniente de contagens: de óbitos, de nascimentos, de chefes de família, etc. Os indicadores estimados serão médias, taxas ou proporções, tão mais úteis quanto menor a dispersão ou a mistura de elementos. Resumindo, é fundamental que as áreas utilizadas como unidade de análise sejam razoavelmente homogêneas quanto ao perfil da população residente. Por razoável entende-se aqui que no mesmo bairro utilizado como unidade de análise não estejam incluídas nos indicadores áreas de residência de classes médias e favelas, por exemplo, o que dificulta a identificação da população alvo para as políticas de saúde (Barcellos & Santos, 1997). Entretanto, quanto maior a homogeneidade interna da área, menor ela será, e conseqüentemente os indicadores sofrerão o efeito de pequenos números, flutuando bastante. É comum que a área com os piores e com os melhores indicadores, como, por exemplo, o de mortalidade, sejam áreas onde a população é tão pequena que um óbito a mais aumenta muito a taxa estimada, ou a ausência em determinado ano de qualquer evento gere indicador igual a zero.

1.1. RELEVÂNCIA DA TESE

No Brasil há grandes diferenças no desenvolvimento de bases cartográficas e de cadastros para apoio aos procedimentos de geocodificação (Barcellos & Ramalho, 2002). Este fato se reflete nos níveis de acesso às informações georreferenciadas. Este problema tem impacto substantivo em diversas questões, que vão de estudos e pesquisas onde o mote é a análise geográfica, às atividades de vigilância epidemiológica, sendo particularmente relevante na perspectiva da construção de uma vigilância em saúde de base territorial, integrando a ocorrência dos agravos registradas nos diferentes SIS com aspectos ambientais relevantes.

A construção de um procedimento padrão de georreferenciamento de endereços é o passo inicial e indispensável para viabilizar a ligação dos elementos da ampla gama de estudos que utilização este tipo de informação. Este procedimento deve ser desenvolvido a partir de uma base de dados pública com um conjunto de elementos básicos, permitindo acrescentar novos elementos disponíveis em cada setor.

1.2. OBJETIVO DA TESE

O objetivo desta tese é propor um método que permita a associação dos endereços dos eventos de saúde pública a uma coordenada ou área pré-definida, utilizando como base os cadastros do Censo 2000 do IBGE, com a finalidade de facilitar análises espaciais destas informações. São discutidos cinco estudos de caso que permitiram a investigação dos procedimentos envolvidos no processo de geocodificação, colocando em evidência as características específicas e dificuldades encontradas neste processo.

1.3. ORGANIZAÇÃO DA TESE

Os resultados das pesquisas bibliográficas utilizados para o desenvolvimento desta tese estão nos capítulos 2 e 3. O capítulo 2 apresenta os fundamentos ligados a espaço, contexto e análise espacial. Começa com o espaço e contexto na Saúde Pública, seguido pelos conceitos de vizinhança, escala e flutuação aleatória, além dos Sistemas de Informações Geográficas (SIG). Ao final, discute os endereços e o modo de apresentação destes pelos informantes. No capítulo 3, são discutidas as metodologias utilizadas na geocodificação, iniciando por seu conceito e uma revisão de sua utilização da Saúde Pública. A seguir são apresentadas as estruturas de bases de dados utilizadas no processo de geocodificação e as rotinas de tratamento dos campos na preparação e comparação de textos. No final do capítulo são discutidos os métodos de tratamento dos textos para a entrada de dados.

O capítulo 4 se refere à proposta de geocodificação que trata esta tese. Nele estão definidas as estruturas do banco de dados e dos arquivos gráficos associados, além dos processos para tratamento dos dados de entrada, de comparação destes com os dados da base de referência e as formas de saída.

Um estudo de casos, com quatro experiências, está no capítulo 5. Estes estudos serviram como fonte de informações para o conhecimento dos problemas encontrados nos processos de geocodificação e como subsídios para a formação da proposta do capítulo 4.

No capítulo 6 são apresentados os comentários finais.

Há 2 anexos neste trabalho:

- Tabelas de Tipos e Títulos de logradouros;
- Artigo: Geoprocessamento dos dados da saúde: o tratamento dos endereços.

2. BASES CONCEITUAIS

"A Epidemiologia é freqüentemente definida em termos do estudo da determinação da distribuição da doença; mas não se deve esquecer que quanto mais espalhada é uma causa particular, menos ela contribui para explicar a distribuição da doença." (Geoffrey Rose)

Neste capítulo são apresentadas as bases teóricas dos conceitos e métodos utilizados nesta tese, relacionadas às necessidades para os estudos que envolvem a análise espacial em Epidemiologia. Ele se inicia com os fundamentos de espaço e contexto em saúde pública (item 2.1), objetivo principal da produção das informações que são objeto desta tese. Para melhor compreensão, esta seção foi dividida em cinco itens. No primeiro são discutidas as possíveis unidades de análise, enquanto o segundo apresenta as relações entre estas unidades. Os conceitos de escala e a flutuação aleatória, importantes na determinação dos parâmetros dos estudos, compõem o terceiro item. O quarto item é composto pelos estudos ecológicos. No último são apresentados os métodos de análise espacial.

Seguindo as ferramentas necessárias para a execução destas análises, são apresentados os Sistemas de Informações Geográficas (item 2.2). A ligação entre os SIG e os registros de eventos de saúde está relacionada com os endereços residenciais dos personagens destes eventos. Neste sentido, o conhecimento dos tipos de endereços e do modo como são informados é necessário para estabelecer uma solução de referenciá-los nos mapas disponíveis (item 2.3).

2.1. ESPAÇO E CONTEXTO NA SAÚDE PÚBLICA

Milton Santos conceituou o espaço como *"um conjunto indissociável de sistemas de objetos e sistemas de ações, [...] um conjunto de fixos e fluxos que se interagem"* (Santos, 1996). Neste contexto, nos estudos a respeito de ocorrência de doenças nas populações e o acesso destas populações aos serviços de saúde, buscou-se estudar sua distribuição como resultado da organização social do espaço, considerando *"o espaço onde se realizam processos econômicos e sociais"* (Sabroza & Leal, 1992).

Tempo, lugar e pessoa compõem a tríade básica da produção/interpretação dos constructos epidemiológicos, dizem os manuais que fundaram as bases metodológicas da disciplina. O que, na verdade, poderia ser escrito como pessoas em lugares/tempos. É a distribuição de ocorrências que define o escopo da epidemiologia, já propõem os textos mais recentes. De qualquer forma está ali, inexorável, o espaço. (Czeresnia, 2000). Epidemiologia é o estudo de saúde e doença em populações humanas e, como as populações estão inexoravelmente localizadas no espaço, parece razoável que a análise espacial dos eventos e os Sistemas de Informação Geográfica (SIG) sejam importantes para o avanço da epidemiologia como ciência (Jaquez, 2000).

Define-se *"análise estatística espacial quando os dados são espacialmente localizados e se considera explicitamente a possível importância de seu arranjo espacial na análise ou interpretação dos resultados"* (Bailey & Gatrell, 1995). A ênfase da análise espacial é medir propriedades e relacionamentos, levando em conta a localização espacial do fenômeno em estudos de forma explícita. Ou seja, a idéia central é incorporar o espaço à análise que se deseja fazer (Câmara & Carvalho, 2004).

O emprego dos métodos de análise espacial é aplicável de forma efetiva em algumas situações tais como: quando a geração do evento estudado for ocasionada por fatores ambientais de difícil determinação pelas variáveis do indivíduo; quando os fatores relacionados com o evento de estudo possuirem distribuição espacial; ou no estudo de trajetórias entre localidades (Carvalho & Cruz, 1998). Neste sentido, com a localização mais exata possível do evento, a determinação de sua vizinhança e da escala utilizada no estudo é fundamental para a utilização dos métodos de análise espacial.

A organização espacial dos indivíduos impõe uma lógica de localização e funcionamento da sociedade. Esta lógica é influenciada por fatores econômicos, culturais e sociais que atuam no espaço. As relações que envolvem este processo não são tão evidentes no espaço de moradia, circulação e consumo. Como o espaço urbano viabiliza a circulação de agentes de doenças e estabelece elos de ligação entre grupos populacionais com características sociais e as fontes de riscos, cabe à investigação epidemiológica e à Geografia da Saúde estabelecer estes elos (Barcellos & Sabroza, 2000).

A seguir são apresentadas noções sobre as unidades de análise utilizadas nos estudos epidemiológicos, suas relações e utilização em estudos ecológicos e análise espacial.

2.1.1. Unidades de Análise

A escolha de uma unidade de análise em um estudo que utiliza técnicas de análise espacial depende de vários fatores, além dos objetivos e o desenho do estudo, entre eles:

- Os elementos disponíveis nas bases cartográficas, contidas no sistema de informações geográficas utilizado, da área de estudo;
- As informações coletadas para o estudo.

A representação cartográfica dos locais de ocorrência de eventos está inserida nos Sistemas de Informação Geográfica (SIG) e obedece aos objetos contidos nestes sistemas (item 2.2). Neste sentido, os eventos são representados de acordo com a representação geométrica do SIG utilizado, podendo ser uma representação vetorial ou matricial. No caso de vetorial, o evento pode ser representado por três formas básicas (Figrua 2.1):

- Ponto – par de coordenadas que localizam o evento;
- Linha – um conjunto de pares de coordenadas, formando uma linha poligonal aberta. Exemplo: trecho de logradouro.

- Área – um conjunto de pares de coordenadas, formando uma linha poligonal fechada. Exemplo: bairro ou setor censitário.

A representação matricial (Figura 2.2) é resultante do tratamento do espaço como uma superfície plana, onde cada célula está associada a uma porção do terreno. Cada célula é representada por um retângulo, também chamado de pixel. Todos os retângulos possuem as mesmas dimensões. Os atributos são associados a cada célula.

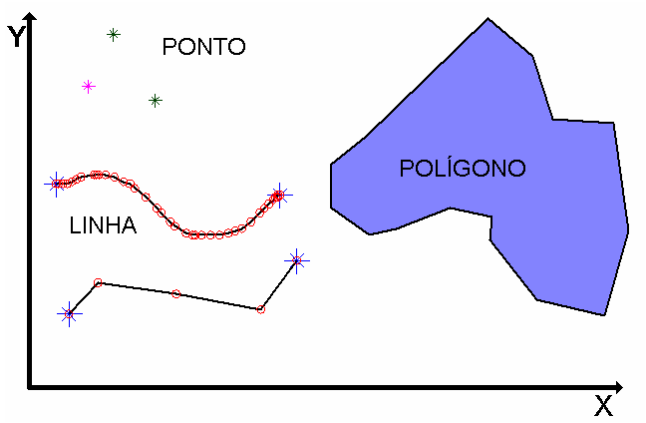


Figura 2.1. Elementos de representação vetorial (Fonte: INPE)

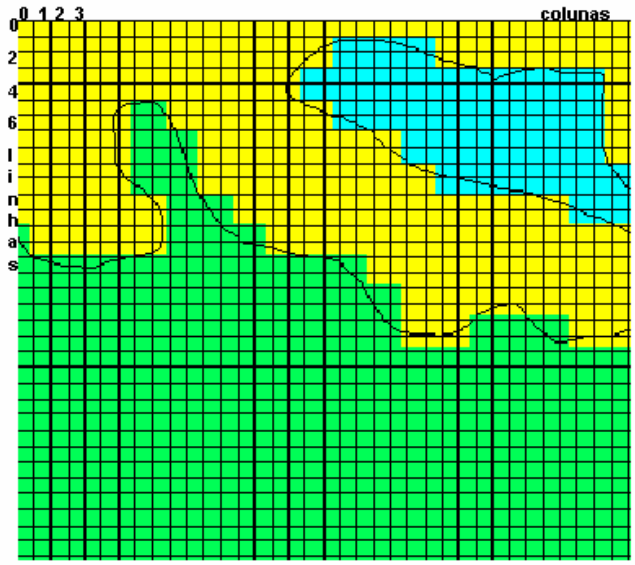


Figura 2.2. Representação matricial de um mapa (Fonte: INPE)

A unidade de maior detalhamento possível é o ponto, mas é o mais difícil de se obter. Outras unidades intra-municipais utilizadas são os bairros e as áreas de censo (setores censitários, census tracts, etc.). Como os setores censitários representam as menores áreas com dados sócio-econômicos disponíveis, é conveniente a utilização de agregados destes setores com características homogêneas, segundo os parâmetros de interesse do projeto (Santos, 2008). Para cada tipo de unidade escolhida existem particularidades quanto às relações entre as unidades (vizinhança, proximidade, interação) e as escalas adequadas aos estudos.

2.1.2. Relações entre as unidades de análise

As relações entre as unidades de análise podem se dar por proximidade (vizinhança, distância) ou por relacionamento. Quando a unidade utilizada é representada por ponto ou linha, as relações são medidas pela distância entre os eventos (buffer). No entanto, quando a área é o tipo de unidade escolhida para estabelecer as relações dos indivíduos no espaço, o conceito de vizinhança (*neighborhood*), ou características de local de moradia, contribui para a determinação dos modelos utilizados na investigação de problemas na área de saúde pública.

Uma definição clássica de vizinhança é apresentada por Keller (1968, *apud* Santos, 2008):

"*áreas distintas nas quais grandes unidades espaciais podem ser subdivididas* (...). *A distinção dessas áreas baseia-se em* (...) *limites geográficos, ou características étnicas ou culturais dos seus habitantes, ou sensação compartilhada de pertencimento pela qual as pessoas se sentem psicologicamente unidas, ou pelo uso concentrado de serviços para compras, lazer, e aprendizado*".

Pickett e Pearl (2001) reforçam a importância dos estudos dos efeitos de vizinhança:

"A evidência de modestos efeitos de vizinhança na saúde é razoavelmente consistente, apesar da heterogeneidade dos desenhos dos estudos [...] e prováveis erros de medida. Ao chamar a atenção da saúde pública para os riscos associados com a

estrutura social e ecológica de vizinhança, ensejam-se possíveis intervenções inovadoras no nível da comunidade."

A relação entre as estruturas espaciais varia conforme o objeto de estudo, enquanto a existência de limites entre as unidades define o tipo de associação. Neste sentido, esta relação pode ser classificada de acordo com os objetivos e o desenho do estudo, do seguinte modo:

- Por distância (buffer) – esta relação é definida pelo conjunto de elementos que estão contidos em uma área definida por uma distância pré-determinada, ou pelo conjunto de elementos mais próximos do evento estudado. Neste caso, a unidade de estudo pode ser um dos três tipos normalmente utilizados: ponto, linha ou área. Exemplo: seleção de residências que se encontram a uma distância de 50m de uma rodovia.

- Por contiguidade (vizinhança) – esta relação se dá quando os elementos envolvidos possuem uma contiguidade física, ou seja, quando as unidades compartilham um lado comum. Neste caso, a unidade utilizada deve ser uma área e esta contiguidade é chamada de primeira ordem. Ampliando este conceito, esta vizinhança pode ser de segunda ordem, quando são selecionados os elementos contíguos aos vizinhos de primeira ordem. Exemplo: municípios que possuem limites comuns.

- Por relacionamento (rede) – os elementos são definidos como vizinhos quando existe alguma interação entre eles. Para o cálculo de distâncias na montagem das redes de relacionamento, o tipo de unidade utilizada é o ponto (ou pixel). No entanto, no caso de utilização de área, é escolhido o centróide ou outro ponto significativo, interno à área. Exemplo: seleção de municípios que utilizam os serviços especializados de saúde, como a cirurgia cardíaca de um determinado município.

Nestes conceitos de relação entre os espaços em estudos epidemiológicos, é utilizada a Matriz de Proximidade Espacial, que é uma matriz quadrada, com dimensão

(linhas e colunas) igual ao número de unidades espaciais de observação. Para determinar a relação entre as unidades, podem ser usados os critérios descritos acima.

O esquema da figura 2.3 apresenta um conjunto hipotético de municípios e a respectiva matriz de ponderação espacial construída pelo critério de contiguidade de primeira ordem, com a vizinhança, entre um município e outro, determinada pela existência de um limite em comum. Nesses casos se atribui o valor 1 aos respectivos elementos w_{ij} , enquanto que os demais elementos da matriz (não vizinhança) são iguais a zero. A diagonal (elementos w_{ii}) é igual a zero por convenção, segundo qualquer critério. Os valores atribuídos aos elementos podem conter pesos, dependendo de atributos associados ao relacionamento entre as unidades.

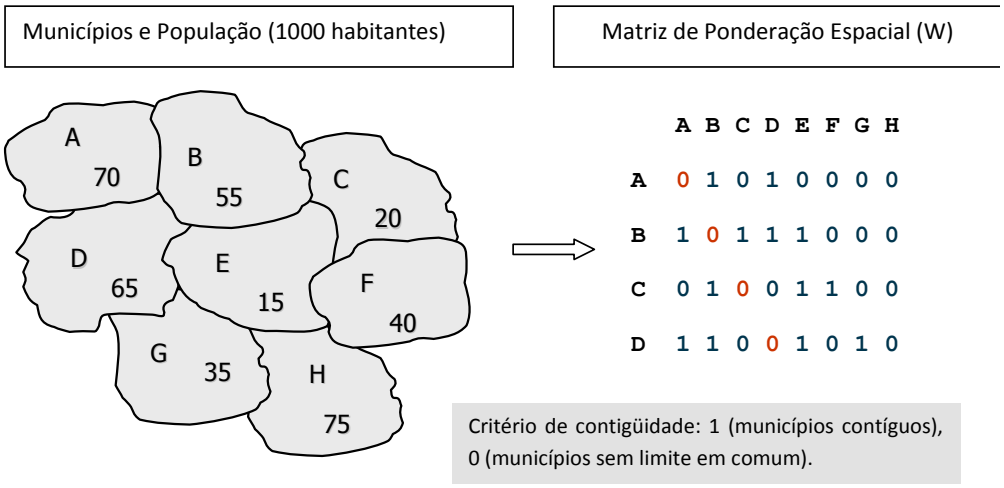


Figura 2.3. Esquema didático da construção da Matriz de Proximidade Espacial pelo critério de contigüidade

Nos espaços ocupados existem fluxos conectando pessoas e lugares, formando as redes geográficas. Na geografia, redes são estruturas de interconexão, constituídas por dois tipos de elementos: pontos (ou nós) e as ligações entre eles. A configuração das ligações revela a estrutura da rede. Nas redes territoriais, os lugares correspondem aos pontos, e as ligações podem ser materiais, como as estradas, ou imateriais, como os sinais eletromagnéticos. O estudo das redes perpassa vários campos do conhecimento, e adquiriu particular relevância nos últimos anos (Oliveira et al., 2004).

2.1.3. Escala

Um padrão espacial bem definido em determinada escala pode ser apenas um conjunto de variações aleatórias em outra. A decisão sobre a escala geográfica de observação, aliada à escolha da unidade espacial de análise, faz parte de um conjunto de questões de ordem prática que estão na base de qualquer análise espacial (Bailey & Gatrell, 1995). Esta decisão depende, claramente, do fenômeno em estudo, dos objetivos da análise, da escala cartográfica na qual os dados foram coletados e, na maioria das vezes, do julgamento e da experiência do analista.

Um ponto não possui dimensões, logo não há escala a definir para uma unidade pontual. Entretanto, dependendo do objetivo e do desenho do estudo, existem dimensões para a área de abrangência ou distâncias de alcance na definição dos eventos relacionados por distância ou relacionamento e, portanto, aqui também, a definição de uma escala de representação é útil. Outro aspecto a considerar para as unidades representadas por pontos ou pixels é a precisão. No caso do ponto, a precisão é medida pela distância entre a localização real do evento e a da representação no mapa, enquanto, para o pixel (representação matricial), representado por um retângulo, a precisão é medida pelo tamanho do pixel que determina a resolução da imagem.

Em uma análise de áreas, os dados utilizados são obtidos, normalmente, em levantamentos populacionais (censos, estatísticas de saúde). A delimitação destas áreas segue frequentemente critérios de limites políticos (bairros, municípios), operacionais (áreas de saúde) ou resultado de interpolação entre pontos amostrais por ferramentas de SIG (mapas isopléticos). Entretanto os estudos pressupõem haver homogeneidade interna nestas áreas, contendo agrupamentos aleatórios de indivíduos semelhantes em relação a outras áreas. Embora tal fato possa ocorrer no campo sócio econômico, demográfico e de variáveis de saúde (Wrigley at al., 1996), na prática, não há garantias da distribuição do evento estudado ser homogêneo dentro destas unidades (Lemos-Dias et al., 2002). No Brasil, com seus contrastes sociais, é comum encontrar grupos sociais distintos agrupados em uma mesma unidade de área como, por exemplo, favelas e áreas de alta renda. Estes agrupamentos apresentam indicadores que representam médias entre populações heterogêneas. Em outras regiões, encontram-se diferenças importantes de

população ou áreas nas unidades amostrais, resultando em distorções nos indicadores obtidos nos cálculos das taxas populacionais. Em áreas com pequenas populações ocorre o problema inverso, obtendo-se maior homogeneidade, mas com taxas variando muito para cada evento encontrado (Carvalho & Cruz, 1998).

A escolha da escala utilizada deve seguir uma avaliação do objeto do estudo. Com escalas maiores obtém-se maior homogeneidade interna com maior flutuação aleatória, enquanto com escalas menores a tendência é de existir maior heterogeneidade, com menor flutuação e a as médias mais próximas da média global. Quanto mais desagregado o dado original, maior será a flexibilidade de se desenhar os modelos por meio de agregação destas áreas. Considerando estes aspectos, uma alternativa é utilizar técnicas de agregação de micro-áreas, com dados na maior escala possível (ex. setor censitário), a fim de obter regiões maiores, preservando o fenômeno estudado do melhor modo possível (Santos et al., 2001). Deste modo, deve-se procurar utilizar a maior escala de levantamento de dados disponível e utilizar técnicas que permitam tratar a flutuação aleatória, sempre buscando critérios de agregação dos dados que sejam consistentes com os objetivos do estudo.

Quadro 2.1. Escala, resolução, homogeneidade e estabilidade e sua relação com a área da unidade de estudo (Carvalho & Cruz, 1998)

Termo	Definição	Variação
Unidade de análise	Menor área para a análise de informações	↘
Escala	Razão entre as medidas no mapa e as distancias reais	↗
Resolução	Capacidade de distinguir pontos adjacentes	↗
Homogeneidade	Características da distribuição estatística	↗
Estabilidade	Presença de flutuação aleatória	↘

No Quadro 2.1 são definidas escala, resolução e homogeneidade, assim como suas relações com a unidade de estudo. Quando a área da unidade de análise diminui, a escala e a resolução aumentam. No entanto, a homogeneidade dos indexadores tende a aumentar e a estabilidade dos indicadores tende a diminuir.

2.1.4. Estudos Ecológicos

Estudos ecológicos avaliam associações entre incidência observada de doenças e potenciais fatores de risco, medidos em grupos populacionais, onde estes grupos são tipicamente definidos por áreas geográficas. A definição clássica de estudo ecológico é apresentada por Morgenstern (1998):

"*Um estudo ecológico ou agregado focaliza a comparação de grupos, ao invés de indivíduos. A razão subjacente para este foco é que dados a nível individual da distribuição conjunta de duas (ou talvez todas) variáveis estão faltando internamente nos grupos; neste sentido um estudo ecológico é um desenho incompleto*".

Nestes estudos somente é possível inferir na escala populacional. Alguns destes estudos procuram estabelecer relações de causa-efeito entre diferentes medidas, como o uso de modelos de regressão; um exemplo clássico é correlacionar anos de estudo do chefe de família e sua renda, que usualmente apresenta forte correlação. Note-se, no entanto, que devido aos efeitos de escala e de agregação de áreas, os coeficientes de correlação podem ser inteiramente diferentes no indivíduo e nas áreas. Este fenômeno, nas ciências sociais e na epidemiologia, é chamado de "falácia ecológica", segundo Schwartz (1994):

"*a falácia ecológica, conforme freqüentemente usada, encoraja três noções interrelacionadas e falaciosas: (1) que modelos em nível individual são mais perfeitamente especificados que os de nível ecológico, (2) que correlações ecológicas são sempre substitutos para correlações de nível individual, e (3) que variáveis de nível de grupo não causam doença.*".

Entretanto, resgatando o conceito da ecologia definido como *"o estudo das complexas inter-relações entre organismos vivos e o seu meio físico"* (Wikipedia). Neste sentido, Rose (2001) evidencia a importância dos estudos ecológicos e a contribuição do contexto na avaliação das causas das doenças, argumentando que sendo a Epidemiologia definida como o estudo da determinação da distribuição da doença, é necessário ter em mente que quanto mais espalhada é uma causa particular desta, menos ela contribui para explicar sua distribuição. Sendo assim, devem-se buscar também nas

diferenças entre as populações ou em mudanças das populações ao longo do tempo, as causas das incidências das doenças. Para Susser (1994), deve-se entender como o contexto afeta a saúde das pessoas através de seleção, distribuição, interação e adaptação. De fato, apenas as medidas de atributos do indivíduo podem não dar conta das explicações dos processos estudados.

Neste contexto, um bom exemplo de estudo ecológico é apresentado por Barcellos et al. (2003) que faz um estudo da distribuição da leptospirose no Rio Grande do Sul comparando com aspectos da ecologia, como as altitudes, as bacias hidrográficas e o uso do solo. Na Figura 2.4 são apresentadas as superposições destes aspectos ambientais com a incidência da doença.

Uma das questões básicas com os estudos ecológicos é que, para uma mesma população estudada, a definição espacial das fronteiras das áreas afeta os resultados obtidos. As estimativas obtidas dentro de um sistema de unidades de área são função das diversas maneiras que estas unidades podem ser agrupadas. Resultados diferentes podem ser obtidos simplesmente alterando as fronteiras destas zonas. Este problema é conhecido como "problema da unidade de área modificável" (*modifiable área unit problem* – MAUP) (Openshaw 1987).

As medidas utilizadas nos estudos ecológicos podem ser divididas nos grupos (Waller & Gotway 2004):

- Medidas agregadas - sumários de distribuição de observações colhidas a nível individual, usualmente proporções, médias, ou percentis da distribuição. (Ex: renda média do chefe da família; % de chefes com renda abaixo de um salário mínimo).

- Medidas ambientais - características físicas do meio onde vivem ou trabalham os indivíduos. Observar que para cada medida ambiental existe um análogo no nível individual (medidas de exposição ou dose) que varia entre os indivíduos do grupo (Ex: poluição do ar, intensidade de UV).

- Medidas globais - não existe análogo individual (densidade populacional; existência de leis, acesso ao serviço de saúde, etc.).

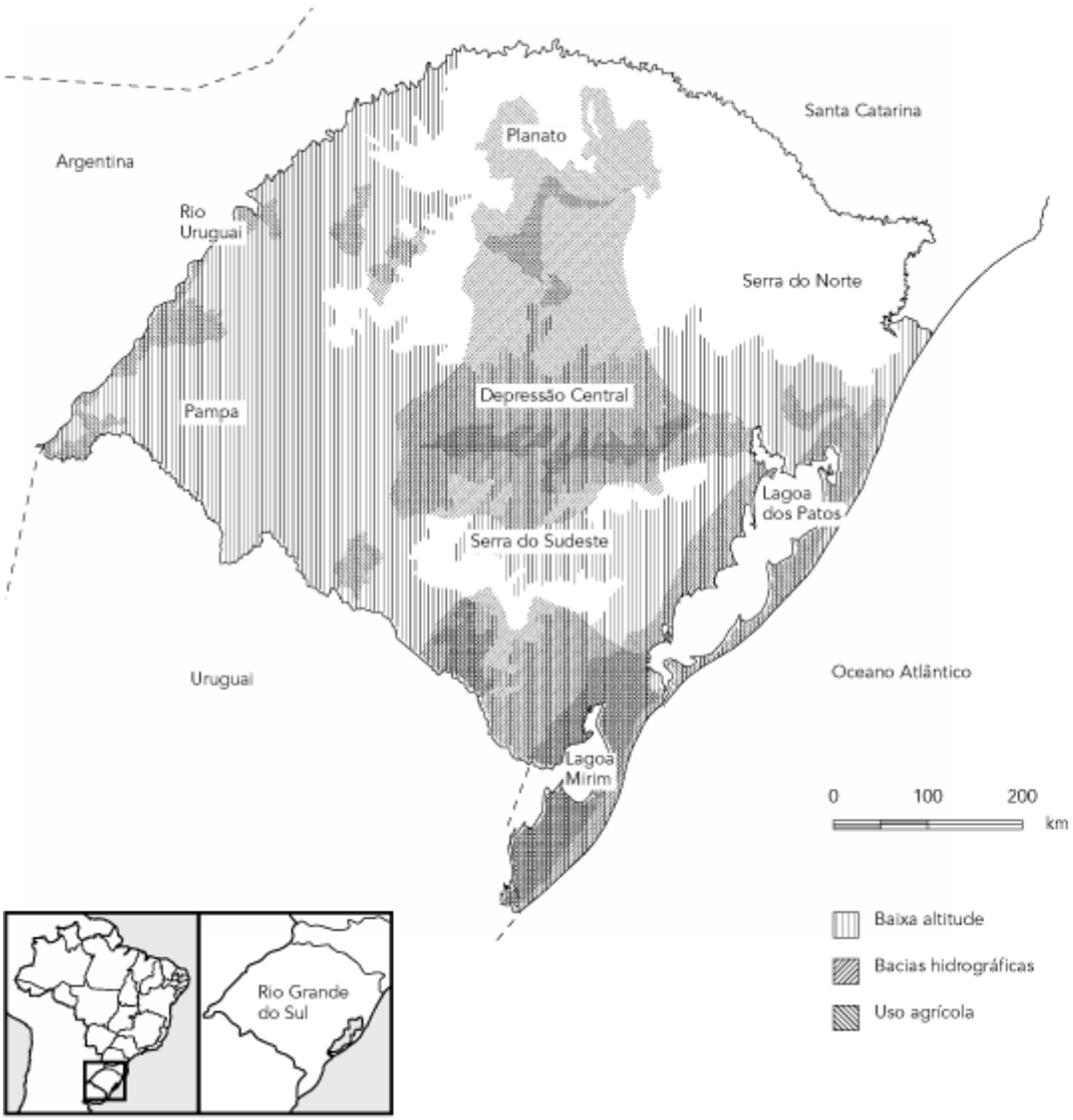


Figura 2.4. Sobreposição de áreas críticas de incidência de leptospirose segundo faixa de altitude, bacias hidrográficas e uso de solo (Barcellos et al. 2003)

2.1.5. Análise Espacial

Os problemas de análise espacial na saúde, em geral, abordam fatores de risco ambientais e aspectos socioeconômicos das populações analisadas. Em geral, o processo de modelagem é precedido de uma fase exploratória, associada à apresentação visual dos dados sob forma de gráficos e mapas e a identificação de padrões de dependência

espacial no fenômeno em estudo. Para estudá-los, utiliza-se um conjunto de procedimentos encadeados com a finalidade de escolher um modelo inferencial que considere explicitamente os relacionamentos espaciais presentes no fenômeno. Na área de saúde o mapeamento de dados georreferenciados levaram a vários achados, possibilitando, por exemplo, a identificação de diversos conglomerados no padrão de mortalidade em diversas doenças (Rushton et al., 2006). Os mapas também facilitam a visualização da associação espacial entre diversos fatores causais, que permitiram a criação de hipóteses etiológicas (Carvalho et al., 1996).

Compreender a distribuição espacial de dados originados de fenômenos ocorridos no espaço constitui um grande desafio para esclarecer questões centrais em diversas áreas do conhecimento, entre elas, a saúde. Além da percepção visual da distribuição espacial do problema, é muito útil encontrar padrões mensuráveis. Os epidemiologistas coletam dados sobre ocorrência de doenças, remetendo a algumas questões (Câmara & Carvalho, 2004):

- A distribuição dos casos de uma doença forma um padrão no espaço? Em que escala?

- Existe associação com alguma fonte de poluição?

- Variou no tempo?

- Há fatores socioeconômicos envolvidos?

Para obter respostas a estas perguntas, podem-se classificar os métodos utilizados em estudos em que existe o componente espacial em três grupos. No primeiro, estão os métodos voltados para a visualização dos dados espaciais. Outro grupo é dos métodos direcionados à investigação de padrões espaciais. O terceiro, concentra os métodos que se ocupam com a especificação de uma modelagem estatística e a estimação de medidas de associação. Em um estudo que envolva análise espacial, estes métodos não são excludentes, sendo comum haver uma interatividade entre os três grupos, com os dados sendo visualizados inicialmente e os aspectos de interesse explorados, gerando modelos. Os resultados da modelagem podem então ser

visualizados novamente, avaliados e, algumas vezes, darem origem a um refinamento dos modelos (Bailey & Gatrell, 1995).

Sob o ponto de vista da fonte da informação os estudos podem ser **agregados**, quando os dados para estudo são relativos a um grupo populacional ou **individuais**, com os dados no nível individual. Quando o enfoque é de representação da informação, podem ser divididos em padrões de **pontos** ou **áreas**.

Visualização de Dados Espaciais

O primeiro requisito para se analisar qualquer dado é a habilidade de "olhar" o dado a ser analisado (Bailey & Gatrell, 1995). O mapeamento dos dados é uma ferramenta fundamental para o pesquisador procurar padrões de distribuição destes dados, gerar hipóteses e avaliar o modelo proposto, ou considerar a validade ou não das predições derivadas deste modelo. Os modos de análise de dados foram alterados nas últimas décadas pela interatividade permitida pelos pacotes de aplicativos computacionais, o que tornou estas análises mais rápidas, simples e flexíveis. Os Sistemas de Informação Geográfica (item 2.2) possuem um ambiente para o mapeamento de dados, exploração dos padrões espaciais e relacionamentos com o objetivo de permitir executar estas funções de um modo rápido, simples e interativo.

Padrões de Pontos

Para a visualização de um padrão espacial de pontos, o caminho mais natural é dispor estes dados em um mapa de pontos. Isto vai dar uma visão geral da disposição destes pontos. Na Figura 2.5 é apresentado um mapa de pontos, apontando as coordenadas das residências de vítimas das principais causas externas de morte em Porto Alegre em 1996 (Santos et al., 2001). Geralmente é difícil tirar qualquer conclusão simplesmente com uma análise visual, como por exemplo, confiar em ideias intuitivas sobre o que considerar padrão randômico. Entretanto, existem técnicas que melhoram nossa capacidade de análise visual, pois alguns fatores podem afetar os

valores das variáveis utilizadas, como a existência de outros atributos relativos ao problema neste mesmo local ou variação da população de cada área.

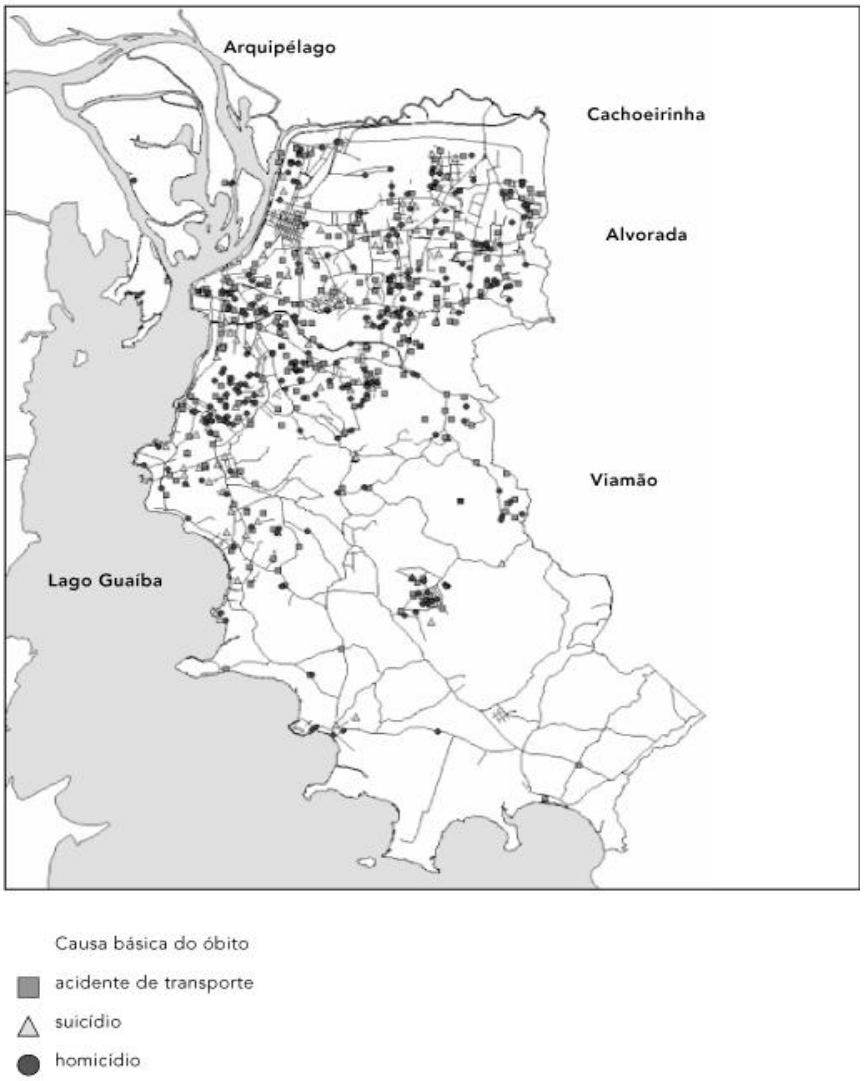


Figura 2.5. Exemplo de mapa de padrão de pontos (Santos et al., 2001)

Mapas Coropléticos

Os mapas coropléticos, um tipo de mapas temático, são a forma mais comum de visualização de dados espaciais. Nestes mapas, a unidade de observação é a área e cada área é colorida ou sombreada de acordo com uma escala discreta baseada no valor do atributo de interesse. Estes mapas são úteis na visualização do comportamento das variáveis escolhidas na região de estudo. O número de classes, a escala de visualização

e o intervalo de cada classe escolhidos podem determinar o tipo de resultado visualizado no mapa. A escolha das cores e símbolos de representação também pode ter relevância na apresentação de um mapa cloroplético. A Figura 2.6 mostra Minas Gerais subdividida nos 756 municípios existentes em 1994. De acordo com o valor da razão de mortalidade infantil padronizada (RMP), os municípios receberam diferentes cores e/ou tonalidades (Assunção et al., 1998). Alguns detalhes devem ser observados na análise destes mapas como, por exemplo, a população residente em cada unidade de área analisada. No Brasil é comum existirem municípios com grandes áreas e população pequena (características rurais) e outros com pequenas áreas e grande concentração de população (aglomerados urbanos, regiões metropolitanas), podendo influenciar na percepção visual dos fatores analisados. Nas áreas com pequenas populações, uma pequena variação absoluta de número de casos representa, muitas vezes, grandes variações nas taxas encontradas (Lemos-Dias et al., 1998).

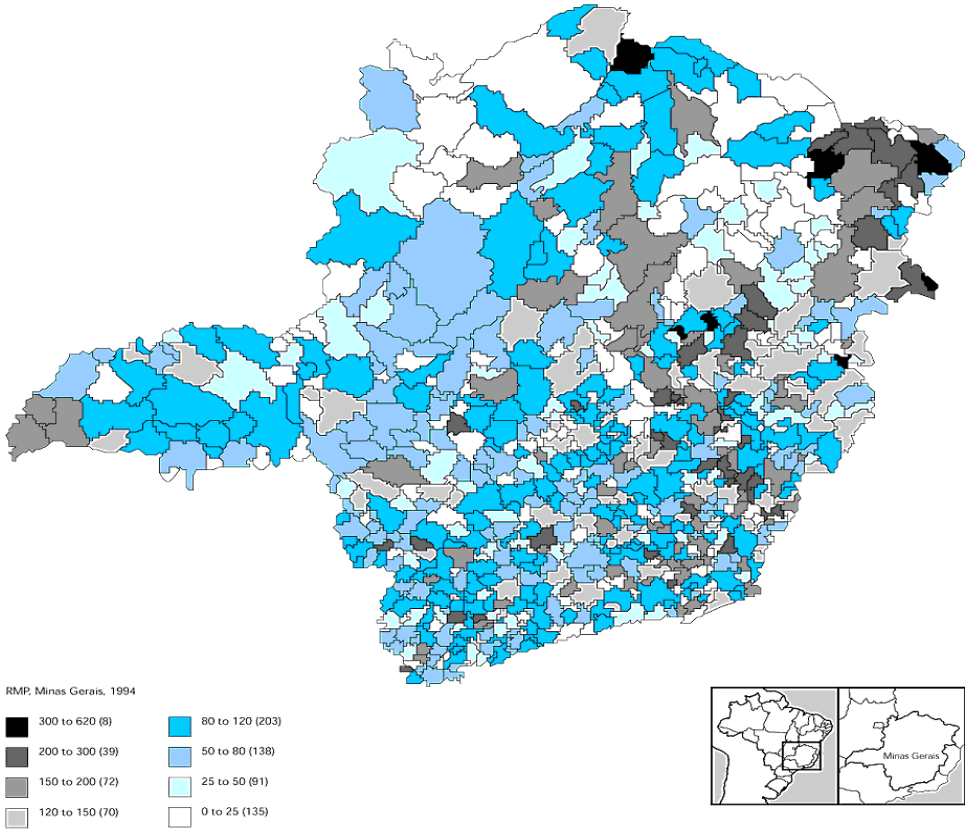


Figura 2.6. Exemplo de mapa cloroplético (Assumpção et al., 1998)

Mapas Animados

Outra questão interessante é a comparação de mapas. Supondo a distribuição espacial de um indicador em diferentes anos: como visualizar a evolução temporal? Certamente os pontos de corte da variável nos diferentes períodos devem ser os mesmos. Uma técnica empregada para descrever a evolução de uma doença no espaço e no tempo é a de utilização dos chamados mapas animados (MacEachren et al., 1998). Para a apresentação destes mapas, deve-se ter o cuidado de utilizar a mesma escala, mesmos intervalos de valores e mesmas cores em todos os mapas individuais da série. Na figura 2.7 é mostrado um desses mapas apresentado por Cruz (2004), ilustrando uma seqüência de mapas de mesmo padrão em três tempos diferentes.

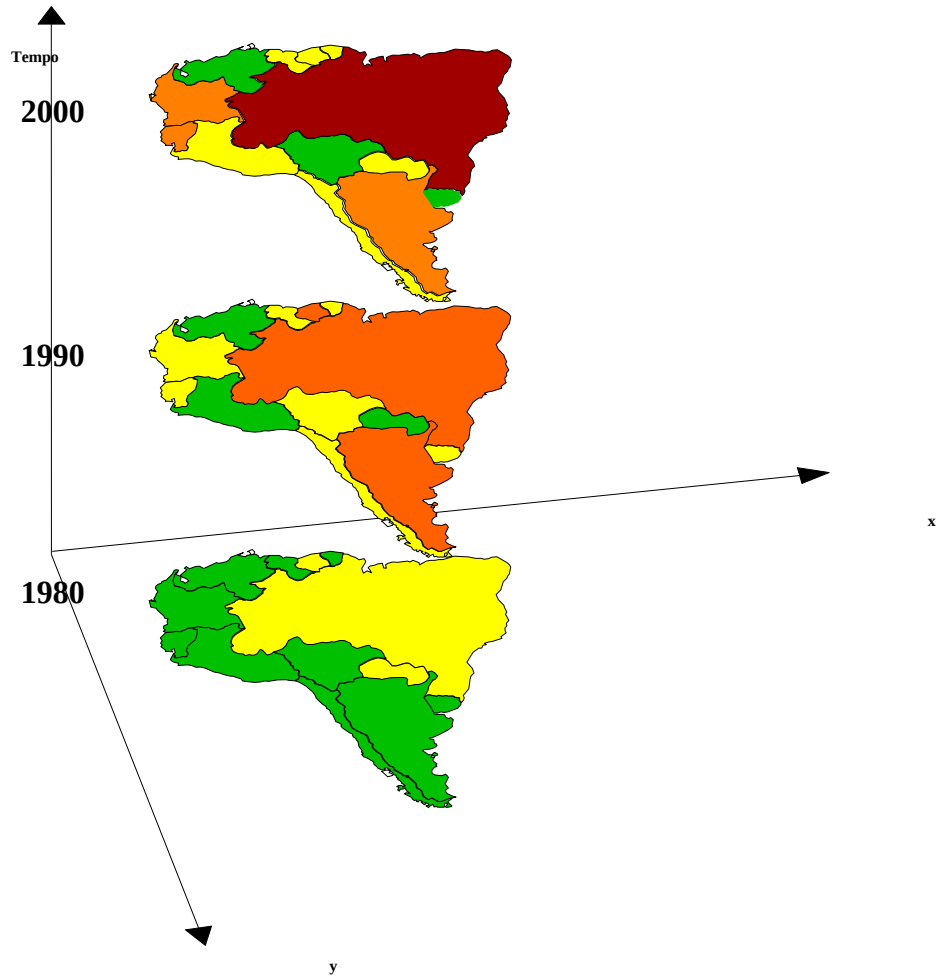


Figura 2.7. Exemplo de mapa animado (Cruz, 2004)

Suavização

Considerando as influências contextuais e as flutuações aleatórias que podem ocorrer em áreas com pouca população ou com doenças de pequena incidência, pode-se também supor que as taxas de diferentes regiões estão correlacionadas e o comportamento dos vizinhos influencia na estimação de uma taxa mais realista para cada área. Estas afirmativas sugerem o uso de técnicas de suavização. Bailey (2001) sintetiza algumas considerações importantes a respeito dos métodos de suavização. A suavização estatística consiste em um grupo de técnicas não paramétricas que permitem a filtragem da variabilidade de um conjunto de dados e que, ao mesmo tempo, retêm as características essenciais locais dos dados. Entre estas técnicas, destacam-se o método de intensidade (Kernel), Loess e o bayesiano empírico, apresentados nos itens abaixo. No contexto espacial, a suavização é uma técnica exploratória particularmente valiosa para a identificação de "áreas quentes".

A idéia básica dos "mapas de **Kernel**" é criar uma superfície contínua sobreposta aos pontos ou polígonos delimitados, formando uma grade regular recobrindo a região estudada. É uma técnica não paramétrica para a filtragem da variabilidade de um conjunto de dados, ao mesmo tempo em que retém as características locais principais dos dados. No caso de utilização de um mapa de pontos, é atribuído um par de coordenadas para cada localização de evento. Para as análises utilizando valores agregados, o mesmo resultado pode ser obtido pela utilização do centróide, centro populacional da área ou a sede do município, associados às taxas, contagens ou outro indicador. O grau de suavização é controlado através da escolha de um parâmetro conhecido como largura de banda (bandwidth), que deve ser definida visando refletir a escala geográfica do fenômeno estabelecido pela hipótese de interesse (Bailey & Gatrell, 1995). Na figura 2.8 é apresentado um esquema do método.

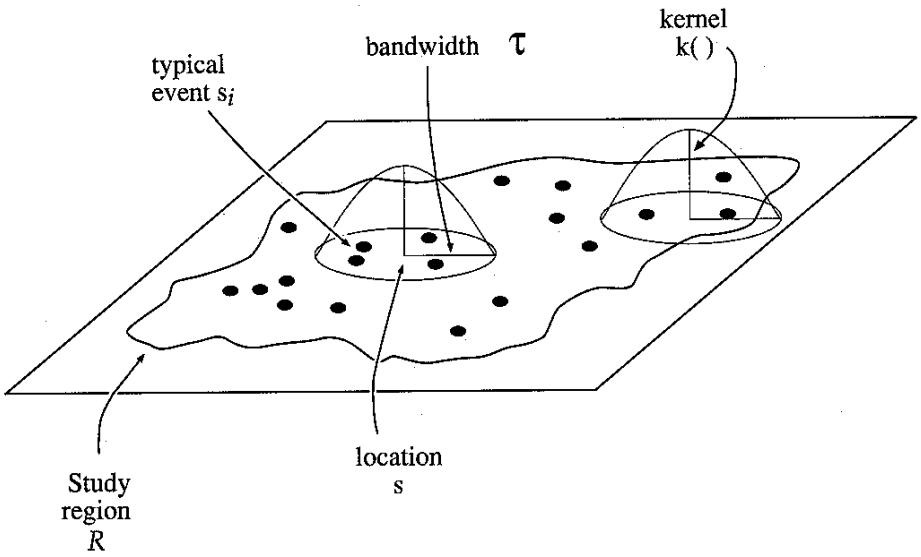


Figura 2.8. Esquema básico do método de kernel (Bailey & Gatrell, 1995)

$$\lambda_\tau(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} \cdot k\left(\frac{(s - s_i)}{\tau}\right) \qquad \text{(Eq 2.1)}$$

Onde

$\boldsymbol{k(\,)}$ - referido como "*kernel*", é a função de ponderação;

$\boldsymbol{\tau}$ - é a largura da banda (*bandwidth*), fator de alisamento;

$\boldsymbol{n}$ - número total de pontos;

$\boldsymbol{s}$ - centro da área;

$\boldsymbol{s_i}$ - local do ponto (*location*);

$\boldsymbol{\lambda(s)}$ - é o valor estimado.

Quando a unidade de análise é representada por polígono, é escolhido um ponto no interior de cada polígono e acrescido um fator (y_i) que corresponde ao número de eventos ocorridos nesta área, resultando na equação:

$$\lambda_\tau(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} \cdot k\left(\frac{(s - s_i)}{\tau}\right) \cdot y_i \qquad \text{(Eq 2.2)}$$

Outra alternativa para mostrar mapas sobre superfícies contínuas, muito similar ao Kernel na idéia, é o **Loess**. Um modelo relativamente simples, e que apresenta ajuste melhor nos extremos da série, é a regressão linear local ponderada. Neste caso, ao invés de se calcular a média em cada janela, como se faria em uma média móvel, estima-se, por mínimos quadrados, os parâmetros de um plano. O peso das observações diminui à medida que se afastam do ponto estimado, sendo então uma regressão local ponderada. Os pesos são atribuídos conforme uma função de decaimento que atua até uma distância pré-definida (Hastie & Tibshirani, 1990). Loess depende fortemente da escolha do fator de suavização, onde quanto maior o número de vizinhos, maior a suavização obtida.

Outra forma de suavização, para dados de área, é o método denominado **estimador bayesiano empírico**. Nesse caso, supõe-se que a taxa real, que de fato expressa a força de ocorrência do evento na área, é uma variável (Assunção, 2001). O melhor estimador dessa taxa é uma combinação linear entre a taxa observada (eventos/população) na área e um valor médio ponderados por um fator. Essa média usada na ponderação pode ser a taxa média da região toda de estudo, quando chamamos o método de bayesiano empírico global, ou pode ser a média dos vizinhos, método ao qual chamaremos de bayesiano empírico local. Regiões com populações muito baixas terão uma correção maior, com maior peso da média da vizinhança, e regiões populosas terão pouca alteração em suas taxas.

Considera-se o método empírico porque a média e a variância serão estimadas a partir dos dados, considerando que cada valor observado é apenas uma realização de um mesmo processo. Alguns cuidados são essenciais na hora de calcular o estimador bayesiano, particularmente o critério de matriz de vizinhança adotado (item 2.1.2). A Figura 2.9 apresenta o mapeamento das taxas de detecção de Hanseníase no período 1993-1997, mostrando as taxas brutas e as suavizadas utilizando o método bayesiano empírico. Esta suavização resolve o problema de haver um peso excessivo das áreas

maiores e menos densas e a grande ocorrência de áreas com taxa zero (indicados pelos círculos na figura).

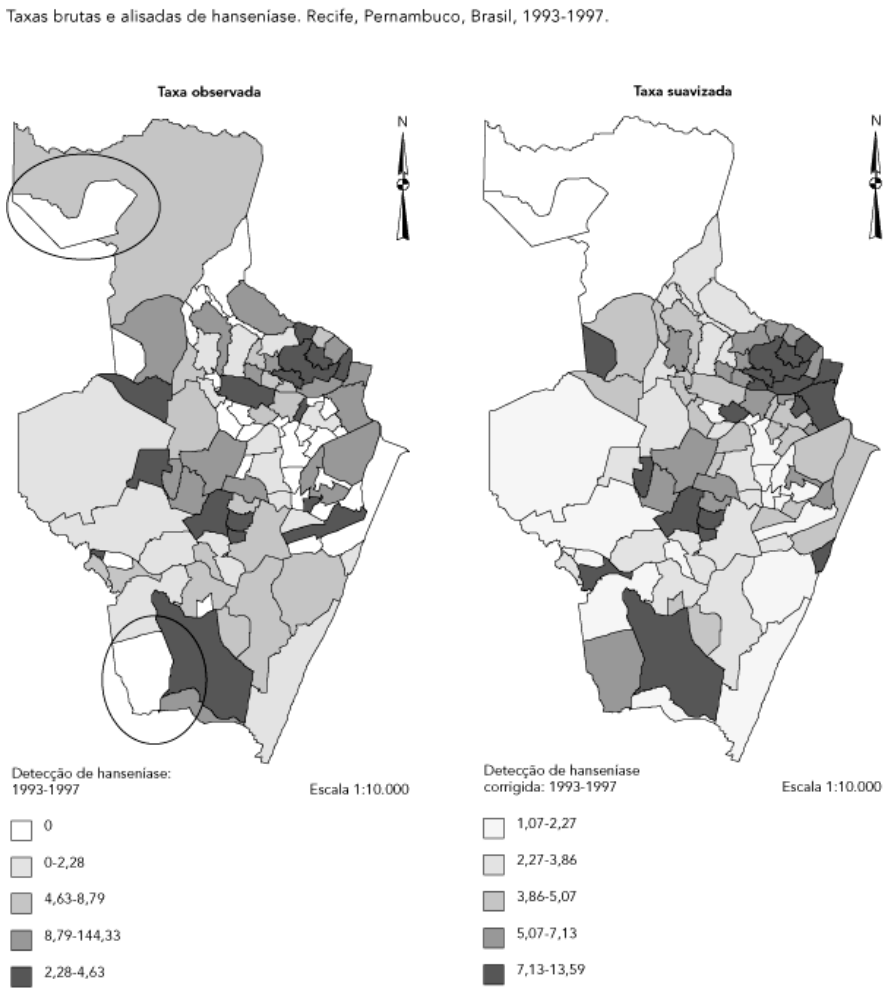


Figura 2.9. Exemplo de estimador bayesiano empírico (Souza et al., 2001)

Avaliação de padrão espacial

No caso de eventos pontuais, o modelo divide a região de estudo em subáreas e analisa a distribuição de eventos pontuais como um processo aleatório. Considera-se o número de eventos que ocorrem na subárea. Estas ocorrências são consideradas como não-correlacionadas e homogêneas, e estão associadas à mesma distribuição de probabilidade de Poisson. Numa visão intuitiva, pode-se considerar que a posição dos

eventos é independente e de que os eventos têm igual probabilidade de ocorrência em toda a região. Isto nos permite estabelecer uma base de comparação entre uma distribuição completamente aleatória (CSR – *complete spatial randomness*), que seria gerada por um processo de Poisson, e os dados coletados em campo (Assunção, 2001). Há um conjunto grande de testes de hipótese que avaliam o afastamento da CSR e que dependem do tipo de dado. Esta distribuição pressupõe as propriedades de estacionariedade e isotropia, ou seja, é invariante sob translação ou rotação, variando apenas com a distância (Figura 2.10). Para avaliar o padrão de pontos são utilizados vários métodos, entre eles os de Knox, Mantel ou Função K, para áreas, Moran e Geary.

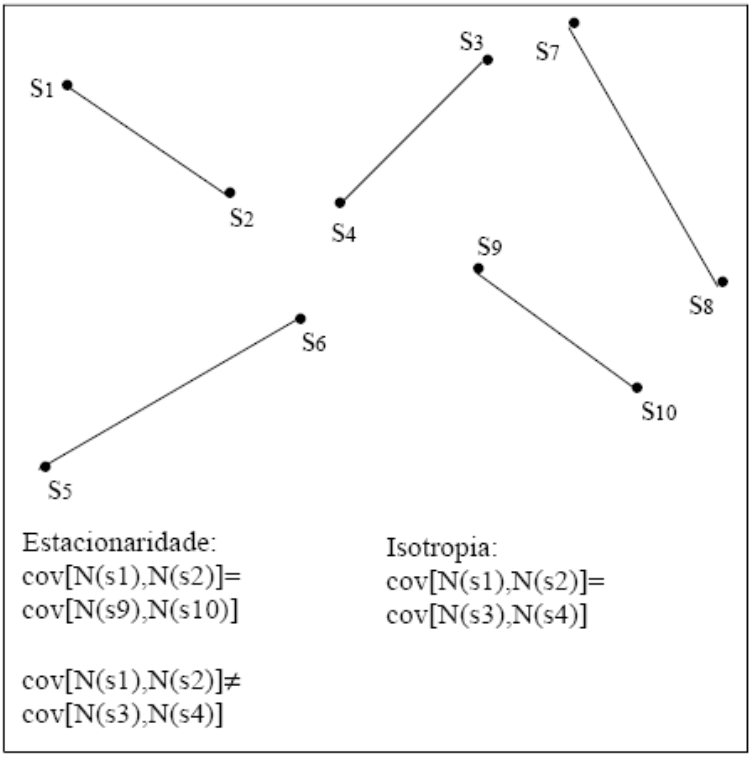


Figura 2.10. Ilustração de processos espaciais estacionários e isotrópicos (Bailey & Gattrel, 1995)

Modelos de Regressão

Do ponto de vista deste trabalho é importante ressaltar que nesse modelo pode-se utilizar informação georreferenciada em diferentes escalas. Os modelos estatísticos de regressão têm o objetivo de determinar um modelo matemático que descreve a

relação entre um desfecho (variável dependente ou resposta) e um conjunto de potenciais fatores de risco (variáveis independentes ou explicativas). Supondo, por exemplo, casos de soro conversão de leptospirose urbana (Assunção, 2001). A variável resposta é sorologia para leptospirose, podendo ser positiva ou negativa. Trata-se então de um modelo logístico do grupo dos GLM (modelos lineares generalizados), pois a resposta é binomial. As variáveis explicativas podem ser: idade, sexo e a posição da residência do indivíduo. Será que o evento varia com o local da residência? Para isso é necessário colocar as coordenadas no modelo de forma a respeitar a continuidade espacial, lembrando que não se tem amostra independente, pois "o mais perto se parece mais". A forma de introduzir é criar ma função não paramétrica a partir das coordenadas que pondera, como um kernel, a densidade espacial dos soropositivos comparados aos soronegativos. Obtem-se um modelo aditivo generalizado - GAM (Wood, 2004) (Figura 2.11).

Outro tipo de dado analisado sob forma de modelo de regressão são as contagens de eventos por unidade de área. Nesse caso temos um modelo linear generalizado (GLM) cuja variável resposta possivelmente tem uma distribuição de Poisson. Como incorporar o aspecto espacial nesse caso? É necessário incorporar o espaço porque "os mais próximos se parecem" e essa semelhança implica em uma dependência que rompe um dos pressupostos importantes no modelo de regressão. Existem várias alternativas como os modelos auto-regressivos (SAR) e o auto-regressivo condicional (CAR), mas todas trabalham sobre uma matriz de vizinhanças. Essa matriz pode entrar como um modelo auto-regressivo onde a quantidade de eventos nos vizinhos mais próximos afeta a quantidade de eventos em cada área estudada ou incorporando ao componente aleatório ou erro do modelo. Uma mistura dessas duas escalas é possível.

Retornando ao exemplo, é possível supor que, além das variáveis e da localização da residência, seja possível ver os efeitos dos vários setores censitários que participam do estudo. Nesse caso, por exemplo, a proporção de casas ligadas à rede de esgoto é a variável de interesse. O setor censitário funcionará como o segundo nível de um modelo multinível. A matriz de vizinhança deverá ser incluída num efeito aleatório associado neste segundo nível.

É possível utilizar métodos análogos àqueles desenvolvidos para o modelo de regressão linear, em situações em que a variável resposta obedece a outras distribuições GLM que não a Normal, ou em que a relação entre a variável resposta e as variáveis explicativas não é linear.

2.2. SISTEMAS DE INFORMAÇÃO GEOGRÁFICA

Os chamados Sistemas de Informação Geográfica (SIG) são a ferramenta utilizada para captura e organização de dados para tornar possível a compreensão dos fenômenos através de uma análise espacial. Para definir SIG, pode-se reportar às definições de Sistema, Informação Geográfica e Sistema de Informação. Para Korte (1994), um sistema é formado por elementos relacionados de maneira a formar uma unidade ou um todo organizado; a informação geográfica é um conjunto de dados que contem associações ou relações de natureza espacial e o sistema de informação é um conjunto de informações relacionadas entre si, com o objetivo de coleta, entrada, armazenamento, análise e manutenção destas informações.

A definição de SIG mais citada é a de Aronoff (1990): "Os Sistemas de Informação Geográfica (SIG) são um conjunto de ferramentas utilizadas para a manipulação de informações espacialmente apresentadas, com capacidade de atualização, divulgação, armazenamento e gerenciamento de dados geográficos e tabulares".

Outros autores apresentam definições que ajudam a compreender a complexidade funcional e estrutural de um SIG. Marble (1994) e Burrough (1992) enfatizam as características de aquisição, armazenamento, manipulação e exibição dos dados espaciais, enquanto Rodrigues e Quintanilha (1991) definem para os SIG como modelos do mundo real úteis a um certo propósito.

SIG são sistemas cujas principais características são integrar, numa única base de dados, informações espaciais provenientes de dados cartográficos, dados de censo e de cadastro urbano e rural, imagens de satélite, redes e modelos numéricos de terreno; combinar as várias informações, através de algoritmos de manipulação, para gerar mapeamentos derivados; consultar, recuperar, visualizar e plotar o conteúdo da base de dados geocodificados (Câmara, 1993). Devido à natureza geográfica dos objetos, os dados de um SIG são referenciados espacialmente. A tecnologia de SIG integra operações convencionais de bases de dados, como captura, armazenamento, manipulação, análise e apresentação de dados, com possibilidades de seleção e busca de

informações e análise estatística, assim como a visualização e análise geográfica e espacial oferecida pelos mapas. Esta capacidade distingue os SIG dos demais Sistemas de Informação e torna-os úteis para organizações no processo de entendimento da ocorrência de eventos, predição e simulação de situações, e planejamento de estratégias. Os SIG permitem a realização de análises espaciais complexas através da rápida formação e alternação de cenários que propiciam a planejadores e administradores em geral, subsídios para a tomada de decisões. A opção por esta tecnologia, busca melhorar a eficiência operacional e permitir uma boa administração das informações estratégicas, tanto para minimizar os custos operacionais como para agilizar o processo decisório (Carvalho et al., 2000).

2.2.1. Histórico

Os sistemas de informações geográficas surgiram há mais de quatro décadas e têm-se tornado ferramentas valiosas nas mais diversas áreas de conhecimento. Helman Hollerith, do Bureau of Census americano (Austrian, 1982) utilizou pela primeira vez o conceito de computação para o processamento de dados geográficos através de cartões perfurados e uma máquina tabuladora, acelerando assim o processamento do censo de 1890, executado em três anos, o que significou um avanço em relação ao censo anterior, que levou 8 anos de processamento para ser completado. Entretanto, só em 1964 no Canadá foi desenvolvido o primeiro SIG que se tem notícia. Nos anos 1970 se iniciou o amadurecimento dos SIGs. As primeiras versões dos sistemas comerciais aparecem no início da década de 1980, com aceitação mundial. Naquela ocasião, os Governos Federais, seja o americano, o canadense e alguns europeus (Suécia, Noruega, Dinamarca) apoiavam financeiramente iniciativas voltadas tanto à Cartografia Assistida por Computador, quanto aos SIG's. Foi naquele período que o USGS (United States Geological Survey) passou a tornar disponíveis ao público bases de dados digitais (USGS, 2009).

O crescimento efetivo das aplicações de SIG ocorreu entre o final da década 1980 e ínício da década de 1990. Este avanço se deve, em parte, ao advento e à disseminação dos microcomputadores pessoais, além da introdução de tecnologia de relativo baixo custo e alta capacidade de performance. Neste novo século, os SIG

assumiram outra dimensão, a partir da disponibilidade de bases cartográficas digitais, públicas ou privadas, e o desenvolvimento da internet como disseminador de informações.

2.2.2. Funções e objetos de um SIG

Segundo Maguirre (1991), um SIG possui três propriedades básicas: a capacidade de apresentação cartográfica; uma base integrada de objetos espaciais e de seus atributos ou dados e um engenho analítico formado por um conjunto de procedimentos; e ferramentas de análise espacial.

Para realizar as operações de georreferenciamento num SIG, é necessária a presença de um código único que associe informações dos arquivos de atributos com os arquivos geográficos. Esta variável deve estar presente nos bancos de dados gráficos e nos bancos de dados não gráficos, estabelecendo uma ligação entre eles (Figura 2.11).

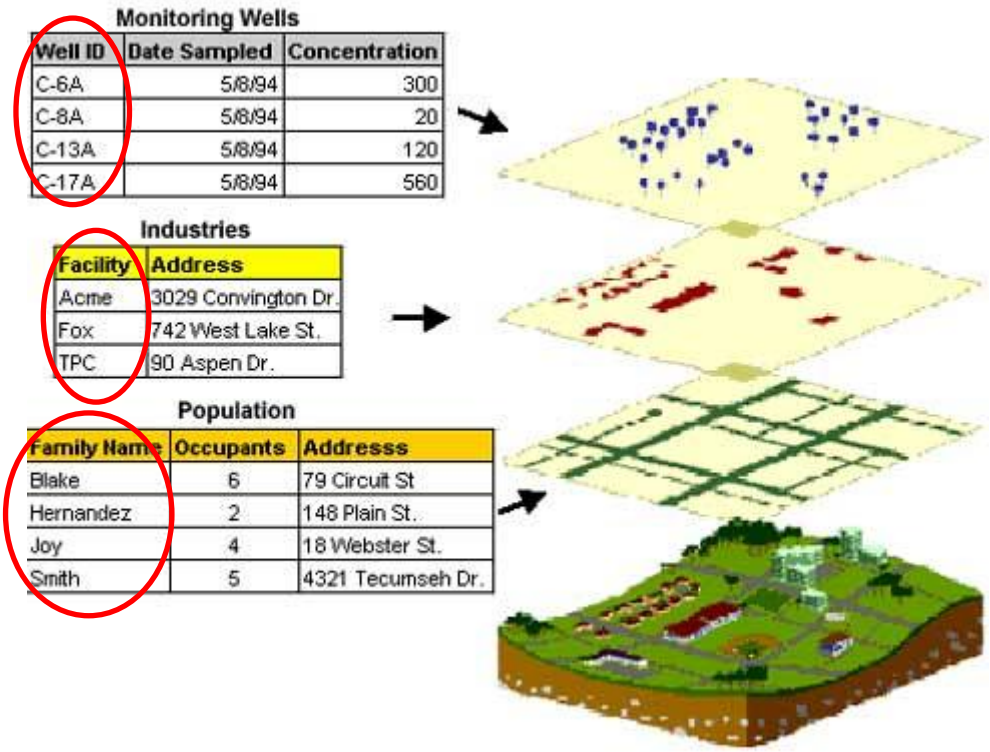


Figura 2.11. SIG - Relação dos elementos com código único (Fonte: www.epa.gov)

2.2.3. Estrutura de um SIG

A estrutura de um Sistema de Informação Geográfica é composta por (Figura 2.12):

- Objetos Geográficos – representações de fenômenos do mundo real. São dispostos em camadas, também chamadas de temas ou planos de informação.

- Atributos – Dados tabulares de um SIG.

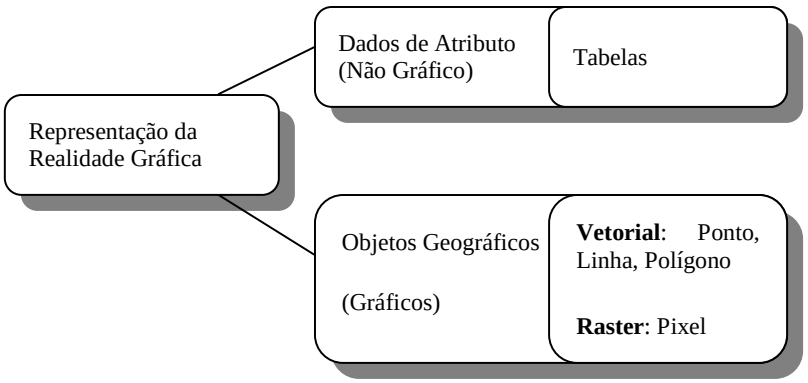


Figura 2.12. Informações de dados espaciais (Adaptado de Scholten & Stillwell, 1990)

A implementação de um SIG é um processo caro e de médio e longo prazo. A decisão de implementá-lo, ou não, deve ser baseada na análise de custo-benefício.

Alguns dos benefícios mais comuns de um SIG são:

- Melhor armazenamento e atualização dos dados;

- Recuperação de informações de forma mais eficiente;

- Produção de informações mais precisas e;

- Rapidez na análise de alternativas.

Estas características propiciam maior probabilidade de tomadas de decisões mais acertadas.

A organização dos dados de atributos é feita segundo as técnicas convencionais de bancos de dados. A grande maioria dos SIG utiliza o modelo relacional, em que a estruturação dos dados se dá através de tabelas, onde cada linha corresponde a uma ocorrência e cada coluna corresponde a um atributo da entidade. O método mais comum de se estabelecer estes relacionamentos é através do armazenamento de códigos comuns, que identifiquem univocamente a entidade, e que recebem o nome de chave primária. Por outro lado, os elementos gráficos são dispostos em camadas superpostas, sendo o relacionamento entre essas camadas possível de ser observado através das coordenadas de localização dos elementos (Figura 2.13).

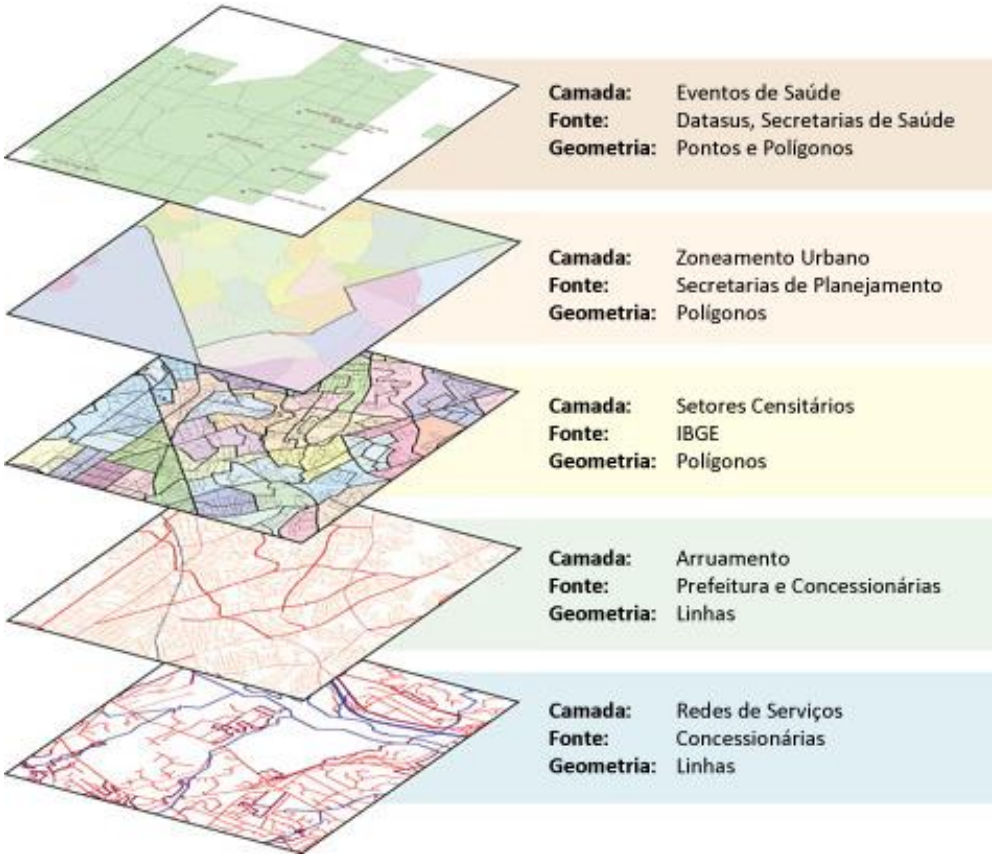
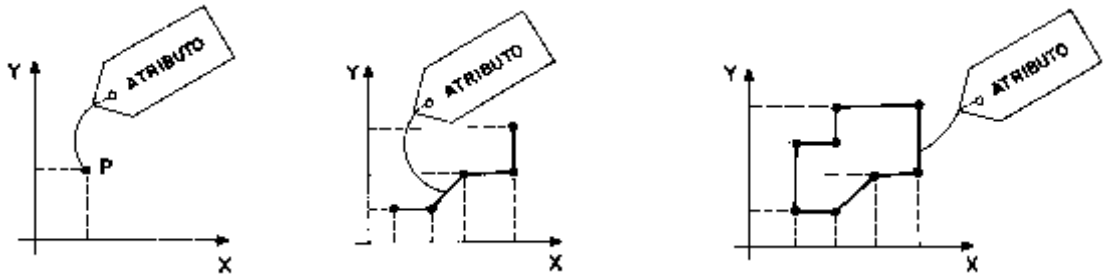


Figura 2.13. Estrutura dos elementos gráficos de um SIG (camadas)

Pontos, linhas e áreas (ou polígonos) são os elementos que permitem a estrutura vetorial representar os dados da forma mais precisa uma vez que suas coordenadas geográficas estão em um espaço contínuo e possibilitam descrição exata de posição,

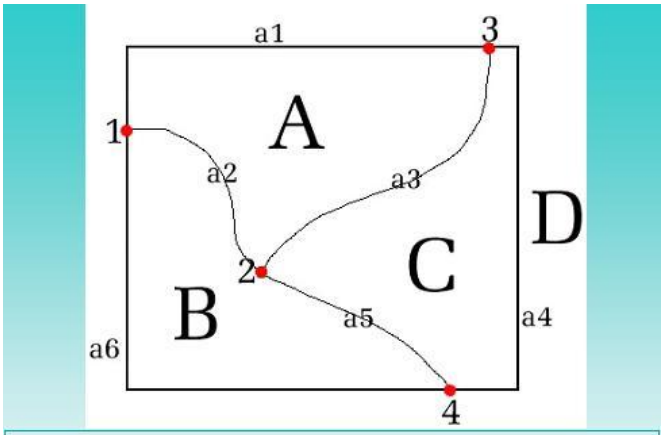
tamanho e dimensão. Ponto é uma entidade que pode ser localizada por um par de coordenadas geográficas. É utilizada para representar a localização da ocorrência de um fenômeno, ou para representação, em um mapa, de uma feição que é muito pequena para ser mostrada como uma área ou linha. Exemplos: localização de um lote de terreno, uma cidade ou o pico de uma montanha. Uma linha é definida por no mínimo dois pares de coordenadas geográficas (dois pontos). Representar feições do mapa que são muito estreitas para serem mostradas como área ou que teoricamente não têm espessura. Exemplos: um logradouro, uma rodovia, ou um limite administrativo. Enquanto a área é uma série de coordenadas geográfica (pontos), formando segmentos de linhas que fecham uma área e freqüentemente representam-se elementos de área por polígonos. Exemplos: e, um lago, ou extensão geográfica de uma cidade.



Ponto Linha Área

Figura 2.14. Elementos gráficos vetoriais de um SIG, com atributo associado (adaptado de Câmara, 1993)

Os elementos gráficos vetoriais podem estar em uma estrutura topológica. Topologia é definida como a parte da matemática que estuda as propriedades geométricas que não variam mediante uma deformação, especificamente o relacionamento espacial entre os objetos, como por exemplo, proximidade e vizinhança. Formas e coordenadas dos objetos são menos importantes que os elementos do modelo topológico como conectividade, contiguidade e continência. A definição da topologia explicita os relacionamentos espaciais entre os objetos através de um processo matemático. Na Figura 2.15 é representada uma estrutura topológica, com os nós (1, 2, 3 e 4), arestas (a1, a2, a3, a4, a5 e a6), polígonos (A, B, C e D) e as relações entre estes elementos.



Topologia					
Polígono: área		Nó: conectividade		Arco: contigüidade	
Polígono	Arcos	Nó	Arcos	Arco	Polígonos Esquerda & Direita
A	a1, a2, a3	1	a1, a2, a6	a1	A D
B	a2, a5, a6	2	a2, a3, a5	a2	A B
C	a3, a4, a5	3	a1, a3, a4	a3	A C
D	a1, a4, a6	4	a4, a5, a6	a4	C D
				a5	B C
				a6	B D

Figura 2.15. Estrutura topológica (UNBC GIS Lab, 2008)

2.2.4. Componentes de um SIG

Um modo útil de organizar os componentes de um SIG é como um núcleo técnico e administrativo cercado por um anel de usuários envolvidos com diferentes aplicações (Figura 2.16). No coração de qualquer SIG estão o hardware, o software, os bancos de dados e o pessoal envolvido na operação, manutenção e administração do próprio sistema.

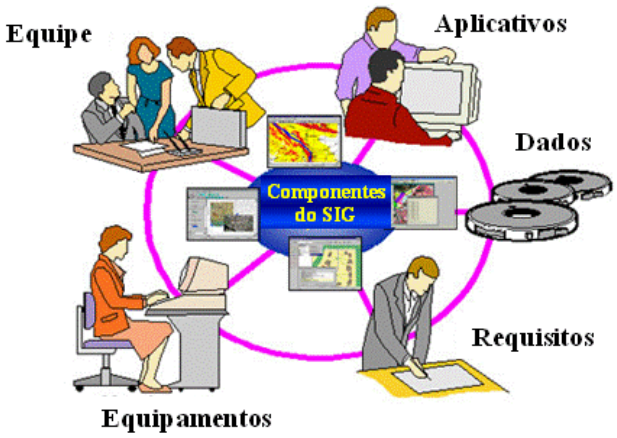


Figura 2.16. Componentes de um SIG

2.2.5. Aquisição de dados

Tradicionalmente, a aquisição de dados representa um papel muito importante em qualquer SIG ou num sistema de mapeamento digital. Sempre foi considerada a atividade mais onerosa e alguns peritos estimam que mais de três quartos do custo de operação de um SIG recaiam na criação de banco de dados. Entretanto, com o aparecimento de imagens orbitais de alta resolução e a disponibilidade, cada vez maior, de bases cadastrais, de utilização de GPS e o acesso a informações de localização e rotas pela internet, fica mais simples, para uma gama de projetos, a obtenção das informações gráficas necessárias.

Os dados gráficos para um SIG são obtidos de muitas formas (Figura 2.17) tais como utilização de GPS para obtenção das coordenadas; levantamento topográfico; fotogrametria, com utilização de fotos aéreas ou por imagens de satélites; utilizando conjuntos de dados previamente existentes, digitalização de mapas, por scanner ou mesa digitalizadora (Carvalho et al., 2000). Os dados não gráficos são obtidos através de cadastros existentes e resultados de censos e pesquisas.

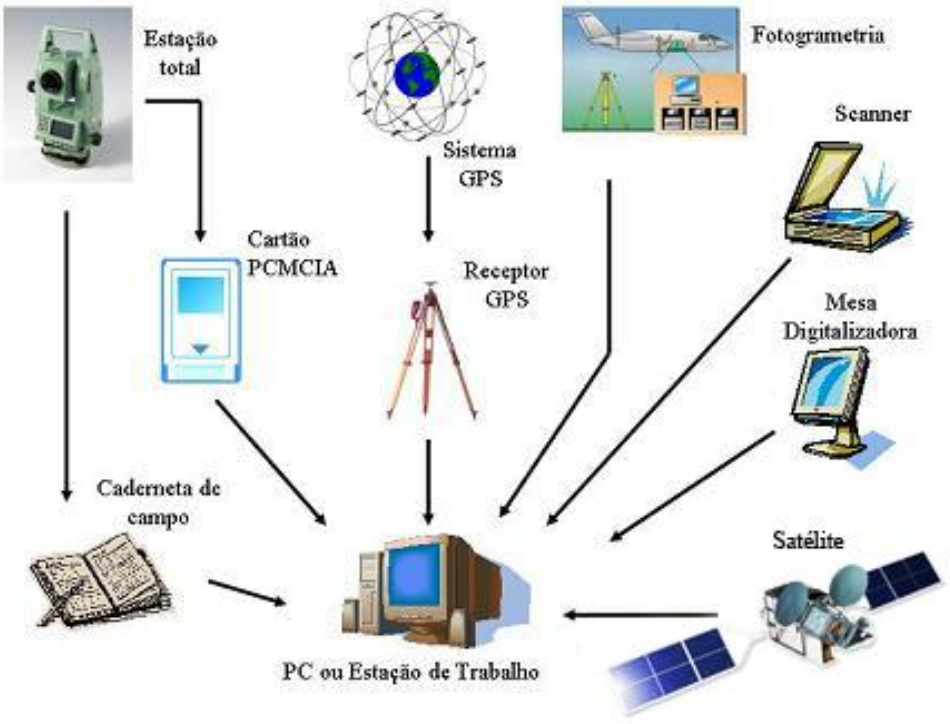


Figura 2.17. Os principais métodos de coleta de dados gráficos utilizados em SIG

2.2.6. Georreferenciamento dos dados

O referenciamento dos dados espaciais à superfície terrestre tem estratégias diferentes para dados gráficos e não-gráficos. O referenciamento de dados gráficos (mapas), chamado de georreferenciamento se dá através da associação a um sistema de coordenadas terrestres quaisquer. Normalmente este processo de georreferenciamento se dá durante o processo de digitalização, garantindo a possibilidade de se associar mapas distintos. Este é um cuidado que deve estar sempre presente no processo de aquisição de bases, pois de outro modo, não é possível sobrepor níveis de informações oriundos de outras fontes. Os *softwares* de SIG possuem funções que auxiliam na tarefa de georreferenciamento de dados tabulares, que pode ser efetuado de diversas maneiras e pode ser entendido como o processo de associar dados a um mapa. Este georreferenciamento pode ser feito através de pares de coordenadas ou através do relacionamento com unidades espaciais (setores censitários, bairros, etc.) presentes no mapa. Além disso, existem programas que permitem localizar eventos em trechos de ruas, através da interpolação entre os números iniciais e finais de cada trecho.

É importante lembrar que a unidade de georreferenciamento deve estar presente tanto na base de dados tabulares quanto no mapa. No caso do uso de trechos de ruas, é necessária a construção de uma base cartográfica correspondente, contendo todos os trechos de rua com o nome do logradouro e numeração, assim como o cadastro de todos os logradouros contendo face de quadra (trecho entre esquinas), lado par, lado ímpar. A construção deste tipo de mapa pode ser muito trabalhosa e de alto custo, dependendo das dimensões da cidade. Além disso, esta estratégia pode ser pouco viável em áreas de ocupação urbana irregular, onde não há seqüência na numeração, e em áreas rurais, onde os endereços raramente são baseados em logradouros.

O georreferenciamento de dados tabulares é ainda um dos fatores limitantes da plena utilização dos SIG na área da saúde, quando se trata de análises em microáreas, em que o endereço do evento é fundamental. Os principais Sistemas Nacionais de Informações da Saúde possuem o campo referente a endereço. O georreferenciamento destas informações é possível a partir deste endereço. Este georreferenciamento é denominado geocodificação e tratado no capítulo 3.

2.3. O ENDEREÇO

O endereço é a forma mais comum de localização de pessoas. Há várias formas de se especificar um endereço, cada uma com seus objetivos e particularidades. Quando o objetivo é de envio de uma mensagem (carta, recado), pode ser utilizado o endereço residencial, endereço comercial, caixa postal dos Correios, o código de endereçamento postal (CEP), um endereço eletrônico (e-mail) ou o número do telefone (fixo ou celular). No entanto, quando tratamos de ações sociais ou especificamente de saúde, o endereço residencial é o mais útil. Necessitamos transformar estes endereços em objetos localizados no espaço geográfico (em forma de ponto, linha ou polígono) de modo a possibilitar sua associação com outros endereços ou a indicadores sócio-econômicos ou ambientais relacionados à área pesquisada. Neste sentido, é necessário conhecer bem a composição da informação de endereço que, na maioria das vezes, é fornecida na forma de texto livre.

O modo como o endereço é apresentado depende de fatores culturais, de organização da ocupação das áreas e dos padrões instalados em cada local. Os padrões utilizados pela população seguem, em cada país, as regras de endereçamento para fins postais. A seguir são apresentadas algumas estruturas de endereço nos diversos países.

Em alguns países orientais, como o Japão e a Coreia, os padrões utilizados não seguem a mesma lógica dos ocidentais (Davis et al., 2003). Em Tóquio, o endereço apresenta uma estrutura que podemos chamar de hierárquica, com o número do prédio sendo relacionado com o bloco ou quarteirão deste e o quarteirão com uma área maior e assim sucessivamente chegando até à cidade. Na Figura 2.18 há um exemplo de endereçamento em Tóquio. Foi utilizado o endereço do JTNO Tourist Information Center (TIC) em Tóquio, como um exemplo. A TIC é na cidade de Tóquio, no ku (distrito) de Chiyoda, na área de Yurakucho; no chome No. 2 (sub-área), no quarteirão número No. 10 e o número do prédio é 1. (http://www.digi-promotion.com/tokyo-info/info-maps-address.html). O sistema de numeração dos prédios é bem antigo e caótico. Em um dado momento há algumas gerações, o primeiro prédio construído numa vizinhança foi identificado com o número 1, o segundo, podendo ser localizado

fora deste quarteirão, foi identificado com o número 2 e assim por diante. Logo não há uma sequência de numeração por local e sim por antiguidade.

Figura 2.18. Estrutura de endereço de Toquio (http://www.digi-promotion.com/tokyo-info/info-maps-address.html)

Na Coreia, a maior parte das cidades não possui números nos prédios nem nomes nos logradouros. Entretanto cada prédio tem um número oficial que é definido quando este é construído, sem uma lei de formação padronizada. Deste modo, um prédio com o número 27 pode estar do lado de um com o número 324. Em relato de um responsável por uma empresa de entrega de encomendas, pode-se ter a ideia da dificuldade da localização destes endereços para a execução de seu serviço (www.teachkoreanz.com/living/address.htm):

"Muitas vezes nossa empresa pergunta ao cliente para ser bem específico sobre a localização de seu endereço e este envia por fax a resposta. No caso da Coreia, normalmente é enviado um mapa para esta identificação. Este é o modo que encontramos para fazermos corretamente as entregas". Em algumas ocasiões, governos locais tentaram desenvolver sistemas mais amigáveis de endereçamento, mas, além de ser muito complexo e caro, esbarra na resistência da população por questões culturais.

Nos Estados Unidos a estrutura do endereço urbano é hierárquica, começando pelo número do prédio, passando pelo nome do logradouro, a direção, a cidade, o estado e complementando com o código postal (*zip code*). Nas áreas rurais, o endereço é

especificado por rotas rurais (Rural Route - RR) ou os chamados endereços de contrato de autovias (Highway Contract Addresses - HC) identificam os pontos de entrega, que possuem conjuntos de caixas individuais onde os residentes buscam suas correspondências (Goldberg et al., 2008).

No Quadro 2.2 são apresentados os endereços das embaixadas brasileiras (obtidas através de links encontrados no site do Ministério de Relações Exteriores – www.mre.gov.br), como exemplo das diferenças regionais de especificação de endereço, inclusive na estrutura dos códigos de endereçamento postais (quando estes existem). Estas diferenças se apresentam também em função da estrutura da língua adotada no país, como no caso de Alemanha e Holanda, com o tipo e o nome do logradouro formando uma só palavra. No caso da Coreia do Sul, o endereço apresentado no site da embaixada (www.brasemb.or.kr) é o mapa da Figura 2.19.

País	Endereço
Alemanha	Wallstrasse 57, 10179 – Berlin
Argélia	55. Bis, Chemin Cheikh Bachir El-Ibrahimi. El Biar. Alger.. BP 246
Paraguai	Calle Coronel Irrazábal c/ Eligio Ayala, Casilla de Correo 22, Asunción
Tailândia	34 Floor Lumpini Tower, 1168/101 Rama IV Road Thungmahamek, Sathorn, Bangkok 10120
Síria	Al-Farabi Street , Building No. 39, Mezzeh – Eastern Villat – Damascus, POBOX 2219
Holanda	Mauritskade, 19, 2514HD Haia
Índia	8, Aurangzeb Road, New Delhi – 110011
Israel	Rechov Yehuda HaLevi, n. 23, 30º andar,Tel-Aviv, 65136
Estados Unidos	3006 Massachusetts Avenue, NW, Washington, DC, 20008-3634
Suécia	Odengatan 3, 114 24 Stockholm, Metrô: Tekniska Högskolan (linha vermelha), Ônibus: 4, 43, 72, 624, 628, 670 e 680

Quadro 2.2. Embaixadas do Brasil (www.mre.gov.br)

No Brasil existe um padrão geral dos endereços urbanos utilizados pela agência de Correios (www.correios.com.br), com o endereço (logradouro, número e complemento) acompanhado da cidade, estado e código de endereçamento postal (CEP). Em aproximadamente 200 municípios brasileiros, existe um CEP para cada logradouro ou parte deste, facilitando assim sua localização quando este é corretamente preenchido. Há cidades em que o padrão geral não é obedecido, como o plano piloto de

Brasilia, que segue uma estrutura hierárquica (setor, super-quadra, quadra, lote e número). Na Figura 2.20 é apresentada uma planta do Plano Piloto de Brasília.

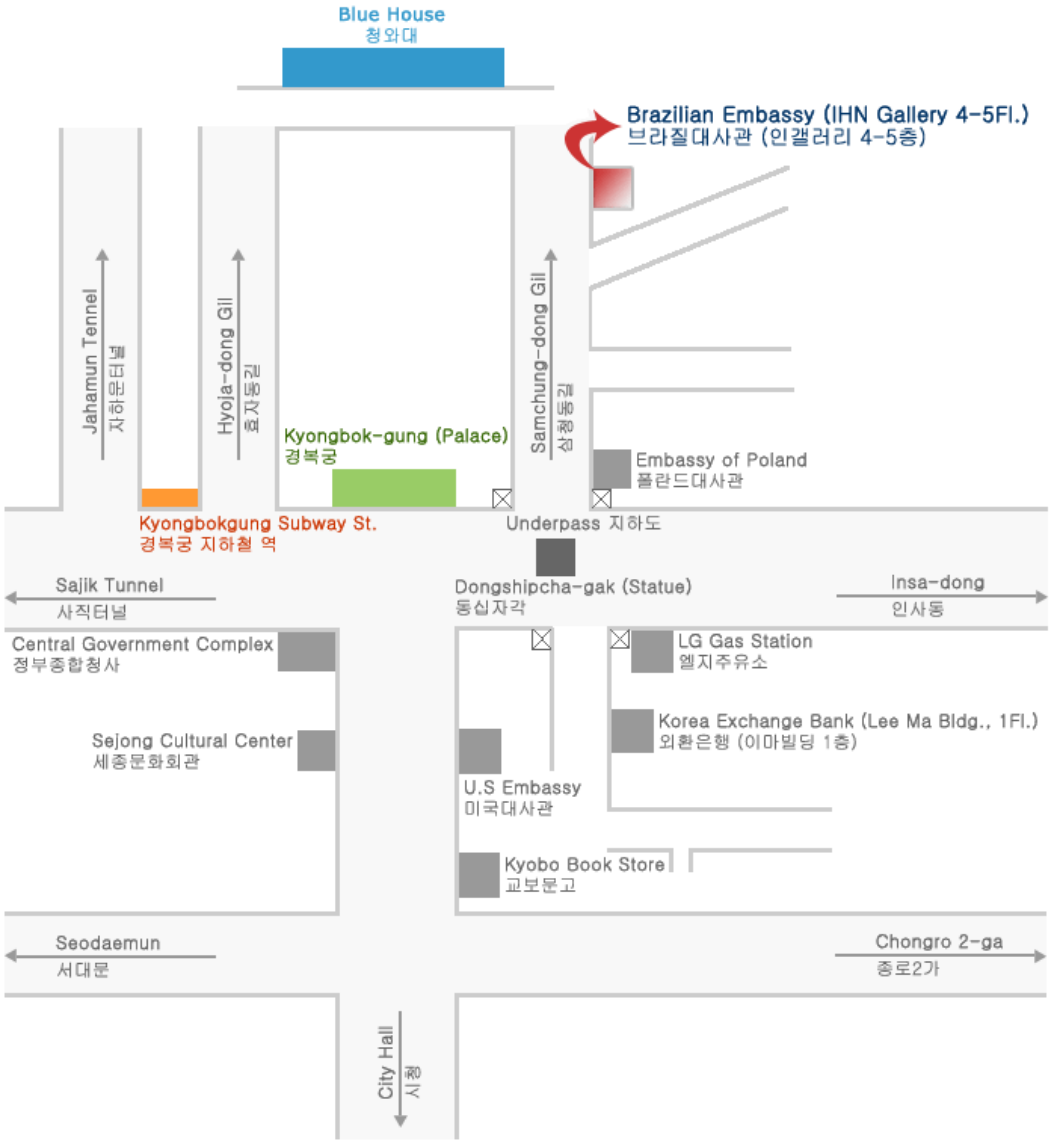


Figura 2.19. Endereço da embaixada brasileira na Coreia do Sul (www.brasemb.or.kr)

Em áreas de ocupação desordenada, como favelas ou invasões (Figura 2.21), os logradouros não são claramente definidos e muitas vezes não identificados com endereços individuais. Para os serviços de entrega de correspondência, muitas vezes é utilizado um endereço único, sendo geralmente o da sede da associação de moradores da comunidade. Outro problema é encontrado em novos loteamentos nas periferias das

cidades, onde os logradouros recebem nomes provisórios como “Rua Projetada” ou “Rua A” e os CEPs oficiais ainda não estão definidos, dificultando assim a localização de um endereço individual. O endereçamento das áreas rurais, no Brasil, é muitas vezes identificado apenas pelo nome da localidade.

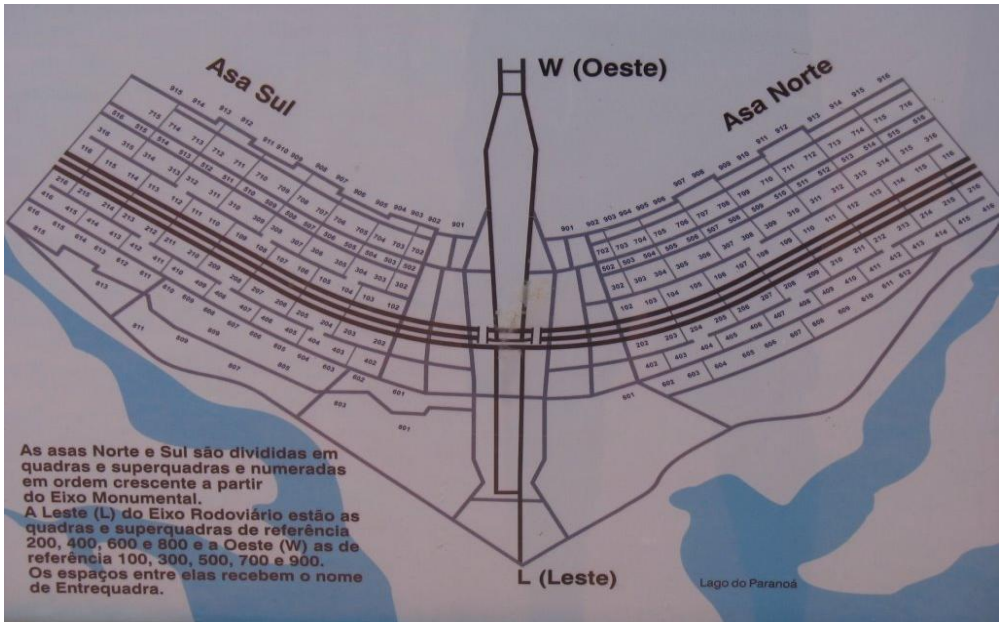


Figura 2.20. Planta do Plano Piloto de Brasilia, com suas quadras e superquadras

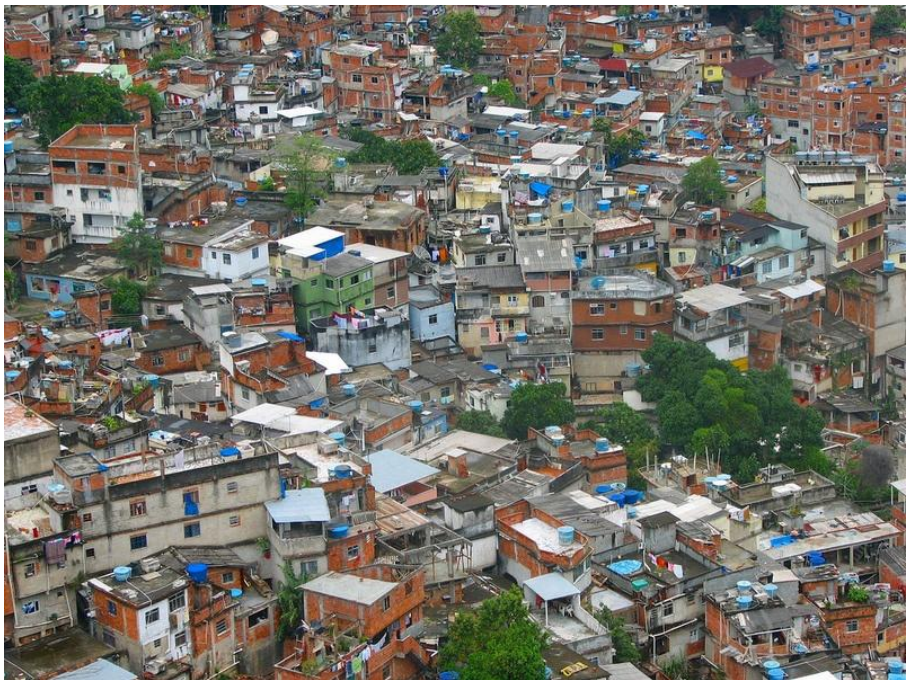


Figura 2.21. Favela do Rio de Janeiro

O CEP

O Código de Endereçamento Postal (CEP), com estrutura de 5 (cinco) dígitos, foi criado pela empresa Brasileira de Correios e Telégrafos, em maio de 1971. Sua divulgação ao público em geral ocorreu com a publicação do Guia Postal Brasileiro, Edição 1971. Em maio de 1992, sua estrutura foi alterada para 8 (oito) dígitos e oficializada junto ao público em geral, com a publicação do Guia Postal Brasileiro, Edição 1992 (www.correios.com.br). Está estruturado segundo o sistema decimal, sendo composto de Região, Sub-região, Setor, Subsetor, Divisor de Subsetor e Identificadores de Distribuição (Figura 2.22).

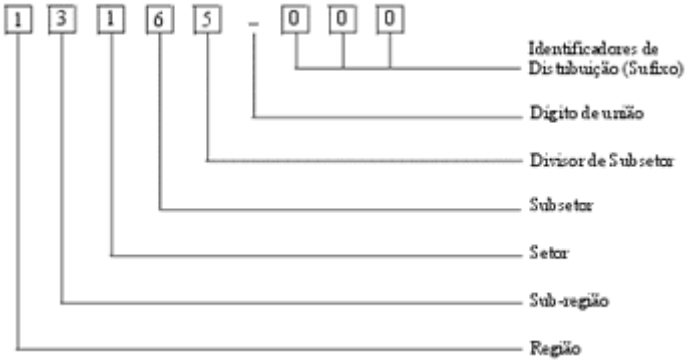


Figura 2.22. Estrutura do CEP

O Brasil foi dividido em dez regiões postais para fins de codificação postal, utilizando como parâmetro o desenvolvimento sócio-econômico e fatores de crescimento demográfico de cada Unidade da Federação ou conjunto delas, sendo:

0 – Grande São Paulo	5 – PE, AL, PB e RN
1 – Interior de SP	6 – CE, PI, MA, PA, AM, AC, AP e RR
2 – RJ e ES	7 – DF, GO, TO, MT, MS, RO
3 – MG	8 – PR e SC
4 – BA e SE	9 – RS

Cada região é dividida em 10 sub-regiões e assim sucessivamente. Os três algarismos após o hífen são denominados de SUFIXO e destinam-se à identificação individual de Localidades, Logradouros, Códigos Especiais e Unidades do Correio, sendo utilizados os valores segundo a codificação abaixo:

- Logradouros: 000 a 899
- Códigos Especiais: 900 a 959
- CEPs Promocionais: 960 a 969
- Unidades dos Correios: 970 a 989 e 999
- Caixas Postais Comunitárias: 990 a 998

3. METODOLOGIAS UTILIZADAS NA GEOCODIFICAÇÃO

"A essência do conhecimento consiste em aplicá-lo, uma vez possuído" (Confúcio)

Neste capítulo são apresentados os resultados da revisão bibliográfica das pesquisas relacionadas às metodologias utilizadas no processo de geocodificação. Algumas destas técnicas devem ser adaptadas, levando-se em conta as diferenças de padrões de endereço (item 2.3) dos diferentes países.

O primeiro item (3.1) apresenta o conceito de geocodificação e a definição assumida nesta tese. O item seguinte (3.2) refere-se a uma revisão da utilização de geocodificação na área da saúde pública, enquanto o terceiro (3.3) às bases de dados que servem de referência para os processos. O item 3.4 apresenta algumas rotinas utilizadas para tratamento dos campos na preparação e comparação dos textos. No último item (3.5) são discutidos os métodos de tratamento dos dados de entrada para o processo.

3.1. O QUE É GEOCODIFICAÇÃO (GEOCODING)?

Literalmente, geocodificação significa atribuir um código geográfico a um elemento cartográfico ou do mundo real. Fazendo uma pesquisa na Wikipedia (utilizando *geocoding*), obtemos: "*Geocoding is the process of finding associated geographic coordinates (often expressed as latitude and longitude) from other geographic data, such as street addresses, or zip codes (postal codes). With geographic coordinates the features can be mapped and entered into Geographic Information Systems, or the coordinates can be embedded into media such as digital photographs via geotagging*"[1]. Esta definição nos remete à associação de coordenadas geográficas a dados geográficos apresentados de forma textual.

Há várias formas de se referenciar a "local" ou "espaço". Na área da saúde, esta informação está tipicamente associada a endereço, bairro, cidade/município e estado do paciente, normalmente conhecido como "endereço residencial". A estes dados podemos incluir o código de endereçamento postal (CEP). Este modo de descrição é facilmente entendido pelas pessoas, mas não é adequado para uso diretamente em um sistema computadorizado. Como qualquer informação geográfica de mapeamento ou consulta com ajuda de um computador, em vez de uma descrição textual, a informação precisa ser discreta, não ambígua, geograficamente determinável.

Neste sentido, precisamos de um processo de conversão da descrição textual em um dado geoespacial válido. Este conceito de transformar informação espacial implícita em explícita, ou de converter uma "informação não-geográfica" em "informação geográfica" é denominado georreferenciamento (Hill, 2006).

---

[1] *"Geocodificação é o processo de encontrar coordenadas geográficas associadas (normalmente expressa como latitude e longitude) por outros dados geográficos, tais como endereços residenciais ou códigos postais (CEP). Com as coordenadas geográficas, os elementos podem ser mapeadas e incorporados a Sistemas de Informação Geográfica, ou as coordenadas podem ser incorporadas a mídias como fotografias digitais".*

Nas últimas décadas, foram desenvolvidas algumas formas de georreferenciamento, cada uma delas adequada a diferentes necessidades ou disponibilidades locais. Uma das formas é a utilização de GPS, determinando as coordenadas do local com a precisão permitida pelo aparelho utilizado. Esta alternativa, apesar de ser a mais simples, é muito dispendiosa, uma vez que é necessária a ida de uma pessoa a cada local para obter estas coordenadas.

Outro método de georreferenciamento é a geocodificação, que pode ser definida como processo de relacionamento de dados tabulares, que contêm informação de endereços, com coordenadas do mundo real. O conceito de geocodificação foi sendo adaptado, através do tempo, a partir das mudanças de disponibilidade de dados geográficos. A crescente disponibilidade, precisão (acurácia) e confiabilidade das séries de dados geográficos digitais contribuíram com que o processo de geocodificação evoluísse continuamente para acompanhar o ritmo de desenvolvimento de produção destas bases de dados. Desta forma, os profissionais têm ampliado os limites sobre quais os tipos de informação podem ser geocodificadas, incorporando informações de várias fontes. Na década de 1960, sistemas de geocodificação utilizados pelo Censo dos EUA apenas inseriam endereços postais e edificações identificadas em zonas geográficas delineadas por códigos numéricos (O'Reagan & Saalfeld, 1987), não os associava com objetos válidos como pontos, linhas ou áreas com que os consumidores de dados geocodificados estão acostumados hoje. O resultado desta evolução é uma certa confusão de conceitos de geocodificação, adaptados às necessidades específicas e disponibilidade de dados. (Goldberg et al., 2008).

Este salto de qualidade nos processos de geocodificação pode ser diretamente creditado aos avanços tecnológicos e ao aparecimento e disponibilidade de bases de dados, às quais estão referenciados estes processos. As tentativas mais antigas de geocodificação eram atrapalhadas pela falta de bases cartográficas digitais para utilização na determinação de localização dos endereços.

Nesta tese será adotada a definição feita por Eichelberg (1993):

“**A geocodificação de um endereço é definida como o processo de associação deste a objetos contidos em um mapa terrestre**”.

3.2. UTILIZAÇÃO DA GEOCODIFICAÇÃO NA ÁREA DA SAÚDE

Com o objetivo de conhecer a evolução dos métodos utilizados para a geocodificação, as preocupações no uso destes métodos e o perfil das análises que utilizam os dados resultantes do processo, foi realizada uma revisão dos artigos publicados na área da saúde que utilizavam métodos de geocodificação.

A metodologia utilizada para a busca de artigos incluiu pesquisas em bases eletrônicas e busca manual de citações nas publicações inicialmente identificadas. Foram utilizadas as bases Medline, Pubmed, Biomed, Lilacs e Scielo, além de acesso a trabalhos obtidos a partir de informações prévias. Os descritores utilizados são "geocoding", "geoprocessing", "address", "health". As referências bibliográficas dos estudos assim localizados foram também rastreadas para localizar outras intervenções de potencial interesse. A partir deste ponto, foram selecionados os artigos que especificaram o método de geocodificação utilizado ou discutiram algum aspecto do processo, como a precisão do posicionamento ou a validade das bases disponíveis.

Como resposta às buscas, foram identificados 130 artigos. A partir destes, foram selecionados 30 trabalhos, seguindo as restrições estabelecidas. Os resultados obtidos estão descritos abaixo.

Quanto ao assunto tratado nos artigos, em 7 destes são discutidos as bases de dados utilizadas para a geocodificação[2] envolvendo tanto um padrão nacional, como é o caso inglês (Morad, 2002), tanto outros tipos de endereço, como o caso de pesquisa a endereços de caixa postal (Hurley et al., 2003) ou as bases resultantes dos Censos demográficos nacionais[3]. Outros trabalhos apresentam a idéia de novos tipos de bases, como as coordenadas de esquinas de logradouros (Levine & Kim, 1998) ou outros tipos de áreas (Gregorio et al., 2005). Em outros 5 artigos[4] é discutida a metodologia utilizada, destacando-se o trabalho de Boulos (2005) que propõe a utilização de mapas e bases disponíveis na Internet, uma nova alternativa que está surgindo com baixo custo.

[2] Levine & Kim, 1998; Gregorio et al., 2005; Rose et al., 2004; Morad, 2002; Boscoe et al., 2004; Hurley et al., 2003; Skaba et al. 2004.

[3] Rose et al., 2004; Boscoe et al., 2004; Skaba et al. 2004.

[4] Levine & Kim, 1998; Skaba et al., 2004; McElroy et al., 2003; Boulos, 2005; Wilmersdorf, 2003.

A precisão da geocodificação utilizando as bases de dados disponibilizadas pelos Censos é o assunto tratado por 4 artigos[5], mostrando as limitações no uso destes dados. Nos 3 primeiros é tratada a precisão em área urbana e no último destes na área rural, em que as distâncias são maiores, gerando maiores distorções.

Em 13 artigos[6] são utilizados os dados geocodificados a áreas de Censos para a utilização dos dados agregados na determinação do nível sócio-econômico ou de condições ambientais e de vizinhança. Nestes casos foram importantes as taxas de endereços encontrados para a determinação do contexto em que se encontravam estas pessoas. Outros assuntos abordados foram o de determinação de rotas para atendimento emergencial[7] e preparação de dados para a análise espaço-temporal (Han et al., 2005).

Quanto ao local de desenvolvimento dos estudos destacam-se 2 países, os Estados Unidos[8] com 16 artigos e Inglaterra[9] com 4. Os outros artigos são divididos entre Brasil (Skaba et al., 2004; Davis et al., 2003), Austrália (Duncan & Mummery, 2004; Hyndman & Holman, 2001), Canadá (Boscoe et al., 2004), Áustria (Wilmersdorf, 2003), Nova Zelândia (Skelly et al., 2002), Costa Rica (Rosero-Bixby, 2004), Bolívia (Kinman, 1999) e Israel (Peleg & Pliskin, 2004).

Analisando-se o ano de publicação dos artigos, nota-se que apenas dois são anteriores ao ano de 2000, sendo publicados em 1998 (Levine & Kim) e 1999 (Kinman). Os outros trabalhos foram apresentados a partir de 2001.

Em 15 artigos são informados o tamanho da amostra de endereços a serem geocodificados e o percentual de acertos no processo. A tabela 1 apresenta estes dados, relacionando-os ao país da pesquisa, ao ano e à unidade referenciada. Esta tabela mostra a diversidade unidades de análise, que variam com os objetivos das pesquisas, assim

[5] Cayo & Talbot, 2003; Davis et al., 2003; Krieger et al., 2001; Skelly et al., 2002.

[6] Samantha & Martin, 2005; Chen et al., 2004; Rosero-Bixby, 2004; Alexander et al., 2003; Duncan & Mummery, 2004; Rutt & Coleman, 2005; Oyana & Rivers, 2005; Laraia et al., 2004; Huff & Gray, 2001; Kaufman et al., 2003; Burdette & Whitaker, 2004; Kinman, 1999; Hyndman & Holman, 2001.

[7] Wilmersdorf, 2003; Boulos, 2003; Peleg & Pliskin, 2004.

[8] Levine & Kim, 1998; Gregorio et al., 2005; Rose et al., 2004; Hurley et al., 2003; McElroy et al., 2003; Cayo & Talbot, 2003; Krieger et al., 2001; Chen et al., 2004; Alexander et al., 2003; Rutt & Coleman, 2005; Oyana & Rivers, 2005; Laraia et al., 2004; Kaufman et al., 2003; Burdette & Whitaker, 2004; Boulos, 2003; Han et al., 2005.

[9] Morad, 2002; Boulos, 2005; Samantha & Martin, 2005; Huff & Gray, 2001.

como as diferenças de eficiências na geocodificação, mesmo quando é utilizada a mesma base de dados.

Como conclusão desta revisão, verificou-se que a escolha do setor censitário ou áreas equivalentes como unidade de análise dos eventos de saúde mostrou-se adequado, como citado em vários artigos já publicados. Além da quantidade de informações indexadas a esta unidade, a existência de cadastros padronizados facilita a recuperação desta unidade a partir do endereço. Ficou nítida a falta de padronização no tratamento dos endereços, mesmo quando utilizadas as mesmas bases de dados como é o caso da base de dados do Censo 2000 dos Estados Unidos. Os resultados obtidos (Tabela 3.1) apresentam grandes diferenças na eficiência.

O nível de desenvolvimento das bases para a geocodificação é um fator determinante da utilização deste processo. Os trabalhos para as montagens das bases de endereços nos Estados Unidos e Inglaterra iniciaram-se na década de 1960, enquanto no Brasil este trabalho começou de um modo mais sistemático no Censo 2000.

Nota-se que, apesar dos cadastros terem sido desenvolvidos neste período, os trabalhos utilizando geocodificação são recentes. Isto pode ser associado ao desenvolvimento tecnológico havido nos últimos anos, principalmente no tratamento dos elementos gráficos.

Outro fator que leva à reflexão é o número de artigos proporcionalmente alto com interesse nas variáveis de determinação de níveis sócio-econômicos e ambientais, associáveis quando a geocodificação permite a localização do endereço nas unidades de levantamento do Censo. Esta associação é possível a partir de dados agregados de cada micro-área existente na área de abrangência da pesquisa.

Fica evidente neste trabalho que existe ainda um grande caminho a ser traçado para o desenvolvimento desta área, tanto na criação de bases de dados gráficas e alfanuméricas quanto nos processos de geocodificação utilizando estas bases.

Autores	Unidade de Análise	País	Ano	Amostra	%
Levine et al.	Áreas do Censo 1990	Estados Unidos	1998	15.975	96,5
Gregório et al.	Áreas do Censo 2000	Estados Unidos	2005	22.562	93,4
Rose et al.	Áreas dos Censos 1960 a 1980	Estados Unidos	2004	24.148	90,0
Skaba et al.	Áreas do Censo 2000	Brasil	2004	4.094	77,0
McElroy et al.	Coordenadas geográficas	Estados Unidos	2003	14.804	97,0
Cayo et al.	Coordenadas geográficas	Estados Unidos	2003	3.000	100,0
Skelly et al.	Áreas do Censo 2000 (rural)	Nova Zelândia	2002	39.757	3,5
Chen et al.	Áreas do Censo 2000	Estados Unidos	2004	117.209	37,0
Duncan, Mummery	Áreas administrativas	Austrália	2005	1.281	94,0
Rutt, Coleman	Coordenadas geográficas	Estados Unidos	2004	942	48,0
Oyana, Rivers	Coordenadas geográficas	Estados Unidos	2005	11.577	90,0
Laraia et al.	Áreas do Censo 2000	Estados Unidos	2004	3.163	100,0
Kaufman et al.	Áreas do Censo 2000	Estados Unidos	2003	1.747	100,0
Burdette, Whitaker	Bairro	Estados Unidos	2004	11.246	90,0
Han et al.	Coordenadas geográficas	Estados Unidos	2005	15.487	82,0

Tabela 3.1. Eficicácia na geocodificação

3.3. BASES DE DADOS DE REFERÊNCIA

As primeiras bases de dados com referências de endereços ou logradouros tinham como objetivo servir de referência para as operações de coleta de dados para os censos demográficos. Estas operações precisam garantir a cobertura de todos os domicílios do país para, a partir destes, conseguir registrar a população residente nestes domicílios. Deste modo, as estruturas organizadas para os censos são baseadas nos domicílios.

O *Census Bureau* dos Estados Unidos deu início, nos anos 1960, à geração de bases de dados com códigos geográficos. Primeiramente foi criado o Dual Independent Map Enconding (DIME) com uma estrutura ainda limitada (Goldberg, 2008). O desenvolvimento de arquivos geográficos com estruturas vetoriais como o banco de dados Topographically Integrate Geographic Encoding and Referencing (TIGER) (US ensus Bureau, 2009) possibilitou o aparecimento de novas gerações de algoritmos de geocodificação utilizando interpolação, aumentando significativamente a resolução da representação geográfica (Ratcliffe, 2001).

Um avanço no desenho e desenvolvimento de bases de dados para geocodificação se deu com a criação de bancos de registros de endereços residenciais geocodificados como o ADDRESS-POINT (www.ordnancesurvey.co.uk) e o G-NAF (Paull, 2003) no Reino Unido e Austrália, respectivamente. Estes bancos facilitam a capacidade de geocodificação mais precisa em escalas nacionais (Christen & Churches, 2005).

No Brasil, o IBGE produziu, no Censo 2000, as malhas de setores censitários urbanos e rurais para todos os mumicípios brasileiros (Skaba & Terron, 2003), além de um cadastro de segmentos de logradouros por setor censitário para as 1058 maiores cidades brasileiras. Este cadastro alfanumérico permite referenciar os endereços identificados às áreas dos setores (Figura 3.1).

Segm2000 : Tabela

UF	MUNIC	DISTR	SDIST	SETOR	SEQ	TIPO	TITULO	NOME	INIPAR	FIMPAR	INIIMP	FIMIMP	IRI	CEP5	CEP
33	04557	05	12	0016	1	TR		ELIESER	116	120	0	0		20930	030
33	04557	05	12	0016	2	R	CAP	FELIX	426	646	305	645		20920	310
33	04557	05	12	0016	3	R		FERREIRA DE ARAUJO	110	138	119	119		20920	350
33	04557	05	12	0016	4	R		LOPES TROVAO	0	0	15	391		20920	340
33	04557	05	12	0016	5	R	PREF	OLIMPIO DE MELO	0	0	1083	1639		20930	001
33	04557	05	12	0016	6	R		UBATINGA	4	150	0	0		20930	021

Registro: 89178 de 166010

Figura 3.1. Cadastro de Segmento de Logradouros

As bases de dados utilizados para geocodificação possuem a estrutura de SIG, contendo bancos de dados alfanuméricos e arquivos gráficos. Os bancos alfanuméricos contêm tabelas que identificam as localidades, logradouros e domicílios com seus relacionamentos, organizados de modo a permitir a exploração dos elementos e referenciá-los aos objetos dos arquivos gráficos disponíveis. Esta estrutura varia com os objetivos do projeto e a existência de cadastros. Na figura 3.2 é apresentado o esquema de relacionamentos das tabelas do G-NAF (www.g-naf.com.au).

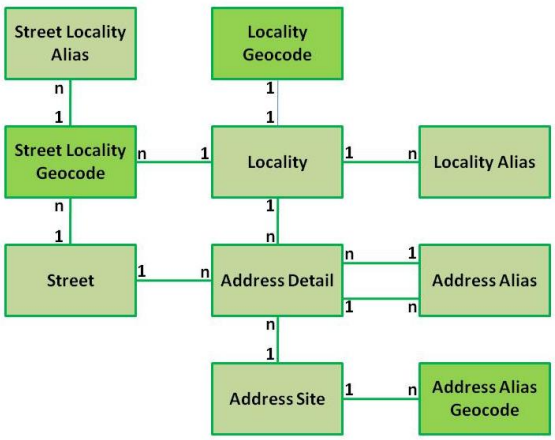


Figura 3.2. Relacionamentos do G-NAF (www.g-naf.com.au)

Os arquivos gráficos podem conter estrutura matricial ou vetorial. Os arquivos com estrutura matricial referenciam cada pixel a uma linha das tabelas associadas, enquanto nos arquivos com estrutura vetorial, os elementos gráficos (ponto, linha ou

polígono) são associados às tabelas de atributos. Nas bases lineares, compostas por polilinhas, os elementos compõem redes de logradouros (grafos). O termo rede se refere à conectividade topológica, com os nós (vértices) identificando os pontos comuns compartilhados (esquinas ou cruzamentos) e os arcos representando os trechos de logradouros. Cada trecho é identificado com seus atributos no banco alfanumérico e a localização de um domicílio pode ser feita pela interpolação a partir das numerações de início e fim do trecho. Um exemplo desta estrutura é o desenho do projeto Master Address File – MAF-TIGER (US Census Bureau 2008). Na Figura 3.3 é mostrada uma associação de endereços por interpolação no elemento linear.

Figura 3.3. Exemplo de localização de endereço por interpolação.
(www.nd.gov/gis/docs/gis-day-2004)

3.4. PROCESSOS DE COMPARAÇÃO

Nesta etapa são apresentados os processos de comparação de registros e de *strings* (conjuntos de caracteres). O método conhecido como Record Linkage, que utiliza combinação de informações para a comparação entre dados de mais de uma base de dados, é mostrado no primeiro item (3.4.1). O item seguinte (3.4.2) descreve os algoritmos mais utilizados para se avaliar a distância entre dois *strings*. As rotinas utilizadas para o método de pesquisa fonética são tratadas no item 3.4.3.

3.4.1. Record Linkage

Para a realização de estudos que permitam as avaliações de saúde a um custo reduzido, uma solução é a reutilização de dados colhidos anteriormente para novas análises, que não foram planejadas anteriormente (Pinheiro et al., 2001). "Redução de custos; Reutilização e Reciclagem", além de um uso Responsável, são os quatro Rs utilizados para este fim (Lazaridis & Mehrotra, 2004).

Fellegi e Sunter (1969) apresentaram uma teoria matemática com o objetivo de se criar uma solução computacional para o problema de comparação de registros em dois arquivos diferentes, relacionando os que tiverem alguma evidência de representarem informação de uma mesma pessoa ou evento. O resultado das comparações entre cada par de registros pode ter uma das três posições para o fato dos registros serem ou não referentes à mesma pessoa ou evento: sim, não ou talvez. Este resultado é decidido segundo um nível de erro previamente estabelecido como tolerado ou não.

O método de Record Linkage é baseado em três processos: a padronização dos registros, a blocagem dos registros e o pareamento dos registros (Jaro, 1989). A padronização tem como objetivo preparar os campos de dados para minimizar a ocorrência de erros durante o processo de pareamento (Camargo & Coeli, 2000). Alguns procedimentos são utilizados neste processo, como por exemplo, transformar os caracteres alfabéticos para maiúsculas e eliminar a pontuação, os acentos e as cedilhas.

O processo seguinte, blocagem dos registros, cria blocos lógicos de registros dentro dos arquivos a serem relacionados, com o objetivo de otimizar o processo de pareamento. Os blocos são montados no sentido de aumentar a probabilidade de os registros contidos representem pares verdadeiros. Os arquivos são indexados segundo uma chave formada por um campo ou combinação de mais de um campo. Quando a chave empregada traz informação relativa a nome, de pessoa ou de logradouro, é comum a utilização de seu código fonético na blocagem de registros. O código fonético mais empregado é o Soundex (Newcombe et al., 1959).

O terceiro processo é o pareamento dos registros e consiste na construção de escores para os diferentes pares possíveis de serem obtidos a partir de determinada estratégia de blocagem (Camargo & Coeli, 2000). A definição do conceito de escore limiar proposto por Fellegi e Sunter (1969) classifica os pares em três categorias: verdadeiros, falsos e duvidosos. Sendo verdadeiros os que apresentarem escores acima do limiar superior predeterminado e falsos quando apresentam escore abaixo do limiar inferior também predeterminado. Os outros pares são considerados como duvidosos e são revisados manualmente.

Vários aplicativos foram desenvolvidos utilizando a técnica de Record Linkage, sendo alguns fazendo parte de projetos nacionais governamentais ou acadêmicos e outros de uso comercial. Entre os de projetos nacionais, podem ser citados:

- GDRIVER – desenvolvido pelo United States Bureau of Census e que se baseia na padronização de nomes e endereços através de uma análise sintática auxiliada por diversos arquivos de referência. Esses arquivos de referência trazem correspondências entre abreviaturas do tipo st para *street* e rd para *road*, para endereços e outras correspondências entre nomes e apelidos ou abreviaturas.

- FEBRL – ou *Freely Extensible Biomedical Record Linkage (*Christen et al., 2004), consiste em outro exemplo de software que realiza remoção de ambigüidades. Para isso, realiza uma padronização de dados através de técnicas supervisionadas implementadas através de modelos escondidos de Markov (hidden Markov models) (Rabiner, 1989). Por utilizar modelos markovianos ocultos, o Febrl necessita de dados de treinamento. O Febrl utiliza métodos que possibilitam que comparações

desnecessárias entre registros possam ser descartadas.

- GRLS - ou *Generalized Record Linkage System* (Fair 2004), criado pela Agência Nacional de Estatística do Canadá, é também um exemplo de software que realiza a remoção de ambigüidades. Para isso, se baseia no método de Fellegi-Sunter (1969), agrupando registros em grupos considerados fracos ou fortes. O sistema disponibiliza uma interface gráfica ao usuário, permitindo que sejam criadas regras que contribuam para a remoção de ambigüidades.

Entre os softwares comerciais podem ser citados o Integrity XE (www.ascentialsoftware.com), que apresenta flexibilidade através da interface com o usuário e utiliza tecnologia de busca probabilística; o Trillium (www.trillium.com), que possui rotinas de separação e verificação de dados de entrada, rotinas de busca e geocodicador; e o i/Lytics (www.innovativesystems.com), com as rotinas de separação de dados e comparação de campos com rotinas definidas pelo usuário.

No Brasil, foi desenvolvido o Reclink (Camargo & Coeli, 2000), com o objetivo de relacionar as informações de pesquisas principalmente com os bancos de dados mantidos pelo Datasus, com seu grande volume de dados sobre natalidade, mortalidade e morbidade. Este software apresenta-se como um sistema de relacionamento de bases de dados fundado na técnica de relacionamento probabilístico de registros (*probabilistic record linkage*). Na figura 3.4 é apresentada a tela inicial da versão 3 do Reclink.

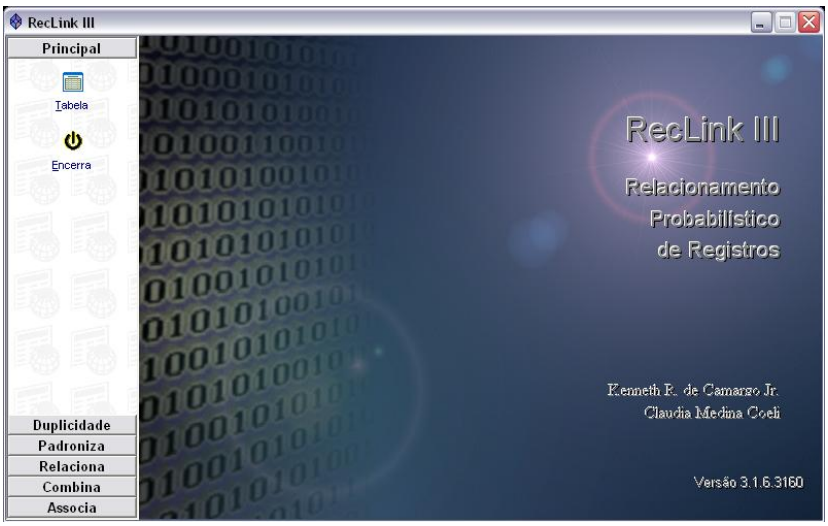


Figura 3.4. Reclink

3.4.2. Algoritmos de comparação de strings

Com o aumento de utilização da WEB e a necessidade crescente de acesso a bases de dados remotos, as rotinas de buscas tornaram-se cada vez mais utilizadas. Para o aprimoramento destas rotinas é necessária a utilização de um algoritmo de comparação de strings. Alguns dos algoritmos utilizados são descritos a seguir. Estes algoritmos recebem duas strings como parâmetros de entrada e retornam um número real, que varia de 0 a 1, sendo que o valor zero representa a falta completa de relação entre as duas strings e o valor 1 significa que os dois valores das strings são idênticos. Para os resultados diferentes, quanto mais próximo de 1, significa que os strings estão mais próximos entre si, enquanto quando estiverem mais próximos de zero, os strings são considerados mais distantes.

Algoritmo de Levenshtein

Este é um dos primeiros algoritmos de comparação de strings e dos mais utilizados. Permite avaliar inserções, remoções e substituições (Navarro, 2001). Para a montagem do algoritmop, é construída uma matriz (M), chamada de Matriz de Levenshtein, com os caracteres das strings a serem comparadas representando suas linhas e colunas. Os valores de cada célula da matriz são:

$M(0, 0) = 0;$

$M(i, j) = Max \{ M(i-1, j) – 1, M(i-1, j-1) + p(i, j), M(I, j-1) – 1\}$

Onde:

$p(i, j) = +2$ se $X_i = Y_j$ ou -1 se $Xi \neq Yj$;

X e Y são as strings a serem comparadas;

i e j são as posições dos caracteres das strings.

No quadro 3.1 é apresentado um exemplo comparando os strings “survey” e “surgery”. Observa-se que o resultado obtido neste caso foi: 8 (último escore da matriz)

/ 13 (total de caracteres dos dois strings) = 0,6154.

Quadro 3.1 – Algoritmo de Levenshtein

	&	**s**	**u**	**r**	**g**	**e**	**r**	**y**
&	**0**	-1	-2	-3	-4	-5	-6	-7
s	-1	**2**	1	0	-1	-2	-3	-4
u	-2	1	**4**	3	2	1	0	-1
r	-3	0	3	**6**	5	4	3	2
v	-4	-1	2	5	**5**	4	3	2
e	-5	-2	1	4	4	**7**	6	5
y	-6	-3	0	3	3	6	**6**	8

Smith Waterman

Apresentado por Smith e Waterman (1981), este algoritmo é muito semelhante ao de Levenshtein, substituindo o escore negativo por zero, deste modo a fórmula geral é a seguinte:

M(i, j) = Max { M(i-1, j) -1, M(i-1, j-1) + p(i, j), M(i, j-1) – 1, 0}

O resultado final é o de maior escore da matriz. Este algoritmo apresenta maiores escores do que o anterior para comparação de strings com seus pedaços como “telefone” e “fone”. No quadro 3.2, um exemplo deste algoritmo.

Quadro 3.2 – Algoritmo de Levenshtein

	&	**s**	**u**	**r**	**g**	**e**	**r**	**y**
&	0	0	0	0	0	0	0	0
s	0	2	1	0	0	0	0	0
u	0	1	4	3	2	1	0	0
r	0	0	3	6	5	4	3	2
v	0	0	2	5	5	4	3	2
e	0	0	1	4	4	7	6	5
y	0	0	0	3	3	6	6	**8**

Resultado final: 8 (maior escore da matriz) / 13 (total de caracteres dos dois

strings) = 0,6154.

Função de distância de Covington

Este método realiza comparações medindo a distância entre duas strings, considerando se o termo comparado é vogal ou consoante. É uma espécie de pesquisa fonética bruta, atribuindo pesos para os pares de segmentos, com custos para substituições, inserções ou remoções (Kondrak, 2003). Sua implementação é bastante simples, identificando 3 tipos de segmentos: consoante, vogal ou espaço. As penalidades são atribuídas aos pares, como mostrado no quadro 3.3.

Quadro 3.3. Penalidades para Distância de Covington

Termo	Penalidade
Consoantes idênticas ou espaços	0
Vogais idênticas	5
Vogais diferentes	30
Consoantes diferentes	60
Termos diferentes	100

Um exemplo deste método é apresentado abaixo com a comparação entre as palavras “paciente” e “patient”, que obtém a distância de Covington (soma das penalidades) de 125. Como o tamanho do maior string é de 8 caracteres, a maior distância seria 800, sendo a semelhança de (800 – 125) / 800 = 0,844

p	a	c	i	e	n	t	e
p	a	t	i	e	n	t	
0	5	60	5	5	0	0	50

Figura 3.5. Distância de Covington

N – grama

Um n-grama é constituído por todas as substrings geradas de uma string, sendo n o tamanho destas substrings. Um exemplo é n-grama da string “conjunto” com n igual a 3, formando o conjunto {##c, #co, con, onj, nju, jun, unt, nto, to%, o%%}.

Segundo Gandrabur e Foster (2003), o objetivo inicial deste algoritmo era de filtragem, descartando áreas onde não pode haver concordância de strings (ou palavras). No entanto pode-se aplicar este método na identificação de sequência de texto que tenham palavras em comum. A partir de dois strings, S e T, pode-se gerar os n-gramas destes e contar os n-gramas comuns. Deste modo é possível calcular a “distância n-grama”, sendo |distância n-grama| = |tamanho do maior n-grama| - |n-gamas comuns|. Como exemplo, o cálculo da distância entre as palavras “paciente” e “patient”:

N-gramas de paciente: {##p, #pa, pac, aci, cie, ien, ent, nte, te%, e%%}
N-gramas de patient: {##p, #pa, pat, ati, tie, ien, ent, nt%, t%%}
N-gramas comuns: {##p, #pa, ien, ent}

O número de n-gramas comuns é de apenas 4, e o tamanho do maior é 10, obtendo-se então: |distância n-grama| = 10 - 4 = 6. Com isso a similaridade é de 4/10 = 0,4.

3.4.3. Pesquisa fonética

A pesquisa fonética é utilizada na identificação de strings que podem ter pronúncia semelhante, mesmo com grafias diferentes. Segundo Zobel e Dart (1996), há duas questões a serem consideradas em um projeto que utilize a pesquisa fonética. Uma delas é a velocidade, as respostas devem ser conseguidas relativamente rápido. A outra questão é a precisão, que possibilita ter um número menor de respostas para um string pesquisado. Existem algumas técnicas desenvolvidas para este processo, três destes algoritmos são apresentados a seguir: Soundex, Phonix e Metaphone.

Soundex

Soundex é o algoritmos de pesquisa fonética mais conhecido. Foi criado,

desenvolvido e patenteado por Odell e Russell em 1918 (Hall & Dowling, 1980), utilizando código baseados no som de cada letra, com o objetivo de transformar o string em um código formado por no máximo quatro caracteres, preservando a primeira letra e compondo com mais até três algarismos numéricos. O algoritmo segue os seguintes passos:

- Transformar todos os caracteres, exceto a primeira letra pelo seu código fonético;
- Eliminar qualquer repetição de código adjacente;
- Eliminar todas as ocorrências de código zero (vogais);
- Retornar os primeiros quatro caracteres como resultado.

Na figura 3.6, são apresentados os códigos fonéticos soundex para as letras do alfabeto. Como exemplo, as palavras “reynold” e “renauld” apresentam o mesmo resultado, r543. Entretanto, não raro, a rotina de Soundex apresenta resultados idênticos para palavras de pronúncias não semelhantes, como “catherine” e “cotroneo” que apresentam o resultado c365. Não existe um ranking para o resultado da pesquisa, sendo apenas semelhante ou não semelhante. Os códigos foram originalmente criados considerando os fonemas das letras para o inglês, deste modo devem ser efetuadas algumas adaptações para a utilização desta rotina na língua portuguesa.

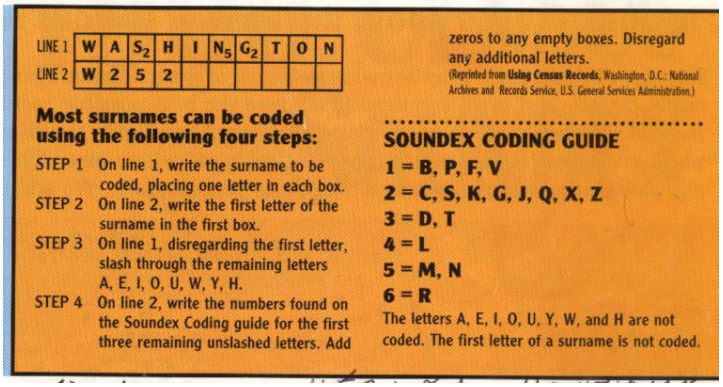


Figura 3.6. Método Soundex (freepages.history.rootsweb.ancestry.com)

Phonix

O Phonix é uma variação do Soundex (Gadd, 1990). As letras são mapeadas em um conjunto de códigos, usando o mesmo algoritmo, mas com conjunto diferente de códigos e tendo anteriormente uma transformação de caracteres, através de cerca de 160 grupos de letras e criando um string padrão de entrada. Por exemplo, a letra c é transformada em s (se anteceder e ou i), k (se preceder (a, o ou u), x (se preceder h) ou outro valor conforme sua posição na palavra. A limitação do código Phonix ou Soundex a quatro caracteres facilita a indexação destes, podendo ser alterados conforme o projeto. Na figura 3.6 são apresentados os códigos fonéticos do Phonix.

Código	Caracteres
0	a e h i o u w y
1	b p
2	c g j k q
3	d t
4	l
5	m n
6	r
7	f v
8	s x z

Figura 3.7. Códigos fonéticos do Phonix

Metaphone

Metaphone é um algoritmo de código fonético descrito por Lawrence Philips (1990). Reduz as palavras por códigos de 1 a 4 caracteres, utilizando regras fonéticas simples para o inglês falado. O Metaphone faz previamente uma transformação dos fonemas para formas padrão, como mostrado na figura 3.8.

Metaphone rushlat 16 consonant sounds:
B X S K J T F H L M N P R 0 W Y
That isn't an O but a zero - representing the 'th' sound

Metaphone uses the following transformation rules:
Doubled letters except "c" -> drop 2nd letter. Vowels are only kept when they are the first letter.

B -> B unless at the end of a word after "m" as in "dumb"
C -> X (sh) if -cia- or -ch-
    S if -ci-, -ce- or -cy-
    K otherwise, including -sch-
D -> J if in -dge-, -dgy- or -dgi-
    T otherwise
F -> F
G -> silent if in -gh- and not at end or before a vowel
      in -gn- or -gned- (also see dge etc. above)
    J if before i or e or y if not double gg
    K otherwise
H -> silent if after vowel and no vowel follows
    H otherwise
J -> J
K -> silent if after "c"
    K otherwise
L -> L
M -> M
N -> N
P -> F if before "h"
    P otherwise
Q -> K
R -> R
S -> X (sh) if before "h" or in -sio- or -sia-
    S otherwise
T -> X (sh) if -tia- or -tio-
    0 (th) if before "h"
      silent if in -tch-
    T otherwise
V -> F
W -> silent if not followed by a vowel
    W if followed by a vowel
X -> KS
Y -> silent if not followed by a vowel
    Y if followed by a vowel
Z -> S
Initial Letter Exceptions
Initial kn-, gn- pn, ae- or wr- -> drop first letter
Initial x- -> change to "s"
Initial wh- -> change to "w"

Figura 3.8. Método Metaphone – esquema de substituições

3.5. TRATAMENTO DE ENTRADA DE DADOS

Após a coleta de dados, alguns procedimentos precisam ser executados para se obter uma estrutura padronizada no texto de entrada, com o objetivo de aumentar a qualidade inicial dos dados. Esta fase é também denominada pré-processamento. Diversas técnicas podem ser aplicadas ou combinadas. A estrutura final deste processo é do tipo atributo-valor. Nos itens a seguir são descritas as técnicas de atomização (*tokenization*), remoção de palavras de acompanhamento (*stopwords*), normalização utilizando os modelos escondidos de Markov (*Hidden Markov Model – HMM*).

3.5.1. Atomização

A atomização (tokenization) tem o objetivo de extrair unidades de texto de um texto livre. Na maioria das vezes um átomo, ou token, corresponde a uma palavra, mas pode ser também um símbolo ou um caractere de pontuação ou de separação. O caractere que normalmente é utilizado como separador dos átomos é o espaço, mas outros delimitadores podem ser encontrados nos textos de entrada como: () <>!-?.;'- "|. A tarefa de identificação de átomos, que é relativamente simples para o ser humano, pode ser bastante complexa de ser executada automaticamente, já que um mesmo delimitador pode assumir outros papéis, como, por exemplo, o caractere traço (-) pode ser um separador de campos, como pode fazer parte de um campo em alguns casos, como o CEP (20270-004). Outro exemplo é o ponto, que pode ser usado como delimitador ou como determinante de abreviação. O uso de dicionários e regras de formação dos campos, a serem comparados na etapa seguinte também é útil nestas rotinas. Para exemplificar esta tarefa, considere o texto:

Rua Mariz e Barros 998/301, Tijuca CEP: 20270-004

Teria o seguinte resultado:

[Rua] [Mariz] [e] [Barros] [998] [/] [301] [,] [Tijuca] [CEP] [:] [20270] [-] [004]

Após a separação dos átomos, estes são classificados nos tipos estabelecidos por tabelas ou padrões, obtendo o resultado:

[Rua] [Mariz] [e] [Barros] [998] [/] [301] [,] [Tijuca] [CEP] [:] [20270] [-] [004]
[TP] [NM] [CJ] [NM] [NU] [PT][NU][PT] [NM] [ID] [PT] [CEP]

Onde: TP = tipo, NM = nome, CJ = conjunção, NU = número, PT = pontuação, ID = identificação de campo, CEP = padrão de Código de Endereçamento Postal.

Outras funções podem ser incorporadas a estas, facilitando as etapas seguintes. Kondrak (2003) apresenta um subsistema unindo funções e camadas, de forma similar a uma "linha de montagem" (Figura 3.9).

Figura 3.9. "linha de montagem" de atomização (adaptado de Kondrak, 2003)

3.5.2. Remoção de palavras

Em uma entrada de dados como texto livre, alguns átomos selecionados não têm valor semântico para a etapa seguinte (comparação com a base de referência), por não terem significado ou por não fazerem parte da base de referência, por uma regra pré-estabelecida. Para o reconhecimento e exclusão destes átomos, é confeccionada uma ou mais tabelas. Normalmente estas tabelas são compostas por preposições, conjunções,

pontuação, artigos e pronomes de uma língua. Aproveitando o exemplo do item anterior, obtém-se o resultado:

[Rua] [Mariz] [Barros] [998] [301] [Tijuca] [20270] [-] [004]
[TP] [NM] [NM] [NU] [NU] [NM] [CEP]

3.5.3. Padronização

A padronização é a etapa da preparação da entrada de dados. Tem por objetivo transformar os textos de entrada no formato padrão existente no banco de dados de referência. Para atingir este objetivo, no caso dos endereços, é necessário utilizar um método que identifique os elementos existentes, uma vez que, como visto no item 2.3, os endereços podem ser escritos de várias formas, sem necessariamente apresentarem todos os elementos.

Modelos Ocultos de Markov (HMM – Hidden Markov Models) são modelos probabilísticos de transições de estados, onde além da função de distribuição de probabilidades associadas aos estados, existe uma função de distribuição de probabilidades para as observações que podem ser realizadas em cada estado. Consiste em um processo duplamente estocástico composto por um processo oculto (os estados não são observáveis), mas que se manifesta através de um outro processo estocástico que produz a sequência de símbolos observados em cada estado. Os dois tipos de parâmetros a que um Modelo Oculto de Markov está associado são: probabilidades de emissão dos símbolos e probabilidades de transição de estados (Rabiner, 1989). HMM são comumente utilizados em ferramentas estatísticas de otimização para controle e reprodução de áudio, aplicativos em ciências biomédicas e bioquímicas, radares, sonares e sinais de imagens, além de predição de informações necessárias aos algoritmos de reconhecimento de frases faladas (Ephraim e Merhav, 2002) e reconhecimento de voz (Lai e Zhao 2002). Este método foi utilizado na padronização de nomes e endereços australianos em dados de saúde com grande êxito (Churches et al., 2002). Na Figura 3.10, o esquema de sequência de decisões utilizada por Churches.

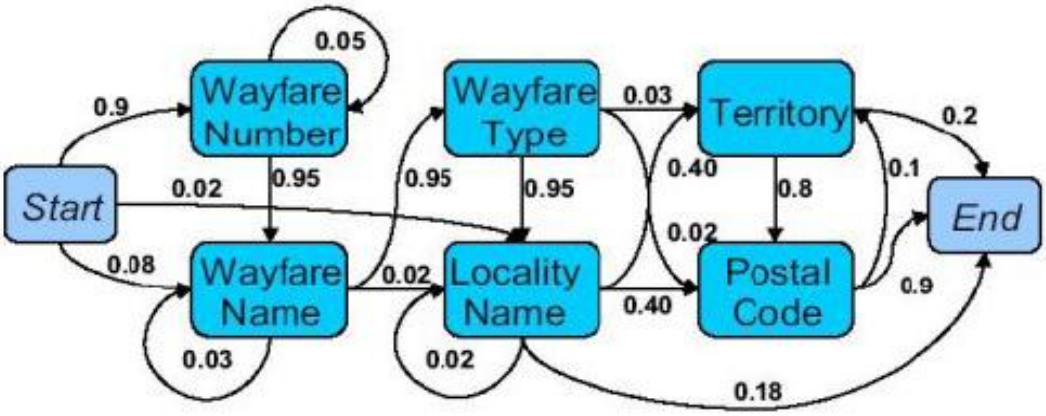


Figura 3.10. Esquema de sequência de decisões (Churches et al. 2002)

4. PROPOSTA DE GEOCODIFICAÇÃO

"A formulação do problema é, frequentemente, mais essencial que solução" (Albert Einstein)

O processo de geocodificação deve ter foco nas características existentes nos cadastros dos Sistemas de Informação em Saúde (SIS), apresentados por Skaba & Terron (2003, Anexo 2) e em experiências anteriores, como os trabalhos desenvolvidos para 5 projetos que utilizaram análise espacial na investigação de doenças, apresentados no capítulo 5 desta tese. Estes trabalhos serviram de laboratório para identificação dos problemas encontrados nas bases utilizadas e no preenchimento das informações de endereços.

A partir das investigações, algumas particularidades foram identificadas, tais como:

- Alguns nomes são informados por sua abreviaturas;
- Tipo e Título são informados de várias formas ou não são informados;
- Diversos separadores de campo foram identificados;
- Utilização de nomes alternativos como, por exemplo, nomes antigos de logradouros;
- Utilização de pontos de referência são importantes para identificação de logradouros com homônimos;
- Grande quantidade de erros na informação do CEP, criando a necessidade de checagem do nome para confirmação;
- A falta de conhecimento dos limites das unidades utilizadas como referência (bairro, RA, etc.) por parte do informante, faz com que haja necessidade de utilização de áreas vizinhas para comparação;

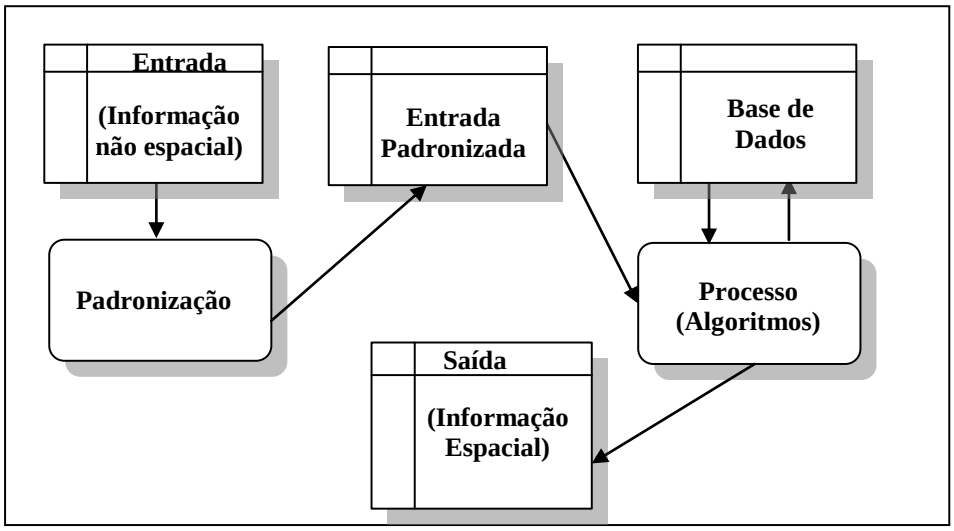


Figura 4.1. Processo de Geocodificação

As definições da base de dados, dos formatos dos registros de entrada de dados e das possíveis informações de saída estão relacionadas, criando um processo iterativo para atingir de modo satisfatório estas definições. Para se proceder a comparação entre o endereço de entrada e o da base de dados disponível, é necessário que esta entrada seja padronizada. Mas como, na maioria das vezes, isto não acontece, alguns procedimentos precisam ser feitos para esta adaptação. Deste modo, o processo geral de geocodificação se dá como mostrado na Figura 4.1. O usuário do processo deve informar alguns parâmetros a serem utilizados e o tipo de resposta que deseja. Nos itens apresentados a seguir são descritos os componentes do processo, com suas composições e suas funções:

4.1. Base de dados - composição da base de dados com seus componentes alfanuméricos, composta por um banco de dados relacional com as tabelas e seus relacionamentos, e os arquivos gráficos, compostos por elementos gráficos nos formatos utilizados pelas demandas deste tipo de informação.

4.2. Entrada de dados – características mais usuais nos dados de entrada e as funções de padronização para a procura e comparação com a base de dados, como a normalização (limpeza e separação de palavras), reconhecimento de padrões (tipo, título, número, cep) e separação nos campos.

4.3. Saída esperada – formato esperado do resultado da operação de geocodificação, constando de elementos gráficos nos formatos mais utilizados para a visualização e análise dos eventos de saúde.

4.4. Processos de comparação – processos utilizados para a procura de campos e comparação de textos, através de técnicas buscas e linkage, como blocagem, fonética entre outras.

4.5. Medidas de qualidade – medidas utilizadas para avaliar a qualidade dos processos frente aos dados de entrada.

4.6. Modelo final – modelo final da proposta de geocodificação, combinando os resultados dos quatro itens anteriores.

4.1. BASE DE DADOS

A base de dados para dar suporte aos processos de geocodificação deve possuir estrutura de SIG, com um banco de dados alfanumérico relacional, contando com os cadastros e relacionamentos para as consultas, e os arquivos gráficos associados a estes cadastros (Figura 4.2).

De acordo com o desenvolvimento desta proposta, as tabelas do banco de dados relacional estão divididas em 5 grupos. O primeiro grupo é composto pelas tabelas que armazenam os endereços, o segundo, pelos cadastros que contêm as informações das áreas ou pontos de interesse para o referenciamento. O terceiro grupo contem as tabelas auxiliares com as relações utilizadas para auxiliar nas rotinas de padronização de entrada de dados e de comparação, enquanto o quarto engloba os relacionamentos entre as tabelas dos três primeiros grupos e representadas pelas setas na Figura 4.2. No último grupo existem duas tabelas que são utilizadas para controle, com a identificação de todas as tabelas, separando as que definem entidades das de relacionamentos.

As tabelas dos grupos de endereços, de controle e auxiliares possuem estruturas fixas, alterando apenas os conteúdos, enquanto as de cadastros e de relacionamentos dependem das disponibilidades de cada projeto. Os arquivos gráficos possuem, para cada elemento, uma chave que relaciona este elemento a uma linha de uma tabela do grupo de cadastro ou diretamente a um trecho de logradouro ou a um endereço. Há vários meios para a construção desta base de dados. Um levantamento total com elaboração do banco de dados e dos arquivos gráficos é, além de muito trabalhosa, bem onerosa, sendo mais viável a obtenção de dados já disponíveis, como os cadastros de censos e pesquisas, e bases gráficas de concessionárias de serviços públicos ou órgãos governamentais. Nesta tese foram escolhidos os produtos do Censo 2000 do IBGE (Skaba & Terron, 2003).

A seguir são apresentadas as tabelas por grupo, com a especificação das variáveis e seus índices, com uma descrição de suas aplicações e importância. Ao final, são feitas algumas considerações sobre os arquivos gráficos.

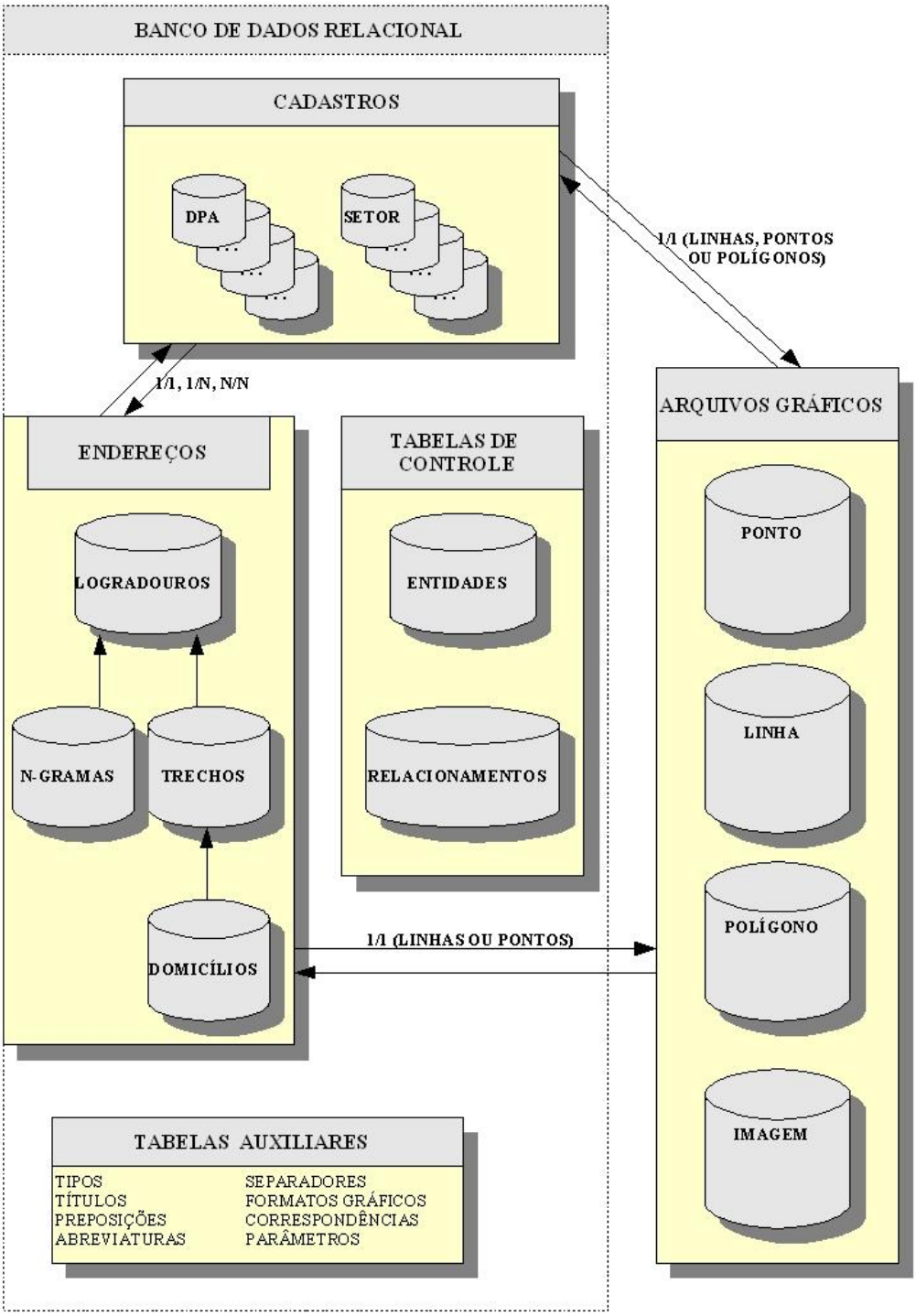


Figura 4.2. Esquema da Base de Dados de Referência

4.1.1. Tabelas de Endereço

As tabelas centrais do banco de dados são as de logradouros e de trechos destes logradouros, a partir das quais são feitas as primeiras pesquisas de comparação com os registros de entrada. Uma terceira tabela apresenta os números de porta de cada trecho. Nos quadros 4.1, 4.2 e 4.3 são apresentadas as variáveis destas tabelas com as respectivas descrições.

Quadro 4.1. Tabela de Logradouros

Variável	Formato	Tamanho	Descrição
Chave_Lograd	Numérico	9	Chave primária da tabela
UF_Munic	Numérico	7	Código de município do IBGE
Nome_completo	Caracter	60	Nome completo do logradouro
Tipo	Numérico	3	Código do tipo do logradouro
Título	Numérico	3	Código de título no nome do logradouro (se não houver = 0)
Nome	Caracter	60	Nome do logradouro, sem tipo, título e preposições
NomeF	Caracter	20	Forma fonética do nome (soundex)
CEP_ini	Numérico	8	CEP inicial do logradouro
CEP_fim	Numérico	8	CEP final do logradouro

Quadro 4.2. Tabela de Trechos de Logradouro

Variável	Formato	Tamanho	Descrição
Chave_trecho	Numérico	9	Chave primária da tabela
Chave_Lograd	Numérico	9	Chave da tabela de logradouros
Seq_Lograd	Numérico	3	Sequencial do trecho no logradouro
CEP	Numérico	8	Código de endereçamento postal
Ini_Imp	Numérico	5	Número ímpar inicial
Fim_Imp	Numérico	5	Número ímpar final
Ini_Par	Numérico	5	Número par inicial
Fim_Par	Numérico	5	Número par final

Na tabela de logradouros é atribuída uma chave primária com código único e, além do nome completo do logradouro (ex. Rua Nossa Senhora de Copacabana), ela contém a separação dos elementos deste, com os campos de códigos de tipo e título e o

nome padronizado, com letras maiúsculas e sem acentos nem cedilhas ou preposições. No exemplo citado acima, tipo = "RUA", título = "NOSSA SENHORA" e nome = "COPACABANA". Os campos de tipo e título recebem um código numérico. Esta tabela recebe também o código *Soundex* (Zobel & Dart 1996) do nome, para possibilitar a pesquisa fonética, além dos CEP inicial e final que, na maioria dos casos, são iguais.

Quadro 4.3. Tabela de Localização de Domicílios

Variável	Formato	Tamanho	Descrição
Chave_domic	Numérico	9	Chave primária da tabela
Chave_trecho	Numérico	9	Chave da tabela de trechos
Numero	Numérico	5	Número de porta
Quant	Numérico	3	Quantidade de domicílios

Associada à tabela de logradouros, possuindo um ou mais registros para cada logradouro, a tabela de trechos de logradouros apresenta os trechos divididos conforme o tipo de informação existente nos cadastros ou arquivos gráficos, como por exemplo, trechos por cada cruzamento ou trechos dentro do setor censitário. Nesta tabela há informação de CEP e numeração inicial e final por cada lado do logradouro, sendo par ou ímpar. Para ligar os endereços às coordenadas geográficas, utiliza-se a tabela de localização de endereços, com uma associação aos trechos de logradouros. Esta tabela possui informações de número de porta e quantidade de domicílios existentes na edificação. Para o teste de aproximação de textos por n-gramas (item 3.4.2.), foi projetada uma tabela com um registro por cada n-grama do logradouro, composto pela chave do logradouro e o n-grama associado. As tabelas de logradouros, trechos, domicílios e dos n-gamas estão relacionadas pelas chaves de cada uma (Figura 4.3). Na Figura 4.4 são apresentados os mapas representando dois modos de relacionamento dos trechos dos logradouros, com vistas a uma operação de coleta do tipo de um Censo. O primeiro (A), por face de quadra, com a identificação das quadras dentro de cada Setor Censitário e, em cada quadra, são identificados os trechos de logradouros que a compõem. No segundo caso (B), são identificados os trechos de logradouros que estão inscritos no Setor Censitário, sem dividi-lo em cada quadra. Nos dois casos, os números

de porta inicial e final são cadastrados, sendo que, no primeiro caso, é representado apenas um lado do logradouro, enquanto no segundo há os dois lados em algumas situações.

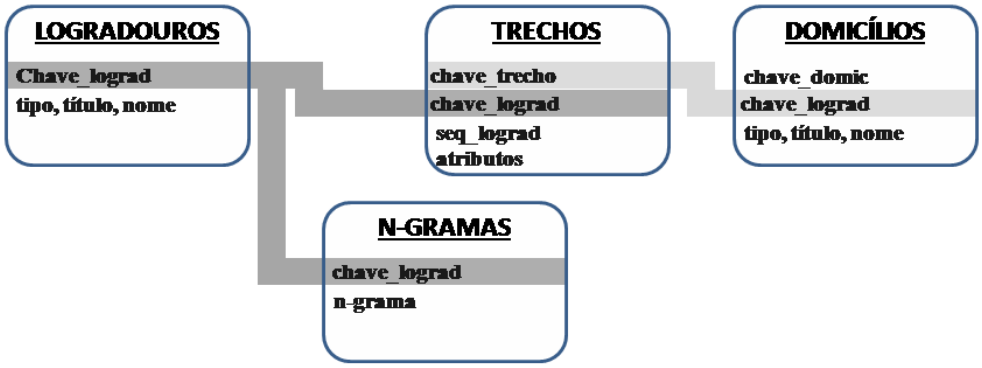
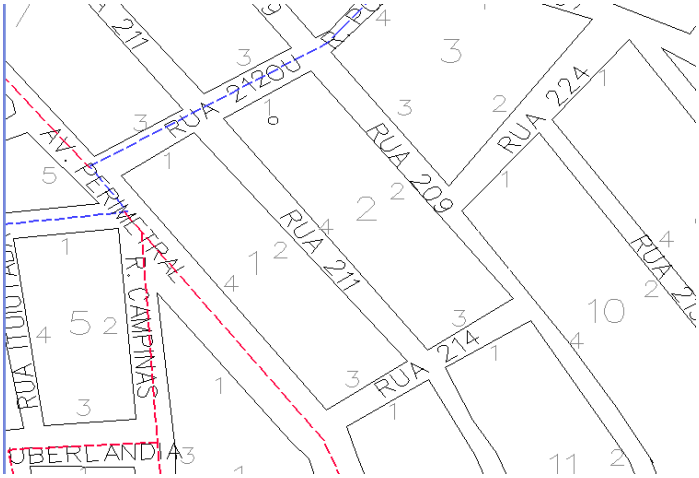


Figura 4.3. Relacionamentos entre as tabelas



A – por Face de Quadra

B – por Setor Censitário

Figura 4.4. Exemplos de determinação de trechos de logradouros (Fonte: IBGE)

4.1.2. Cadastros Associados

Este grupo é composto pelos cadastros utilizados na localização dos endereços, segundo a disponibilidade ou desenho do projeto. Eles são compostos por áreas intramunicipias, pontos de referência ou até tabelas de trechos ou endereços com atributos de interesse das pesquisas. Os cadastros mais comuns são os de bairros, regiões administrativas, setores censitários, áreas de saúde ou agregações destas áreas, além dos pontos de referência associados. O quadro 4.4 mostra um exemplo da definição de uma tabela de cadastro, tendo como unidade o setor censitário 2000. As tabelas deste grupo têm como objetivo a associação aos arquivos gráficos, através de uma chave comum.

Quadro 4.4. Tabela de Cadastro (Setor Censitário 2000)

Variável	Formato	Tamanho	Descrição
Chave_setor	Numérico	9	Chave primária da tabela
Setor2000	Numérico	15	Chave do setor censitário 2000
Pop2000	Numérico	5	População do setor
Domic2000	Numérico	4	Número de domicílios ocupados em 2000

4.1.3. Tabelas Auxiliares

As tabelas auxiliares são utilizadas nas rotinas de normalização de entrada e de comparação da entrada padronizada com as tabelas do grupo de endereços. Cada uma tem um objetivo definido como descrito abaixo.

- Tipos – é composta pelos tipos de logradouros (ex. Rua, avenida, rodovia, beco) utilizados nos endereços cadastrados.

- Títulos – contém os títulos existentes nos nomes de logradouros (ex. Presidente, padre, dona, princesa) e tem o objetivo de isolar o nome.

- Preposições – comporta as preposições que devem ser retiradas do nome do logradouro. Ex: de, do, da, e.

- Abreviaturas – relação das abreviaturas encontradas nos registros de entrada de dados para o processo de geocodificação. Ex. R (rua), NS (nossa senhora), AV (avenida), JK (Juscelino Kubitschek).

- Separadores – caracteres especiais e palavras que se caracterizam como separadores de campos. Ex. /, ;, perto, ao lado, entre.

- Parâmetros – parâmetros utilizados para definir as comparações e decisões na padronização dos campos.

- Formatos gráficos – com definição dos formatos utilizados da base de dados.

As tabelas de tipo e título possuem um código numérico como chave, enquanto as abreviaturas possuem a abreviatura encontrada e o extenso correspondente. As preposições e separadores são apenas relações com campo único.

No Anexo 1 são apresentadas as tabelas de tipo e título utilizadas pelo IBGE no projeto CNEFE (IBGE, 2005) e as relações de preposições e separadores selecionados. Estas tabelas devem ser revistas a cada projeto, segundo necessidades destes.

4.1.4. Relacionamentos

Ao serem definidos os relacionamentos entre as entidades do projeto ou entre uma entidade e um arquivo gráfico, são criadas tabelas com a definição destas relações. O objetivo destes relacionamentos é a associação de endereços ou trechos a um objeto significante para o projeto desenvolvido ou a elementos gráficos definidos. Na Figura 4.5 há dois exemplos de esquema de relacionamentos, sendo um com uma entidade, neste caso o setor censitário, e o outro ligando diretamente a um arquivo gráfico com a representação de eixos de rua.

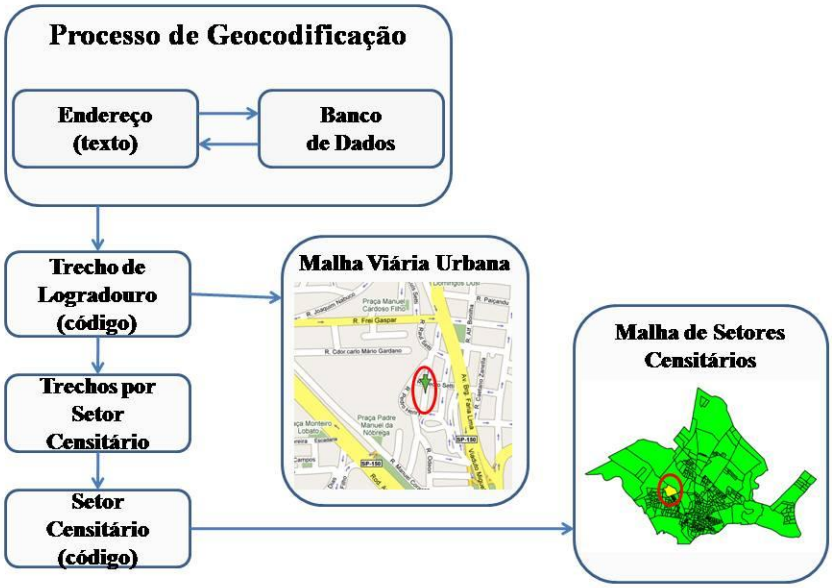


Figura 4.5. Exemplos de esquemas de relacionamento

4.1.5. Tabelas de Controle

Com o objetivo de registrar as tabelas e relacionamentos existentes no banco de dados, foram projetadas três tabelas chamadas de “Entidades”, “Relacionamentos” e “ArqGraficos”. Estas tabelas, além do registro de existência, servem de orientação para as rotinas desenvolvidas. A tabela “Entidades” (quadro 4.5) define os cadastros existentes, seus campos de chave primária e atributos, enquanto a “ArqGraficos” (quadro 4.6) contém os arquivos gráficos existentes, com a especificação de tipo de arquivo, tipo de elemento gráfico e o campo chave. Na tabela “Relacionamentos” (quadro 4.7) estão representadas as tabelas de relacionamentos entre tabelas dos três grupos: endereços, cadastros e arquivos gráficos.

Quadro 4.5. Tabela Entidades

Variável	Formato	Tamanho	Descrição
Chave_entidade	Numérico	9	Chave primária da tabela
Nome	Texto	25	Nome da entidade
Descrição	Texto	100	Descrição da entidade

Quadro 4.6. Tabela ArqGraficos

Variável	Formato	Tamanho	Descrição
Chave_grafico	Numérico	9	Chave primária da tabela
Nome	Texto	25	Nome do nível
Nome_arquivo	Texto	40	Nome do arquivo gráfico
Formato	Numérico	9	Chave_formato
Tipo_elemento	Numérico	1	1=ponto, 2=linha, 3=polígono, 4=pixel

Quadro 4.7. Tabela Relacionamentos

Variável	Formato	Tamanho	Descrição
Chave_relacionamento	Numérico	9	Chave primária da tabela
Chave_tabela1	Numérico	9	Chave primária da primeira tabela
Chave_tabela2	Numérico	9	Chave primária da segunda tabela
Tipo_relacionamento	Numérico	1	1=1/1, 2=1/N, 3=N/N

4.1.6. Arquivos Gráficos

Os arquivos gráficos utilizados para a associação dos endereços geocodificados, possuem uma estrutura com elementos gráficos, contendo identificação e atributos associados. A identificação ou algum atributo possui uma relação com uma tabela de cadastro ou de endereço. Os formatos possíveis destes arquivos estão descritos em uma tabela auxiliar. O formato padrão de armazenamento escolhido é o Shape (ESRI 1998), por ter um formato aberto e ser de fácil manuseio, além de ser aceito em vários aplicativos. Para os resultados da geocodificação, outro formato utilizado é o KML (Google Maps 2009), por sua ampla utilização e disponibilidade de *software*.

4.2. ENTRADA DE DADOS

O elemento utilizado para a entrada de dados neste processo é o endereço residencial. Com referência aos tipos de endereço apresentados na seção 2.3.1, neste estudo, foi dado ênfase ao endereço urbano com logradouros (não considerando os de Brasília), que correspondem a mais de 80% dos domicílios brasileiros, segundo o IBGE (2002). Um endereço residencial típico descreve uma localização em termos de uma posição (número, complemento) em um logradouro. Este formato de endereço pode ser descrito como consistindo de um número de atributos que, quando se agregam, identificam um local.

Para um melhor aproveitamento na comparação com os endereços que constam da base de dados apresentada no item anterior, o ideal seria que os dados de entrada tivessem o mesmo formato, com os campos informados separadamente. Entretanto, a maior parte dos registros de endereços encontrados nas fontes de informação utilizadas nos projetos na área da saúde tem uma forma textual livre. No Quadro 4.8, estão alguns exemplos de endereço residencial encontrados nestas fontes de dados. Para o desenvolvimento do projeto, os exemplos são apresentados prevendo a situação mais complexa, ou seja, os endereços são apresentados com o formato de um campo texto único.

Quadro 4.8. Endereços Residenciais

	Endereços
1	Rua Maris e Barros, 1052/503, Tijuca, RJ, 20270-004
2	Estrada dos Bandeirantes, km 4 casa 3, São Paulo
3	Avenida Rio Branco 156 apto 101, Centro, Porto Alegre, CEP 71000-310
4	R Projetada c/ 25 Meier 20000150
5	Rua JK 232, Pelotas, RS
6	Av. Pres. Jucelino Kubitchek de Oliveira 503 – Pelotas – RS

Nestes exemplos, pode-se notar algumas características comuns como o uso de abreviaturas no nome e no tipo; o preenchimento do CEP não constante, sem padrão e algumas vezes não correto, utilização de separadores de campos não padronizados (vírgula, traço, barra); falta de padrão no uso de referências, como bairro, cidade, UF.

No Brasil, os componentes mais comuns de figurarem nas informações de endereço são o logradouro e o número do prédio. No entanto, bairro, cidade e estado também aparecem com freqüência. O CEP é mais comum de ser preenchido nos endereços utilizados para correspondência, por ser requisito básico para operação pelos Correios (www.correios.com.br), enquanto nos sistemas de saúde (www.datasus.gov.br), o município é sempre codificado e especificado em campo próprio.

Para atingir um grau maior de acertos no processo geral de geocodificação, é necessário tratar o endereço informado com o objetivo de obter o formato dos dados armazenados no banco, com o maior número possível de campos preenchidos. Este tratamento pode ser dividido em três funções básicas, sendo:

- Normalização – transformação dos caracteres em formato padrão;
- Separação e identificação – identificação dos conjuntos de caracteres (strings) de uma forma siginificativa e;
- Padronização – preenchimento dos campos padrões para comparação com o banco de dados.

4.2.1. Normalização

Esta fase tem como objetivo fazer o tratamento de cada caractere, transformando-os, quando necessário, em formas identificáveis na formação dos conjuntos de caracteres. Os procedimentos a serem realizados são:

- transformação dos caracteres alfabéticos em maiúsculas;
- transformação de Ç em C;
- exclusão de acentuação gráfica (~, ^, ´, ` e ¨);
- transformação dos caracteres não alfanuméricos e que não sejam utilizados como separadores de campo (vírgula, traço ou barra) ou indicador de

abreviatura (ponto ou barra) em espaço simples, enquanto os separadores de campos ganham um espaço;

- eliminação dos espaços duplos.

Como exemplo da execução desta fase, o resultado da transformação do Quadro 4.8 é apresentado no Quadro 4.9.

Quadro 4.9. Endereços Residenciais Normalizados

	Endereços
1	RUA MARIS E BARROS, 1052 / 503 , TIJUCA , RJ , 20270-004
2	ESTRADA DOS BANDEIRANTES , KM 4 CASA 3 , SAO PAULO
3	AVENIDA RIO BRANCO 156 APTO 101 , CENTRO , PORTO ALEGRE , CEP 71000 - 310
4	R PROJETADA C/ 25 MEIER 20000150
5	RUA JK 232, PELOTAS, RS
6	AV. PRES. JUCELINO KUBITCHEK DE OLIVEIRA 503 – PELOTAS – RS

4.2.2. Separação e identificação

Após a normalização, os conjuntos de caracteres são separados pelos espaços e classificados a partir de critérios definidos. No Quadro 4.10 são apresentadas as classes definidas para este primeiro passo.

Quadro 4.10. Códigos para classificação de conjunto de caracteres

Código	**Descrição**
PA	Palavra
Nx	Número com x dígitos
SP	Separador
PR	Preposição

Depois desta classificação, algumas identificações são feitas, tais como, preposição, abreviatura, sigla de UF e CEP. As ações a serem efetuadas a partir destas identificações são: eliminar as preposições, separar CEP e UF e expandir as abreviaturas. No Quadro 4.11, o resultado desta operação para o exemplo seguido neste desenvolvimento.

Quadro 4.11. Resultado da operação separação e identificação

1		2		3		4		5		6	
RUA	[PA]	ESTRADA	[PA]	AVENIDA	[PA]	R	[PA]	RUA	[PA]	AV.	[PA]
MARIZ	[PA]	DOS	[PR]	RIO	[PA]	PROJETADA	[PA]	JK	[PA]	PRES.	[PA]
E	[PR]	BANDEIRANTES	[PA]	BRANCO	[PA]	C	[PA]	232	[N3]	JUCELINO	[PA]
BARROS	[PA]	,	[SP]	156	[N3]	/	[SP]	,	[SP]	KUBITCHEK	[PA]
,	[SP]	KM	[PA]	APTO	[PA]	25	[N2]	PELOTAS	[PA]	DE	[PR]
1052	[N4]	4	[N1]	101	[N3]	MEIER	[PA]	,	[SP]	OLIVEIRA	[PA]
/	[SP]	CASA	[PA]	,	[SP]	20000150	[N8]	RS	[PA]	503	[N3]
503	[N3]	3	[N1]	CENTRO	[PA]					-	[SP]
,	[SP]	,	[SP]	,	[SP]					PELOTAS	[PA]
TIJUCA	[PA]	SÃO	[PA]	PORTO	[PA]					-	[SP]
,	[SP]	PAULO	[PA]	ALEGRE	[PA]					RS	[PA]
RJ	[PA]			,	[SP]						
,	[SP]			CEP	[PA]						
20270	[N5]			71000	[N5]						
-	[SP]			-	[SP]						
004	[N3]			310	[N3]						

4.2.3. Padronização

Nesta fase são identificados os possíveis elementos existentes no texto de entrada, seguindo o padrão utilizado no banco de dados: Tipo, título, nome, número, complemento, referência, cidade e UF, a partir dos elementos separados e identificados no item anterior.

Para esta pesquisa de elementos, é utilizada o método de Cadeia Escondida de Markov (HMM), que trabalha com a proporção de alternativa de caminhos, conhecida a priori. Para este método, os parâmetros foram escolhidos tendo como base a proporção de ocorrência de títulos no cadastro de endereços do estado do RJ (IBGE) e a ocorrência de tipo e outros elementos nos estudos de casos apresentados no capítulo 5. A partir destes parâmetros foi montado o diagrama (Figura 4.6).

Antes de seguir o fluxo do diagrama apresentado, alguns procedimentos são executados:

- Identificação da abreviaturas;
- Identificação dos elementos padrão como a sigla da UF (tabela);
- Identificação de CEP (8 dígitos numéricos ou sequência de 5 dígitos numéricos, um hífen e outros 3 digitos numéricos;
- Eliminação das preposições (já identificadas);
- Identificação dos separadores de campos;

- Determinação de identificadores de complementos.

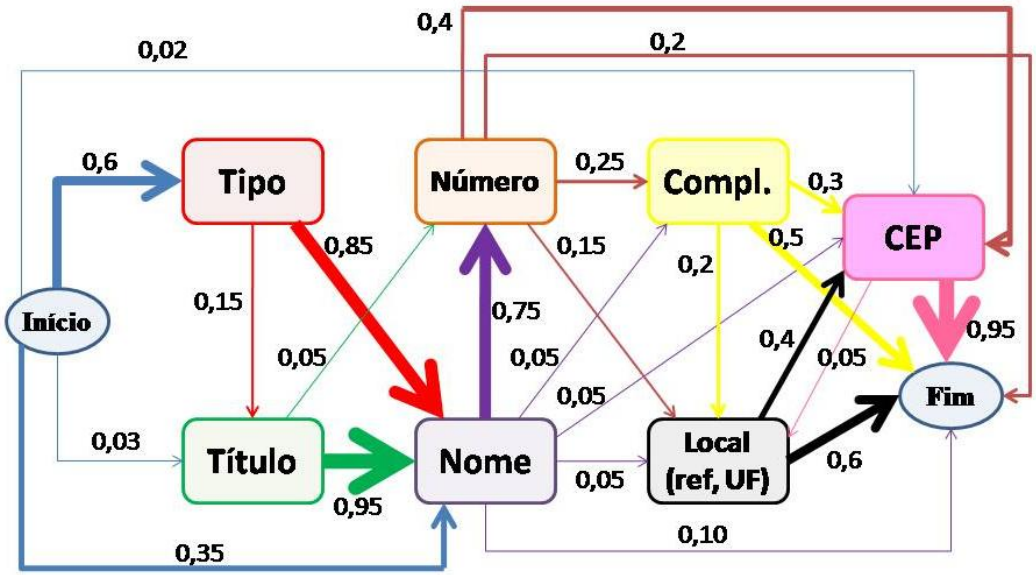


Figura 4.6. Diagrama HMM para endereços do RJ

Os percentuais do diagrama definem a prioridade para a determinação dos campos. Como resultado da padronização, os elementos utilizados na próxima etapa (comparação) se apresentam como no quadro 4.12.

Quadro 4.12 – Endereços padronizados

Tipo	Título	Nome	Núm.	Compl.	Refer.	Cidade	UF	CEP
RUA		MARIZ BARROS	1052	503	TJUCA		RJ	20270004
ESTRADA		BANDEIRANTES	4	CASA 3		SAO PAULO		
AVENIDA		RIO BRANCO	156	APTO 101	CENTRO	PORTO ALEGRE		71000310
RUA		PROJETADA	25		MEIER			20000150
RUA		JUSCELINO KUBITSCHEK	232			PELOTAS	RS	
AVENIDA	PRESIDENTE	JUSCELINO KUBITSCHEK OLIVEIRA	503			PELOTAS	RS	

4.3. RESULTADO DA GEOCODIFICAÇÃO

O resultado da operação de geocodificação é composto por três produtos: um texto (TXT), com a posição obtida no processo; uma tabela (DBF) com os possíveis resultados; e um arquivo gráfico (SHP ou KML), com a localização e forma deste resultado.

O arquivo texto contem um relatório com as informações relevantes para cada elemento de entrada (Figura 4.7), tais como:

- Identificação do endereço de entrada
- Quantidade de possíveis resultados
- Posição de concordância – representado pela posição de saída no fluxo da Figura 4.x (item 4.4)
- Tipo do elemento

PROCESSO DE GEOCODIFICAÇÃO
RESULTADOS

IDENTIFICAÇÃO	N.RESULT.	POSIÇÃO	TIPO ELEM.
99990001	3	42	Linha
99990002	0	50	
99990003	1	10	Polígono
99990004	1	31	Ponto

Figura 4.7. Relatório dos resultados da Geocodificação

A tabela dos resultados possibilita ao usuário ter as informações dos elementos gráficos obtidos no processo de digitalização, com o conteúdo de cada ocorrência de saída referente aos endereços de entrada, em formato DBF, a saber:

- Código do endereço de entrada

- Sequencial do resultado
- Nome do arquivo gráfico
- Código do elemento gráfico
- Tipo do elemento gráfico
- Coordenadas do retângulo envolvente

Este arquivo pode ser utilizado para uma pesquisa manual e escolha do resultado mais apropriado.

O terceiro produto da geocodificação é o arquivo ou arquivos gráficos com os elementos identificados na tabela descrita acima. O formato destes arquivos pode ser SHP ou KML, de acordo com a necessidade de utilização do usuário.

4.4. PROCESSO DE COMPARAÇÃO

Nesta etapa do processo são comparadas a entrada de dados padronizada e as informações do banco de dados relacional. Para esta comparação, os elementos centrais são o município, o nome do logradouro e o CEP, enquanto os outros elementos são utilizados para a resolução da ocorrência de múltiplos resultados, como um filtro mais fino. Após a constatação de ocorrência ou não de concordância de nome ou CEP para o município, é seguida uma sequência lógica de comparações, com o objetivo de obtenção do melhor resultado possível. Esta sequência é mostrada na Figura 4.8. As técnicas de record linkage, com a combinação de campos para a composição de índices e comparação de registros, e de pesquisa fonética (soundex e metaphone) e n-gramas, para a comparação de palavras, descritas no item 3.4, são utilizadas nas rotinas deste processo.

Algumas observações importantes podem ser notadas:

- A simples existência de um CEP válido não é suficiente para servir como resultado, pois como pode ser visto no item 5.4, é comum o erro de preenchimento deste campo. Neste caso é utilizado o nome e suas variações (forma fonética e n-gramas) para a confirmação.

- Como resultado da busca, pode-se obter uma ocorrência, nenhuma ou mais de uma. A forma de apresentação deste resultado está especificada no item 4.3.

- Quando o resultado de apenas uma ocorrência é obtido com poucos testes, os outros campos que completam o endereço são testados com o objetivo de avaliar a qualidade deste resultado (item 4.5). Para isto, são utilizados os números de saída do fluxo como indicadores.

- Para o teste de nome, quando nenhuma ocorrência é encontrada, são utilizadas suas variações como forma fonética ou testes com grupos de caracteres (N-gramas).

- Na pesquisa em tabelas que servem de referências com representação de área, são pesquisados, além das unidades onde se encontram os possíveis endereços, seus vizinhos ou outras unidades com relacionamento de importância hierárquica. Como exemplo, pode-se observar os bairros, para os quais, em muitas vezes, não há um nível de informação exato (O Globo 2009), sendo informado um bairro vizinho ou o mais conhecido da região.

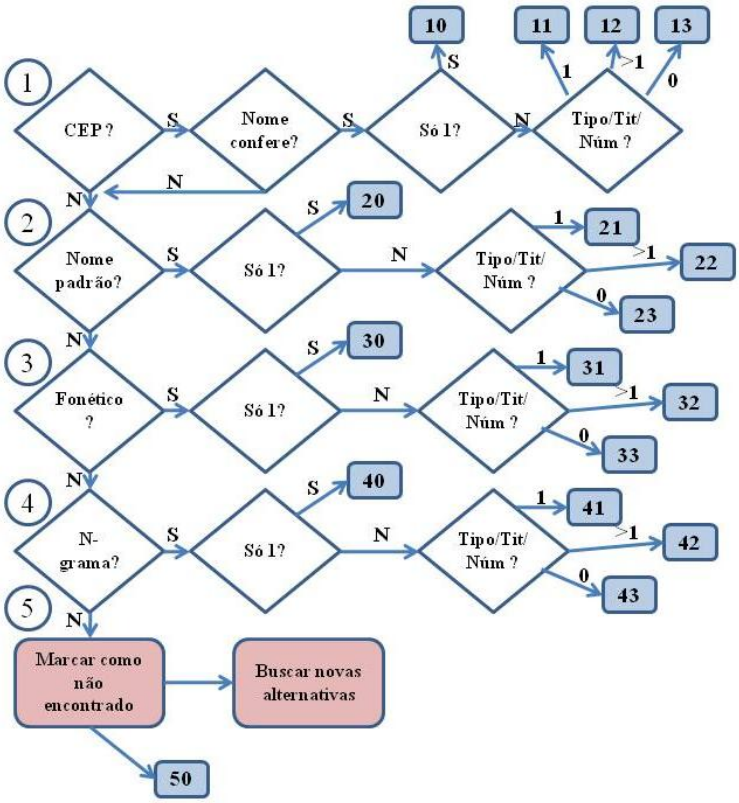


Figura 4.8. Fluxo de decisões para comparação

4.5. MEDIDAS DE QUALIDADE

Para medir a qualidade do processo de geocodificação, são utilizados dois tipos de avaliação. A primeira visa avaliar o processo e os dados de entrada em termos do resultado deste, a partir de três medidas: eficácia, eficiência e precisão.

- Eficácia é representada pela determinação ou não do setor censitário onde está localizado o endereço pesquisado. Possui valores 0 (não localizado) ou 1 (localizado) para cada registro e de 0 a 100%, dependendo das perdas de localização em uma base de dados.

- Eficiência é representada pela velocidade em que o setor é localizado. Escala baseada na seqüência de pesquisa apresentada no item 4.4. A eficiência é máxima quando o processo é realizado automaticamente. A eficiência é menor quando o processo de geocodificação exige a correção de endereços e o trabalho do técnico e da máquina.

- Precisão é representada pela distância entre a localização real de um evento e a obtida pelo sistema de georreferenciamento. Quando a unidade utilizada é pontual, utiliza-se a distância absoluta. Entretanto, para as linhas, pode-se utilizar o ponto central ou um ponto interpolado, a partir da numeração do logradouro representado. Para as áreas, é escolhido um ponto interno da unidade encontrada como resultado, podendo ser seu centróide. Do mesmo modo, quando o arquivo gráfico é composto por imagens, deve-se usar o ponto médio do pixel encontrado.

Outra avaliação de qualidade do processo pode ser obtida calculando a sensibilidade, fração dos que obtiveram resposta positiva entre aqueles bem definidos, e

a especificidade, fração dos endereços que não foram encontrados entre os que não têm endereço bem definido. Para os cálculos destas medidas assim como das de precisão é utilizado um padrão-ouro que, neste caso, deve ser obtido a partir da medição das coordenadas dos locais com utilização de GPS, em visitas aos locais dos endereços.

4.6. MODELO FINAL

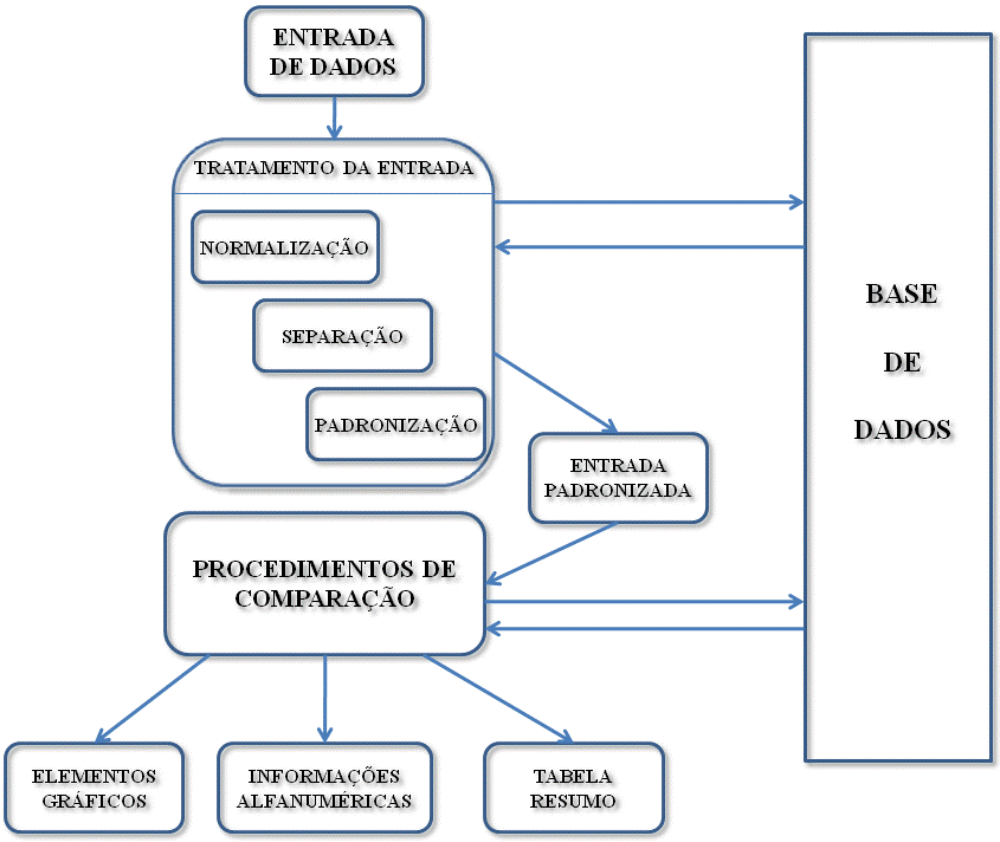


Figura 4.9. Modelo Final

A Figura 4.9 apresenta a sequência simplificada dos procedimentos descritos nos itens anteriores. Esta sequência tem início com o tratamento dos dados de entrada, com interação com as tabelas da base de dados e, a partir da entrada padronizada são efetuadas as comparações com os cadastros da base, resultando nos arquivos gráficos e tabelas de resultados.

5. ESTUDO DE CASOS

"Há verdadeiramente duas coisas diferentes: saber e crer que se sabe. A ciência consiste em saber; em crer que se sabe reside a ignorância" (Hipócrates)

Na busca por definição dos aspectos relevantes ao processo de geocodificação, alguns trabalhos desenvolvidos com utilização desta técnica foram analisados. Neste capítulo são apresentados 5 destes trabalhos com suas características próprias e que revelam diferentes respostas às questões apresentadas.

Nos dois primeiros trabalhos foram pesquisadas áreas predominantemente de ocupação desordenada, como favelas ou periferias, em estudos das ocorrências de leptospirose, sendo o primeiro no município do Rio de Janeiro e o outro em Salvador. Nestes dois estudos os maiores problemas ocorreram no fato destas áreas, em alguns pontos, não terem endereço definido. O terceiro trabalho se refere a uma coorte de nascimento, tendo como característica uma defasagem de tempo entre a época de ocorrência dos eventos (nascimentos) e a data de referência das bases de dados utilizadas na geocodificação, havendo ainda uma mudança na divisão político-administrativa entre estas duas épocas no município pesquisado, Pelotas. O quarto trabalho foi de exploração, realizado na pesquisa de localização dos endereços dos 500 primeiros participantes do projeto de Estudo Longitudinal de Saúde dos Adultos (ELSA Brasil) no Rio de Janeiro, compreendendo a região metropolitana do Rio de Janeiro e com o objetivo de localização a partir do CEP informado. O último trabalho analisado focaliza a evolução dos acessos à informação e utilização da Internet neste trabalho de geocodificação direta pelo usuário final, como uma alternativa viável.

5.1. LEPTOSPIROSE EM SALVADOR

A leptospirose teve um aumento de notificação em vários países do mundo nos últimos anos (Nunes, 2007). Os roedores desempenham o papel de principais reservatórios da doença, sendo as áreas de ocupação urbana desordenada, como as favelas, as de maior risco, por apresentarem condições sanitárias precárias, com falta de tratamento de esgoto em sua maioria. Com o objetivo principal de analisar a distribuição espacial da leptospirose, na cidade de Salvador, no período de 1996 a 2006 e caracterizá-la segundo fatores sócio-econômicos e ambientais, Nunes (2007) investigou os casos de leptospirose em Salvador, pesquisando os 1762 eventos confirmados ou suspeitos referentes ao município, ocorridos no período de março de 1996 a março de 2006. Os dados para este estudo são provenientes da Vigilância Epidemiológica da Leptospirose (LVIGI). Entre os objetivos específicos relatados na dissertação, destacam-se: georreferenciar os casos de leptospirose em Salvador neste período e descrever a distribuição destes casos utilizando o método de suavização kernel nos períodos secos e de chuvas.

Para atingir estes objetivos foi feita a geocodificação utilizando-se o cadastro de segmentos de logradouros do Censo 2000 (Cadlog), que relaciona os logradouros com numeração de início e fim aos setores censitários correspondentes, segundo o Censo Demográfico 2000, realizado pelo IBGE. Para esta geocodificação, foram seguidas as seguintes etapas:

- Formatação dos endereços – para compatibilização com o Cadlog, separando o endereço nos campos: Tipo do logradouro (rua, avenida, rodovia, ...), Título (presidente, padre, coronel, ...), Nome do logradouro, Número de porta;

- Busca automática ao Cadlog – esta busca segue uma sequência lógica, da mais completa para as menos completas, dando-se prioridade aos campos de melhor preenchimento segundo experiências anteriores. A pesquisa passa ao passo posterior quando não encontrado no atual. A Figura 5.1 apresenta a sequência de busca seguida nesta etapa;

- Busca manual – neste trabalho optou-se por fazer uma busca manual dos casos não encontrados na automática utilizando os nomes completos. O objetivo seria de identificar os problemas encontrados e criar rotinas para enriquecer a busca automática.

1. Tipo, Título, Nome, Número
2. Tipo, Título, Nome
3. Tipo, Nome, Número
4. Tipo, Nome
5. Título, Nome, Número
6. Título, Número
7. Nome, Número
8. Nome

Figura 5.1. Sequência de busca automática

Como resultado do processo de geocodificação (Figura 5.2), foram localizados automaticamente 1114 (63%) endereços nos setores correspondentes. Dos 648 endereços restantes, 392 (22%) foram localizados na primeira etapa da procura manual em que houve a comparação em termos de nomes com grafia não coincidente e diferenciação, pelas informações de referência de local, tais como bairro ou pontos de referência, foram localizados os setores censitários de mais 150 (9%) endereços. Os 106 (6%) endereços restantes apresentavam problemas relacionados à ausência de dados ou endereços semelhantes em diferentes setores censitários, o que exigiu uma revisão dos mesmos. Foram realizadas visitas domiciliares pela equipe do CPqGM/ FIOCRUZ, com o objetivo de melhorar a qualidade do dado. Após as visitas domiciliares e a revisão dos endereços obteve-se 100% dos endereços geocodificados.

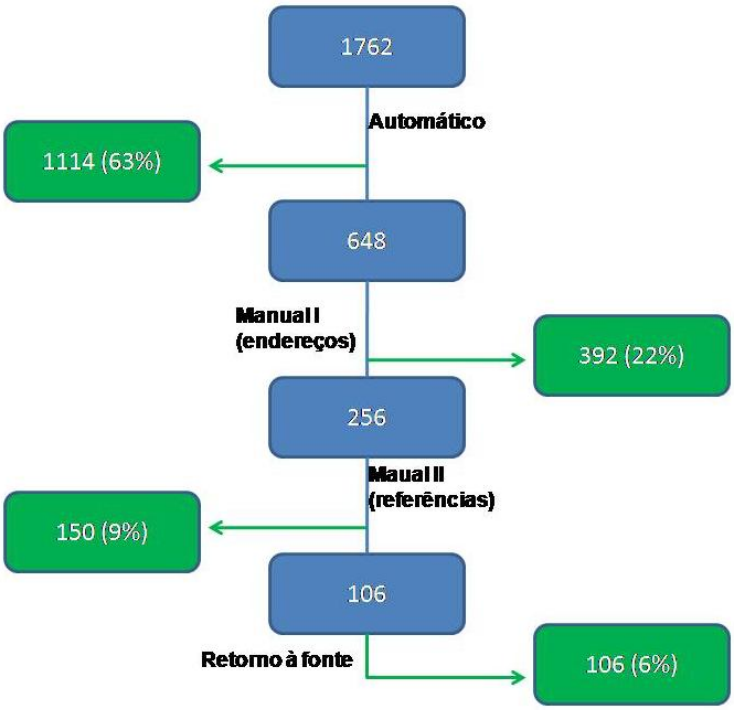


Figura 5.2. Resultado do processo de geocodificação

A Figura 5.3 apresenta os mapas que constam da dissertação, utilizando as coordenadas dos centróides dos setores censitários dos endereços geocodificados. São mapas da razão de *Kernel* dos casos de leptospirose em Salvador, sobrepostos às áreas de favelas, no período de 1996 a 2006, sendo o primeiro no período de seca e o segundo no período de chuva.

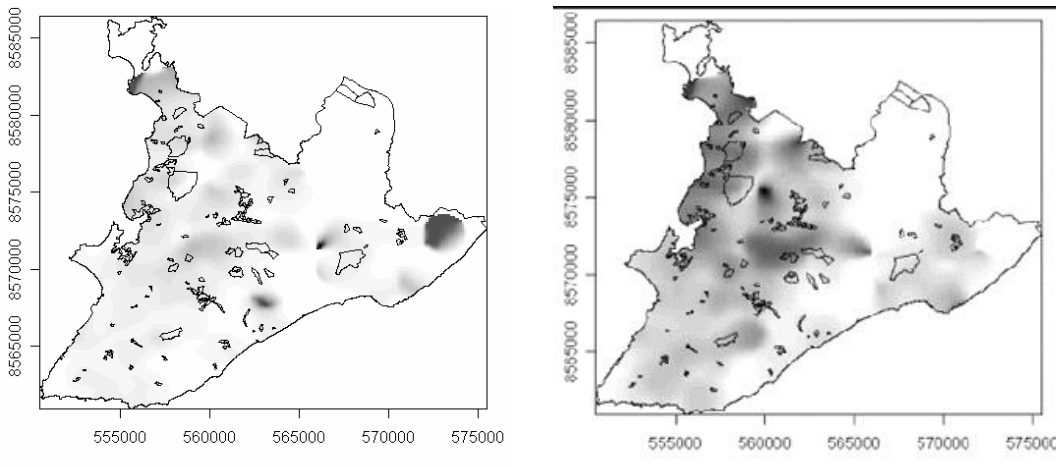


Figura 5.3. Razão de Kernel dos casos de leptospirose em Salvador entre 1996 e 2006, nos períodos de seca e de chuvas (Nunes, 2007)

CONCLUSÕES

Neste trabalho, deve-se ressaltar a qualidade da informação de entrada, fato não comum para o tipo de evento estudado, por ser localizado, em sua maioria, em áreas com endereços de difícil definição, como as favelas ou periferias. Mesmo nessas condições, foram localizados os setores censitários de 63% destes endereços de forma automática, com a pesquisa de nome inteiro do logradouro. Após este primeiro processamento, foram feitas duas buscas manuais, sendo que a primeira localizou os logradouros com erros de grafia, tanto por questões fonéticas, com trocas de letras com o mesmo fonema (como s e z ou sh e ch), como por erro de preenchimento, com inclusão, exclusão ou troca de letras. Ao final, para os 106 endereços restantes e sem condições de localização, por falta de preenchimento ou preenchidos com textos não identificados, foi feito um retorno às fontes de informações, com o preenchimento correto do endereço ou localizando os setores censitários diretamente nos mapas.

Considerando-se a busca automática, o número de casos que podem ser avaliados como verdadeiros positivos (VP) é de 1114. Neste sentido, os falsos negativos (FN) são os 542 localizados manualmente e os verdadeiros negativos, os 106 só localizados voltando à pesquisa de campo. Assim pode-se calcular o valor de sensibilidade (VP / (VP + FN)), como sendo: 1114 / (1114 + 562) = 0,67. Entretanto para o cálculo de especificidade necessita-se do número de falsos positivos. Este valor só pode ser calculado utilizando o padrão-ouro, ou seja, o verdadeiro local do endereço pesquisado, como, por exemplo, a medição das coordenadas com GPS, o que não há neste caso.

Como contribuição para a definição da metodologia, com o objetivo de obter um melhor resultado nas rotinas de busca automática, foram identificadas as necessidades de utilização de busca fonética e do cálculo de aproximação de textos para indicar possíveis concordâncias de nomes e criação de rotinas para a utilização de pontos de referência no processo de refinamento de localização. Algumas perdas podem ser inevitáveis quando não há possibilidade de retorno à fonte de informação.

5.2. LEPTOSPIROSE NO RIO DE JANEIRO

Outro trabalho com pesquisa em áreas predominantemente de ocupação desordenada para apoio a um estudo de análise espacial é o de ocorrência de casos de leptospirose no Rio de Janeiro no período de 1997 a 2002. Este trabalho serviu de apoio à tese de doutorado de Wagner Tassinari (2009), a partir de seu primeiro artigo (Tassinari et al., 2007). Foram coletados 514 endereços residenciais correspondentes às notificações no Sistema Nacional de Agravos de Notificação (SINAN) de casos de leptospirose no Rio de Janeiro entre os anos de 1997 e 2002. Em uma primeira fase foram localizados automaticamente 165 endereços utilizando o sistema de localização do Laboratório de Geoprocessamento da CICT/Fiocruz (Labgeo, 2001), outros 164 endereços foram localizados manualmente, através de guias rodoviários.

Para os 185 endereços restantes, com maior complexidade, por não conter associação direta dos nomes de logradouros, com abreviações ou nomes de logradouros homônimos, foi desenvolvido um procedimento manual utilizando o cadastro de Folhas de Coleta do Censo 2000 do IBGE (Figura 5.5). Este cadastro contem todos os endereços visitados pelos recenseadores, organizados por Setor Censitário. Os campos obtidos no cadastro de Folha de Coleta são:

- Setor Censitário – código do Setor Censitário do Censo 2000, com 15 dígitos. (campos 1.01, 1.03, 1.04 e 1.05).
- CEP – Código de Endereçamento Postal, preenchido pelo recenseador (campo 1.08).
- Localidade – nome local. Pode ser preenchido com bairro, distrito ou nome mais conhecido, pesquisado pelo recenseador (campo 1.10). Este campo é importante na localização das áreas com nome local não oficial que, em alguns casos, expressa como é conhecido pela população e muitas vezes utilizado no endereço informado por esta população.
- Logradouro – nome completo do logradouro, incluindo tipo e título (campo 1.11).
- Número do logradouro – número de porta do logradouro (campo 2.01).

- Complemento – complemento do domicílio, composto por abreviaturas pré-estabelecidas e o complemento propriamente dito. Ex. AP301, representando apartamento 301 (campo 2.02).
- Espécie – define o tipo de domicílio ou determina unidade não residencial (campo 2.03).

Este cadastro se encontra em fase de consolidação com vistas à criação de um Cadastro Nacional de Endereços para Fins Estatísticos – projeto CNEFE (IBGE, 2005). Destes 185 endereços pesquisados, foram localizados os setores censitários de 133 endereços ou referências e indicado um setor entre dois ou três possíveis com mesma característica para 26 endereços. Os 26 endereços restantes foram identificados como não localizados no município do Rio de Janeiro, mas em municípios vizinhos. Neste trabalho os nomes locais foram de grande utilidade, servindo como referência na definição dos setores censitários, principalmente em áreas de favela. A Figura 5.6 apresenta a localização dos 488 casos de leptospirose identificados.

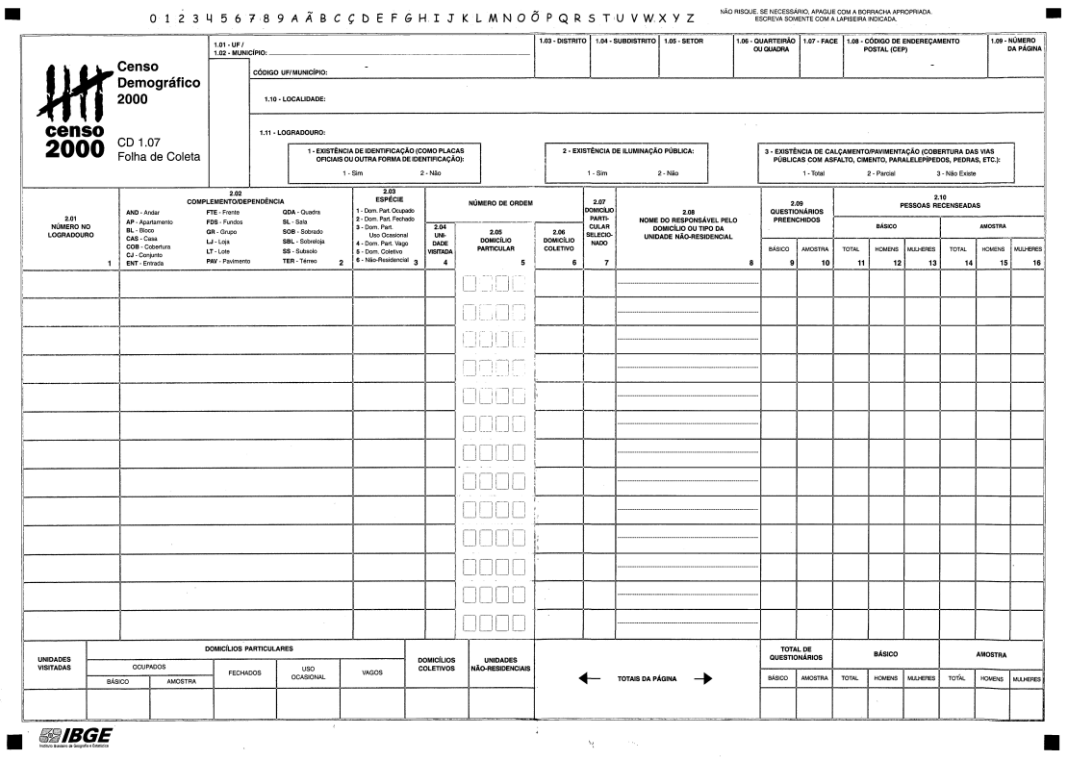


Figura 5.4. Folha de Coleta do Censo 2000

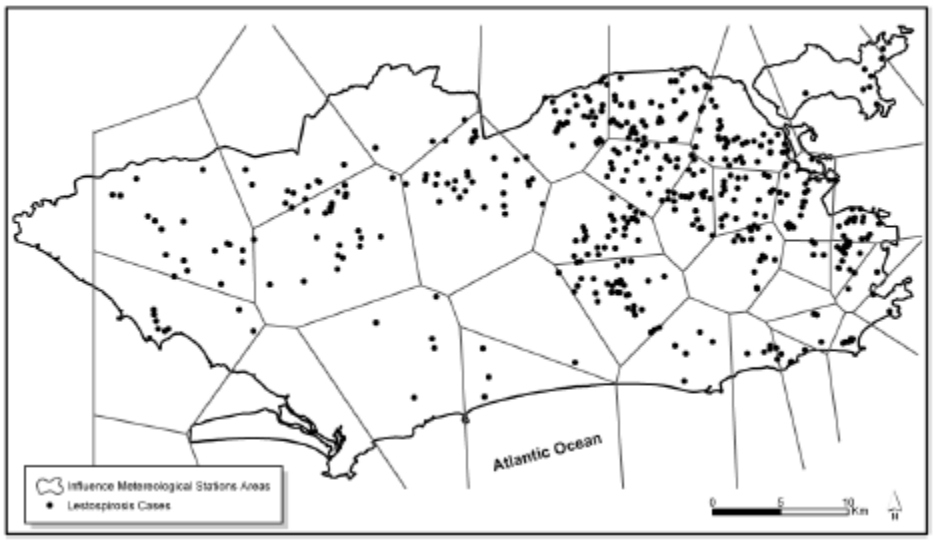


Figura 5.5. Distribuição dos casos de leptospirose no Rio de Janeiro e os polígonos de Voronoi com cada uma das 32 estações meteorológicas

Comparando-se os resultados dos 488 casos identificados no final do processo com os dos 165 obtidos na geocodificação automática (Tabela 5.1), pode-se ter a dimensão da importância deste trabalho adicional. Enquanto no estudo com os 165 casos obteve-se um resultado somente marginalmente significante (p-valores < 0.10), com os 488 casos houve uma definição significativa da diferença entre os dois primeiros anos e os outros.

Tabela 5.1. Comparação das características por cluster espaço-tempo

ANO	1997	1998	1999	2000	2001	2002
Geocodificação automática (165 casos)						
Casos no ano	33	34	27	26	23	20
Casos por cluster	6	5	2	2	2	2
p-valor	0.06	0.08	0.12	0.49	0.58	0.25
Geocodificação ampliada (488 casos)						
Casos no ano	114	111	64	65	71	63
Casos por cluster	13	19	2	2	2	5
p-valor	< 0.001	< 0.001	0.291	0.161	0.590	0.973

CONCLUSÕES

Neste trabalho ficou evidenciada a importância de processos para a localização do maior número possível de eventos nos estudos epidemiológicos, principalmente quando a doença estudada tem característica de ocorrência em áreas de ocupação desordenada, onde o endereçamento não é bem definido.

Como contribuição deste trabalho na definição do processo proposto, fica fortalecida a necessidades de pesquisa fonética, rotinas de aproximação de texto, e pesquisa de abreviaturas, além de utilização de locais de referência, tanto de distribuição política (bairro, RA, etc.) como de nomes não oficiais (favelas, pontos de referência, nomes locais).

5.3. COORTE DE NASCIMENTO DE PELOTAS

Este trabalho de geocodificação serviu de apoio para o desenvolvimento da tese "Peso ao nascer e determinantes ecológicos nos padrões nutricionais de crianças" (Martins, 2007).

A fonte de dados para estes estudos é a Coorte de Nascimento de Pelotas de 1993 (Victora et al., 2006). As coortes de nascimento de Pelotas estudam os nascidos vivos no município, dos anos de 1982, 1993 e 2004, acompanhando as crianças e mães. Deste modo, fazem um mapeamento da saúde no município, contribuindo com as políticas públicas.

A coorte de 1993 é formada por 5249 crianças nascidas neste ano. Como base de referência para as rotinas de geocodificação foram utilizados os arquivos de Folha de Coleta do Censo 2000 (IBGE, 2002). Como o objetivo deste processo era obter os códigos de setores censitários do Censo 1991, para a utilização das informações sócio-econômicas desta pesquisa, e não bases de referência desta data, estes setores foram obtidos a partir dos códigos dos setores 2000 e das tabelas de comparabilidade inter-censos, ou seja, entre 2000 e 1996 e entre 1996 e 1991 (IBGE, 2002). Estas tabelas retratam os relacionamentos entre os setores censitários de dois censos consecutivos, composto pelos códigos dos setores e o código de formação, com o tipo de relação: manutenção, agregação ou desmembramento (Figura 5.6).

Na criação da comparabilidade de setores, alguns fatores são considerados, além do aumento da população na área em questão, tais como mudança de divisão político-administrativa, alteração de perímetro urbano ou surgimento de áreas de favelas.

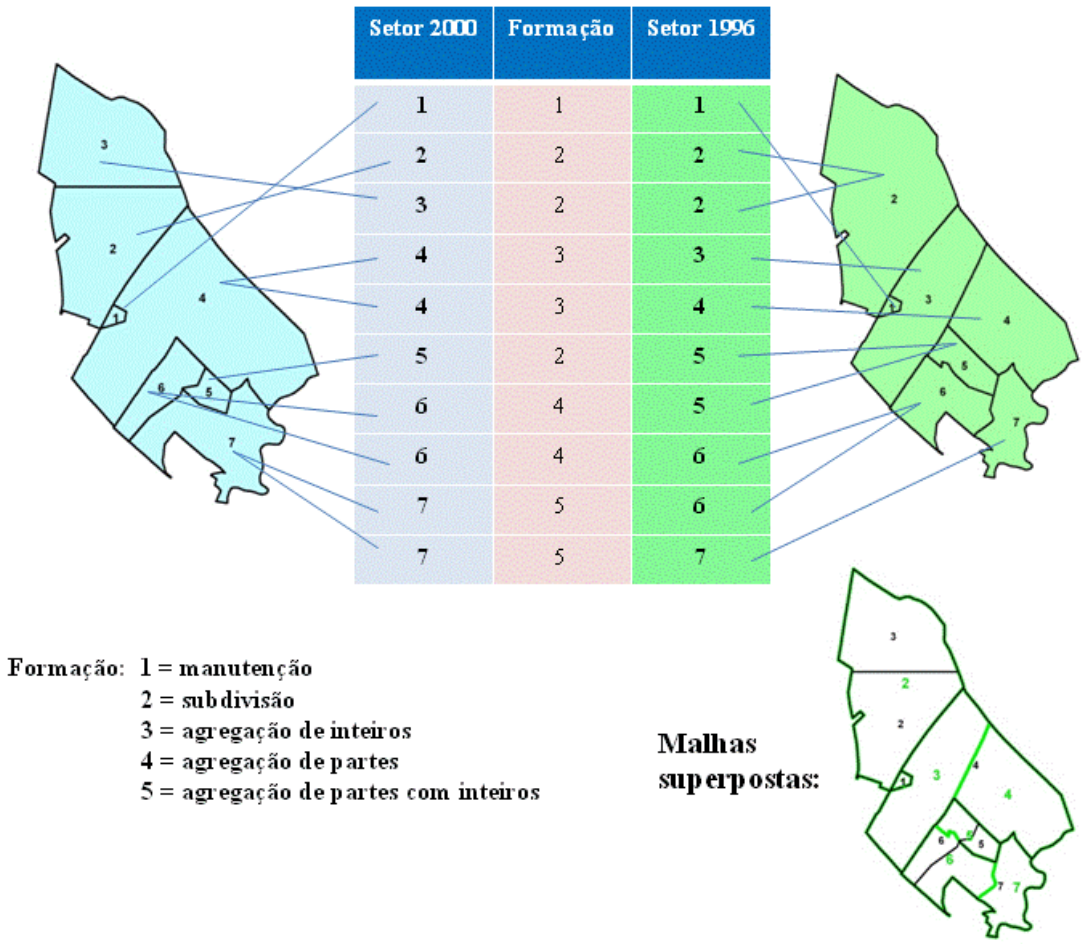


Figura 5.6. Comparabilidade de Setores

No processo de geocodificação propriamente dito, outros fatos relacionados ao preenchimento foram percebidos, tais como:

- Endereços em branco – estes registros não têm recuperação, a menos de um novo contato à fonte;
- Endereços ilegíveis – registros preenchido com um texto que não define um endereço. Exemplo encontrado na base de dados: “A mãe não sabe o nome da rua, número 22”. Do mesmo modo do grupo anterior, não há recuperação;
- Endereços incompletos e sem campo de referência – nestes pode haver mais de um setor censitário com possibilidade de ser o local procurado;

- Um mesmo logradouro com vários modos de apresentação da grafia. Na Figura 5.7 estão as 30 formas em que foi encontrada a informação do logradouro "Avenida Juscelino Kubitschek de Oliveira" nos endereços. Nestes foi retirada a informação de número de porta para não haver identificação individual.

AVENIDA JUSCELINO KUBITSCHECK
AV.JUSCELINO KUBISCHEK
AV JUSCELINO KUBITSCHEK
AV.JUSCELINO KUBITSCHECK
JOSSELINO CUBICKE
JUCELINO KUBICHEK OLIVEIRA
AV.JUSCELINO KUBITSCHECK DE OLIVEIRA
AV.JUSCELINO KUBITSCHECK
AV.JUCELINO KUBISHECK DE OLIVEIRA
Av. J.K. OLIVEIRA
AV.JUSCELINO KUBITSCHEK
AV.JUSCELINO K.DE OLIVEIRA
AV.JUSCELINO KUBITSCHEK
AV.JUSCELINO KUBITSCHECK DE OLIVEIRA
JUSCELINO KUBITSCHEK DE OLIVEIRA
JUSCELINO KUBISCHECK DE OLIVEIRA
AV.JUSCELINO K.DE OLIVEIRA
AV. JK DE OLIVEIRA
AV.JK
AV.JUCELINO KIBICKEK OLIVEIRA
AV.JUCELINO K.DE OLIVEIRA
AV.JUSCELINO K.DE OLIVEIRA
JK DE OLIVEIRA (ANTIGA AVENIDA 41)
AV.JUSCELINO KUBISCHECK DE OLIVEIRA
JK
JUSCELINO KUBICHEQUE
AV.JK DE OLIVEIRA
AV. JUSCELINO KUBICHTKE DE OLIVEIRA
RUA JUSCELINO KUBITSCHEK DE OLIVEIRA
JUCELINO KUBICHEK

Figura 5.7. Formas de apresentação de um logradouro

A metodologia utilizada apresenta os seguintes passos:

- Normalização do endereço – neste procedimento, foram retirados acentos e cedilhas e os caracteres alfabéticos transformados apenas em letras maiúsculas;

- Separação dos campos – divisão do campo de endereços em tipo de logradouro (rua, avenida, estrada, ...), título (presidente, padre, coronel, ...), nome, número e complemento;
- Procura do endereço – segue a sequência descrita em 5.1;
- Pesquisa fonética de nome do logradouro (utilizando Soundex) – para os não relacionados a nenhum setor censitário;
- Pesquisa de referência (bairro) – para os relacionados a mais de um setor censitário;
- Retorno à fonte – para os não relacionados ou relacionados a mais de um setor, para acerto de endereço ou obtenção de referência;
- Nova sequência total.

Os resultados encontrados estão representados na Figura 5.8, sendo que os endereços incompletos (224) são os endereços em branco ou sem nome do logradouro; nenhum setor localizado (162) diz respeito a não haver setor censitário com o registro do logradouro preenchido; setor não definido (120) se refere aos endereços em que há mais de um setor censitário em locais não vizinhos com logradouro de mesmo nome e sem informação de referência que possa dirimir a dúvida. Deste modo, dos 5105 endereços de entrada, 4291 (84,1%) foram geocodificados. A Figura 5.9 apresenta um resultado obtido na pesquisa utilizando os dados geocodificados. Trata-se de um mapa obtido pelo método bayesiano empírico.

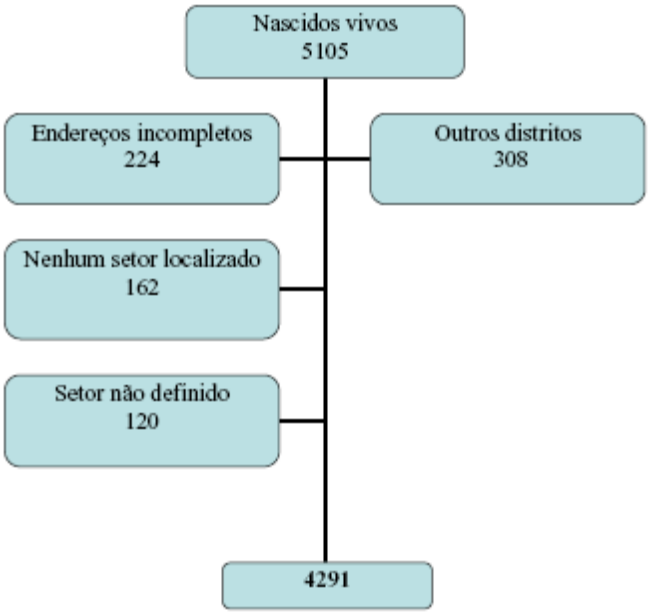


Figura 5.8. Resultado da geocodificação (adaptado de Martins, 2007)

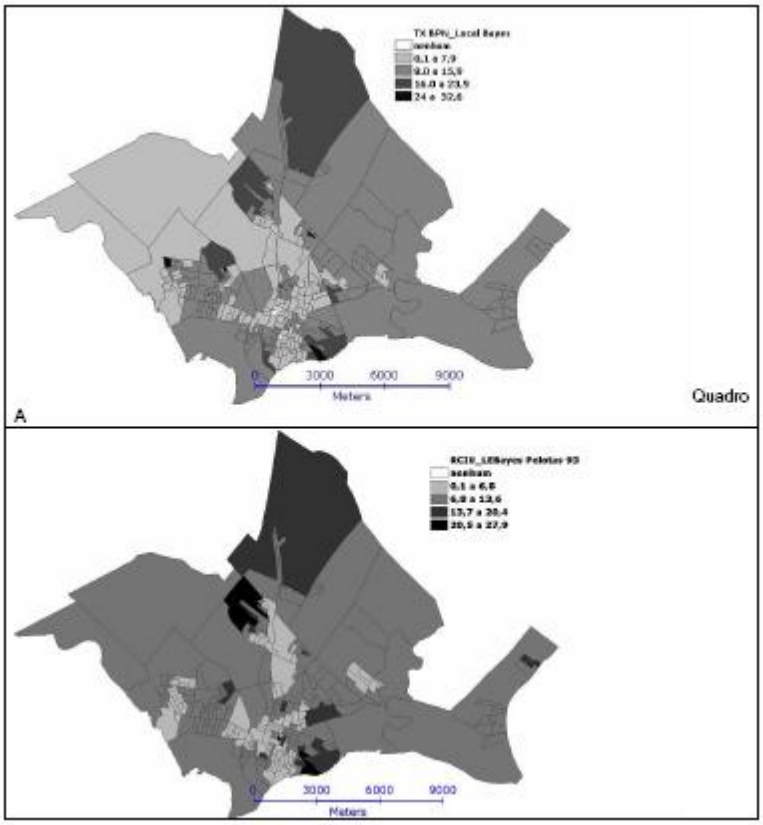


Figura 5.9. Bayesiano empírico (Bender & Carvalho, 2006)

CONCLUSÕES

Este trabalho tem algumas características peculiares. Algumas perdas acontecem neste processo, por falta de contato com alguns componentes, por mudança de cidade de moradia ou, simplesmente, por desistência. Pode-se notar que entre os participantes é criada uma identidade, como pode ser demonstrado na comunidade criada no site de relacionamento Orkut (www.orkut.com.br), cuja descrição é: "*Essa é uma comunidade para todos que nasceram em Pelotas/RS no ano de 1982 e fazem parte do projeto de acompanhamento da UFPel. Tem como objetivo trocar experiências sobre este projeto*".

Outra característica importante foi a necessidade de se reportar aos setores censitários de 1991, a partir de um cadastro de endereços por setor de 2000 com, inclusive, alteração na divisão territorial urbana do município, havendo ampliação do perímetro urbano e mudança da estrutura de alguns bairros.

Como contribuições para o processo proposto no capítulo 4, podem-se apontar as necessidades de: utilizar outras formas de aproximação de texto além da busca fonética; utilizar áreas vizinhas quando a área de referência for bairro ou outra divisão territorial; utilizar abreviaturas comuns no campo de nome de logradouro na base de dados.

5.4. PROJETO ELSA - CEP

O Estudo Longitudinal de Saúde do Adulto - ELSA Brasil - é uma investigação multicêntrica, realizada em 6 locais (Rio de Janeiro, São Paulo, Bahia, Minas Gerais, Espírito Santo e Rio Grande do Sul). Trata-se de coorte composta por 15 mil funcionários, com idade entre 35 e 74 anos, de seis instituições públicas de ensino superior e pesquisa. A pesquisa tem o propósito de investigar a incidência e os fatores de risco para doenças crônicas, em particular, as cardiovasculares e o diabetes (www.elsa.org.br).

Para uma melhor performance no relacionamento de um banco de dados é desejável a existência de códigos com característica de chave primária, ou seja, não nula e única. No endereço, a informação que tem estas características é o Código de Endereçamento Postal – CEP – que, para os 200 maiores municípios brasileiros, possui um código para cada logradouro ou trecho destes. Com o objetivo de testar a utilização deste campo como informação básica para a pesquisa de localização de endereço, foi feito um teste de relacionamento dos endereços dos 500 primeiros participantes do Estudo Longitudinal de Saúde do Adulto – ELSA – no Rio de Janeiro a partir do CEP. Como resultado deste teste, foram encontrados os seguintes dados:

- CEP em branco – 2 (0,4%)

- CEP inválido – 64 (12,8%)

- CEP válido com logradouro não coincidente – 32 (6,4%)

- CEP correto – 402 (80,4%)

CONCLUSÕES

Analisando os casos de CEP inválido ou não coincidente, não foi encontrado um padrão de bairro ou área da cidade, sendo que os 98 casos localizaram-se em 49 bairros diferentes (Figura 5.10), com a maior incidência no bairro das Laranjeiras, com os casos

localizando-se na Rua das Laranjeiras que possui vários CEPs, não coincidindo, neste caso, os três últimos dígitos.

Como conclusão deste trabalho, fica claro que, caso exista um CEP válido nos dados de entrada, deve-se testar o nome do logradouro para a confirmação deste.

BAIRRO	QUANT.	BAIRRO	QUANT.
ANDARAÍ	3	LARANJEIRAS	7
AREIA BRANCA	1	MARACANA	1
BANCO DE AREIA	1	MARAMBAIA	1
BANGU	1	MEIER	1
BARRA DA TIJUCA	2	NOVA CIDADE	1
BRAZ DE PINA	1	OLARIA	1
CACHAMBI	1	PANTANAL	1
CENTRO	3	PAVUNA	1
CENTRO / VILA EMIL	1	PECHINCHA	1
CIDADE DE DEUS	1	PIABETÁ	1
CONRADO	1	PIEDADE	1
COPACABANA	1	PILARES	1
CORDEIRINHO	1	PIRATININGA	1
COSMOS	1	PORTUGUESA	1
EDSON PASSOS	1	PRACA CRUZEIRO	1
ICARAI	1	RAMOS	2
ILHA	1	ROCHA SOBRINHO	1
IPANEMA	1	SANTA TERES	1
ITAIPUAÇU	2	SÃO DOMINGO	1
ITAPEBA	1	SARACURUNA	1
JACAREPAGUA	2	SENADOR CAMARA	1
JACON	1	SURUI	1
JARDIM 25 DE AGOSTO	1	TANQUE	1
JARDIM BOTANICO	2	TIJUCA	1
JARDIM GUANABARA	1		

Figura 5.10. Quantidade CEPs inválidos por bairro

5.5. ADESÃO AO HAART

A terapia antirretroviral altamente ativa (HAART – *higly active antiretroviral therapy*) é constituída pela combinação de, pelo menos, três drogas que atuam sobre diferentes partes do HIV e impedem que o vírus entre nas células sanguíneas.

Este trabalho é parte da tese de doutorado em Saúde Pública de Dayse Pereira Campos e foi utilizado para a construção do artigo "Avaliação da associação da adesão ao HAART na evolução clínica dos pacientes HIV positivos", com o objetivo de avaliar a evolução clínica dos pacientes em uso de HAART, em função da adesão em pacientes acompanhados em uma instituição de excelência (Instituto de Pesquisa Clínica Evandro Chagas – IPEC) na prestação de assistência ao portador HIV/AIDS, controlando por fatores sócio-demográficos, assistenciais, clínicos e relacionados ao tratamento. Uma das variáveis independentes utilizadas nesta avaliação foi a distância da residência do paciente ao IPEC. Para obter esta variável, foi utilizado o endereço completo e mais recente do paciente (disponível no sistema de informações do IPEC – SISPEC) e o endereço do IPEC. Estas duas informações foram incluídas no Google Earth (http://earth.google.com.br), utilizando uma função de busca de trajeto (de/para). Nos casos de endereços não localizados diretamente pelo nome do logradouro do endereço do paciente, foi utilizada a distância ao município ou bairro da residência deste. Foram calculadas as distâncias de 1.738 paciente, e em apenas 97 (5,6%) não foi possível localizar o endereço. Estas perdas foram causadas, principalmente, por trocas de nomes de municípios (emancipados) ou endereços em áreas de ocupação desordenada (favelas ou periferias). O CEP, presente na maioria dos endereços, facilitou a busca. O tempo gasto nesta atividade foi de 72 horas de trabalho, espalhadas em 3 semanas.

Para o artigo, foram analisados 711 pacientes. Destes, a distância mínima foi de 1 km e a máxima de 323 km. A média foi de 28,97 km com desvio padrão de 25 km e mediana de 25 km. A categorização proposta no trabalho foi de distância até 5 km e de mais de 5 km, considerando as áreas de baixa renda localizadas no entorno do IPEC/Fiocruz. A Figura 5.11 mostra um gráfico com a distribuição dos pacientes por

distância ao IPEC e outro com a probabilidade de falhas, em função do número de dias do tratamento, separando os pacientes nas duas categorias de distância do IPEC.

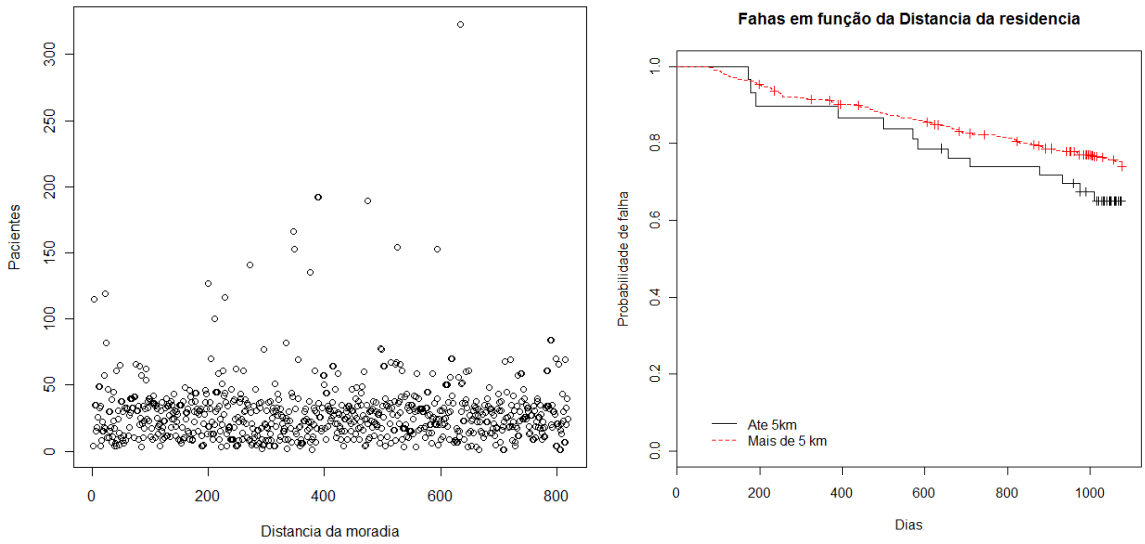


Figura 5.11. Distribuição dos pacientes por distância da moradia e probabilidade de falhas em função da distância da residência (Campos 2009)

CONCLUSÕES

Este trabalho mostra que, a partir das possibilidades de acesso à informação espacial e com a criatividade, que é fator constante no processo científico, novas soluções são criadas. Esta solução envolveu apenas o acesso a uma ferramenta disponível na Internet e o usuário final, sem interferências de processo específico ou de profissional especialista.

Como contribuição ao processo proposto nesta tese, sugere que um bom caminho para apresentação dos resultados da geocodificação é a criação de dados gráficos em formatos utilizáveis em aplicativos de uso público, como o KML do Google Earth.

6. COMENTÁRIOS FINAIS

"A coisa mais bela que o homem pode experimentar é o mistério. É essa emoção fundamental que está na raiz de toda ciência e toda arte" (Albert Einstein)

O processo de geocodificação abrange vários aspectos metodológicos, concentrando conceitos e práticas de uma gama extensa de especialidades. Começando pelos objetivos do processo, neste caso, a Epidemiologia, sendo fundamental para as análises espaciais. Os Sistemas de Informação Geográfica estão na base destas áreas de interesse, passando também por aspectos culturais que se refletem na construção e determinação dos endereços residenciais ou postais. A área política, com seus projetos e prioridades, tem um efeito marcante nestas abordagens, pois as atividades de construção dos cadastros e bases territoriais gráficas digitais demandam grandes somas de recursos, além de dependerem das definições das divisões político-administrativas deste território. Juntam-se a estes aspectos, os recursos computacionais e a disponibilidade, hoje cada vez mais crescente, de dados digitais de representação geográfica.

A definição e o modo de informação do endereço sofre influência da estrutura administrativa, com a divisão político-administrativa (bairros, distritos) e a identificação e sinalização dos logradouros e locais públicos, além da produção e fornecimento de mapas de localização; e da cultura de cada local, com o hábito de fornecer informações claras e consistentes sobre os locais de moradia ou trabalho. Somam-se a estes dois aspectos, as leis locais sobre individualidade do cidadão. Enquanto na Finlândia existe um ato na constituição que diz: *"Every resident of Finland has a unique personal identification number which can be used for linking records between various national databases on an individual basis. This also allows every individual to be located by means of the map coordinates of his or her place of residence and postal address"*[10] (Rytkonen et al., 2001), em outros, como o Brasil, é respeitado o sigilo individual. Neste sentido, apresenta-se a questão ética, que se mostra de maneiras distintas em cada país, determinada por leis e culturas.

As dificuldades encontradas para geocodificação no Brasil, além do mal preenchimento da informação, devem-se às diferentes estruturas existentes de endereço. Enquanto a maioria das áreas urbanas possui estrutura de logradouro e número de porta,

[10] "Todo residente da Finlândia tem um número de identificação pessoal único que pode ser usado para combinar entre várias bases de dados nacionais e uma base de dados individual. Isso também permite que todos os indivíduos possam ser localizados por meio de coordenadas geográficas de seu endereço de residência".

outras, como Brasília, são organizadas por quadra e lote. Esta estrutura também é utilizada em novos loteamentos, enquanto não há definição dos logradouros. Na maioria das áreas rurais e nas áreas de ocupação desordenada, não há definição de logradouros ou quadras, o que dificulta ou até inviabiliza a definição de localização do domicílio, obtendo-se, como alternativa, a identificação por localidade, oficial ou não. Neste caso, a unidade geométrica de representação espacial será ponto ou polígono, dependendo da escala cartográfica utilizada.

A etapa de entrada de dados tem grande importância no êxito do processo. Nela é gasto grande parte do esforço de desenvolvimento, quando não há padrão de apresentação dos dados, por conter campos de textos e necessidade de rotinas de comparação destes. O CEP seria o número de identificação do local de geocodificação para uma melhor performance do processo, mas ainda existem muitos erros nesta informação, por desconhecimento dos informantes, necessitando que se combine este campo com o nome do logradouro, voltando aos problemas decorrentes da comparação de campos de texto. Uma alternativa para minimizar este problema é a criação de uma entrada de dados assistida por computador na unidade de saúde, com tabelas das estruturas de domicílios (logradouros, quadras ou nomes locais) válidos para as localidades trabalhadas.

Para a apresentação do resultado da geocodificação, devem-se considerar os objetivos do projeto que a demandou. Uma maior necessidade de precisão no resultado, diminui a eficiência do processo e aumenta o risco de não se conseguir este resultado. Além da escolha do nível de precisão, o usuário final deve especificar o tipo de informação e o formato do dado para utilização. Entretanto, um fator importante para se avaliar um sistema de geocodificação é a qualidade do processo, que pode ser medida quanto ao processo em si (eficiência, eficácia ou precisão) ou quanto aos resultados (sensibilidade ou especificidade). Estas medidas são calculadas utilizando-se um padrão-ouro, que pode ser obtido por meio de coordenadas tiradas através de GPS.

Os recursos necessários para o desenvolvimento de um projeto de geocodificação (bases de dados e tecnologia) estão em ampla expansão. As bases de dados estão sendo produzidas em grande escala, tanto por órgãos públicos como na área

comercial. Na área governamental, grandes projetos nacionais estão produzindo bases de dados de endereços, associando cadastros a arquivos gráficos, com projetos de uso geral ou de uso específico para as áreas de interesse. A citar:

- Estados Unidos: Master Address File (MAF) – arquivo arquivo nacional de endereços atualizado a cada ano.
- Australia: Geocoded National Address File (G-NAF) – base de dados confiável de referência para dados de endereço.
- Reino Unido: ADDRESS-POINT – uma ferramenta definitiva para a identificação e localização precisa dos endereços.
- Brasil: Cadastro Nacional de Endereços para fins estatísticos (CNEFE) – base de dados de endereços para censos e pesquisas.

Estes cadastros são georreferenciados e criam condições de geocodificação. Na área comercial, o surgimento de aplicativos on-line, do tipo Google Earth (earth.google.com.br), facilitam o acesso a rotinas de localização de endereços a qualquer usuário com acesso à Internet. Na área tecnológica, os recursos computacionais estão cada vez mais potentes e financeiramente mais acessíveis, tornando mais viável o manuseio de arquivos gráficos de mapas e de imagens. Somando-se a estes fatores, o aparecimento de softwares livres e abertos para o gerenciamento de bancos de dados e de arquivos gráficos, torna mais viável o desenvolvimento e implantação de aplicativos.

Neste cenário, projeta-se a implementação da proposta desta tese por meio de uma aplicação de uso público, através de software livre, com possibilidade de agregar bases de dados disponíveis.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Alexander GL, Kinman EL, Miller LC, Patrick TB. Marginalization and health geomatics. Journal of Biomedical Informatics 2003. 36: 400-407.

Almeida-Filho N. Modelos de determinação social das doenças crônicas não-transmissíveis. Ciência & Saúde Coletiva; 2004. 9(4): 865-894.

Aronoff S. Geographic information systems: a management perspective. Ottawa: WDL Publications. 1990.

Assunção RM, Barreto S, Guerra H, Sakurai E. Mapas de taxas epidemiológicas: Métodos estatísticos. Cadernos de Saúde Pública. 1998. 14: 713-723.

Assunção RM. Estatística Espacial com Aplicações em Epidemiologia, Economia e Sociologia. São Carlos: Associação Brasileira de Estatística. 2001.

Austrian GD. Herman Hollerith: Forgotten Giant of Information Processing. New York. Columbia University Press. 1982.

Bailey TC, Gatrell AC. Interactive Spatial Analysis. Longman. 1995.

Bailey T. Spatial statistical methods in health. Cad. Saúde Pública. 2001. 17(5): 1083-1098.

Barcellos C, Ramalho W. Situação atual do geoprocessamento e da análise de dados espaciais em saúde no Brasil. Revista IP – Informática Pública 2002. 4: 221-30.

Barcellos C, Santos SM. Colocando dados no mapa: A escolha da unidade de agregação e integração de bases de dados em saúde e ambiente através do geoprocessamento. Informe Epidemiológico do SUS; 1997. 6: 21-29.

Barcellos C, Sabroza P. Socio-environmental determinants of the leptospirosis outbreak of 1996 in western Rio de Janeiro: a geographical approach. Int J Environ Health Res. 2000. 10(4): 301-313.

Barcellos C, Lammerhirt CB, Almeida MAB. Distribuição espacial da leptospirose no Rio Grande do Sul, Brasil: recuperando a ecologia dos estudos ecológicos. Cadernos de Saúde Pública. 2003. 19: 1283–1292.

Barcellos C, Ramalho WM, Gracie R, Magalhaes MAF, Fontes MP, Skaba DA. Georreferenciamento de dados da saúde na escala sub-municipal: algumas experiências no Brasil. Epidemiologia e Serviços de Saúde; 2008. 17(1): 59-70.

Barret FA. "SCURVY" Linds Medical Geography. Social Science and Medicine. 1991. 33: 347-353.

Boscoe FP, Ward MH, Reunolds P. Current practices in spatial analysis of cancer data: data characteristics and data sources for geographic studies of cancer. International Journal of Health Geographics 2004. 3: 28.

Boulos MNK. Location-based health information services: a new paradigm in personalized information delivery. International Journal of Health Geographics. 2003. 2: 2.

Boulos MNK. Web GIS in practice: creating a simple interactive map of England's stractegic health authorities using Google Maps API, Google Earth KML, and MSN Virtual Map Control. International Journal of Health Geographics 2005. 4: 22.

Burdette HL, Whitaker RC. Neighborhood playgrounds, fast food restaurants, and crime: relationships to overweight in low-income preschool children. Preventive Medicine 2004. 38: 57-63.

Burrough PA. Principles of geographic information systems for land ressource assessment. Oxford University press. Oxford. 1992.

Câmara G. Anatomia de sistemas de informação geográfica, visão atual e perspectivas de evolução. In: Sistemas de informação geográfica e suas aplicações na agricultura. Brasilia – DF. 1993. 37-59.

Camara G, Carvalho MS. Análise espacial de eventos. Análise espacial de dados geográficos. In: Câmara G, Monteiro AM, Fucks SD, Carvalho MS. Spatial Analysis and GIS: A Primer. EMBRAPA. Brasília. 2004.

Camargo Jr KR, Coeli CM. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic Record linkage. Cad. Saúde Pública. 2000. 16(2): 439-447.

Carvalho MS, Cruz OG, Nobre FF. Spatial partition using multivariate cluster analysis and contiguity algorithm: application to Rio de Janeiro, Brazil. Statistics in Medicine. 1996. 15: 1885-1894.

Carvalho MS, Cruz OG. Análise espacial por microáreas: métodos e experiências. In: Veras RP, Barreto ML, Almeida Filho N, organizadores. Epidemiologia: contextos e pluralidade. Editora Fiocruz. Rio de Janeiro: 1998. 79-89.

Carvalho MS, Pina MF, Santos SM. Conceitos Básicos de Sistemas de Informações Geográficas e Cartografia Aplicados à Saúde. Organização Panamericana de Saúde / Ministério da Saúde. Brasília. 2000.

Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. International Journal of the Health Geographics. 2003. 2: 20.

Chen W, Petitti DB, Enger S. Limitations and potential uses of census-based data on ethnicity in a diverse community. Ann Epidemiol. 2004. 14: 339-345.

Christen P, Churches, T, Hegland M. A Parallel Open Source Data Linkage System. Proceedings of the 8th PAKDD'04 (Pacific-Asia Conference on Knowledge Discovery and Data Mining), Sydney. Springer LNAI-3056. 2004. 638-647.

Christen P, Churches, T. A probabilistic reduplication, record linkage and geocoding system. In Proceedings of the Australian Research Council Health Data Mining Workshop (HDM05), Canberra, AU. 2005.

Correios - Empresa Brasileira de Correios e Telégrafos. Guia Postal Brasileiro. Correios. 1992

Churches T, Christen P, Lim K, Zhu J. Preparation of name and address data for record linkage using hidden Markov models. BMC Medical Informatics and Decision Making. 2002. 2(1): 9.

Cruz OG. Modelagem Espaço Temporal dos Homicídios da Região Sudeste do Brasil, 1979-1998. Tese de Doutorado – Universidade Federal do Rio de Janeiro, Rio de Janeiro. 2004.

Curtis S, Jones IR. Is There a Place for Geography in the Analysis of Health Inequality? Sociology of Health & Illness. 1998. 20(5): 645-672.

Czeresnia D, Ribeiro AM. O conceito de espaço em Epidemiologia: uma interpretação histórica e epistemiológica. Cadernos de Saúde Pública; 2000. 16(3): 595-617.

DATASUS. Informações em saúde: bancos de dados do Sistema Único de Saúde. 2003. Disponível em: http://tabnet.datasus.gov.br/tabnet/tabnet.htm. Acesso em: 10/04/2008.

Davis CA, Fonseca F, Borges KAV. A flexible addressing system for approximate urban geocoding. In: V Simpósio Brasileiro de Geoinformática (GeoInfo 2003). Campos de Jordão. 2003.

Duncan M, Mummery K. Psychosocial and environmental factors associated with physical activity among city dwellers in regional Queensland. Preventive Medicine. 2004. 40: 363-372.

Ephraim Y, Merhav N. Hidden Markov processes. IEEE Transactions on Information Theory. 2002. 48(6): 1518-1569.

Eichelberger P. The Importance of Adresses – The Locus of GIS. Proceedings of the URISA 1993 Annual Conference. 1993. 200-211.

ESRI. White Paper. ESRI Shapefile Technical Description. United States of América, 1998. Disponível em: <http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf>.

Fair M. Generalized Record Linkage System – Statistics Canada's Record linkage software. Austrian Journal of Statistics. 2004. 33(1): 37-53.

Fellegi L, Sunter A. A Theory for Record Linkage. Journal of the American Statistical Society. 1969. 64: 1183–1210.

Francis AM, Schneider JB. Using computer graphics to map origindestination data describing health care delivery system. Social Science and Medicine. 1984. 18: 405-420.

Gadd TN. PHONIX: The algorithm. Program: electonic library and information systems. 1990. 24(4): 363-366.

Gandrabur S, Foster G. 2003. Confidence estimation for text prediction. In: Procedures of Conference on Natural Language Learning (CoNLL). Canada. 2003. 95–102.

Goldberg D, Wilson J, Knoblock C, Ritz B, Cockburn M. An effective and efficient approach for manually improving geocoded data. International Journal of Health Geographics. 2008. 7(1): 60.

Google Maps. Referência do KML. Disponível em http://code.google.com/intl/pt-BR/apis/kml/documentation/kmlreference.html. Acessado em 01/06/2009.

Gregorio DI, Dehello LM, Samociuk H, Kulldorff M. Lumping or splitting: seeking the preferred areal unit for health geography studies. International Journal of Health Geographics 2005. 4: 6.

Hall PAV, Dowling GR. Approximate String Matching. ACM Computing Surveys (CSUR). 1980. 12(4): 381-402.

Han D, Rogerson PA, Bonner MR, Nie J, Vena JE, Muti P, Trevisan M, Freudenheim JL. Assessing spatio-temporal variability of risk surfaces using residential history data in a case control study of breast cancer. International Journal of Health Geographics. 2005. 4: 9.

Hastie TJ, Tibshirani RJ. Generalized Additive Models. London: Chapman & Hall. 1990.

Hill L. Georeferencing. Cambridge, MA, MIT Press. 2006.

Hippocrates et al. Airs, Waters, Places. In: Lloyd Ger (Ed.). Hippocratic Writtings. London: Penguin Books, 1983; 148-169.

Huff NC, Gray D. Coronary heart disease inequalities: deaths and the socio-economic environment in Nottingham, England. Health & Place 2001. 7: 57-61.

Hurley SE, Saunders TM, Nivas R, Hertz A, Reynolds P. Post Office Box Address: a challenge for Geographic Information System-based studies. Epidemiology. 2003. 14: 386-391.

Hyndman JCG, Holman CDJ. Accessibility and spatial distribution of general practice services in an Australian city by levels of social disadvantage. Social Science & Medicine 2001. 53: 1599-1609.

IBGE (Fundação Instituto Brasileiro de Geografia e Estatística). Censo demográfico do Brasil 2000. IBGE. Rio de Janeiro. 2002.

IBGE (Fundação Instituto Brasileiro de Geografia e Estatística). Cadastro Nacional de Endereços para Fins Eststísticos CNEFE – Manual do Cadastro. IBGE. Rio de Janeiro. 2005.

Jacquez GM. Spatial analysis in epidemiology: Nascent science or a failure of GIS? Journal of Geographical Systems. 2000. 2(1): 91-97.

Jaro MA. Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Society. 1989. 84(406): 414–420.

Kaufman JS, Dole N, Savitz DA, Herring AH. Modeling community-level effects on preterm birth. Ann Epidemiol 2003. 13: 377-384.

Kinman EL. Evaluating health service equity at a primary care clinic in Chilimarca, Bolivia. Social Science & Medicine. 1999. 49 :663-678.

Kondrak G. Phonetic alignment and similarity. Edmonton, Canada, 2003. Disponível em: < http://www.cs.ualberta.ca/~kondrak/papers/chum.pdf >.

Korte G. The GIS book. On World Press, Santa Fé. 1994.

Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. American Journal of Public Health. 2001. 91: 1114-1116.

Krieger N, Chen T, Waterman D, Chen JT, Soobader M, Subramanian SV. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). J Epidemiol Community Health. 2003. 57: 186-199.

Krieger N, Waterman PD, Chen JT, Soobader M, Subramanian SV. Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis, and violence: geocoding and choice of area-based socioeconomic measures--the public health disparities geocoding project (US). Public Health Rep. 2003. 118(3): 240-260.

Krieger N, Chen JT, Waterman PD, Rehkopf H, Subramanian SV. Painting a truer picture of US socioeconomic and racial/ethnic health inequalities: the Public Health Disparities Geocoding Project. Am J Public Health. 2005. 95(2): 312-323.

Laraia BA, Siega-Riz AM, Kaufman JS, Jones SJ. Proximity of supermarkets in positively associated with diet quality index of pregnancy. Preventive Medicine. 2004. 39: 869-875.

Lazaridis I, Mehrotra S. Approximate Selection Queries over Imprecise Data. Pages 140–152 of: Proceedings of the 20th International Conference on Data Engineering, ICDE 2004. Boston, MA, USA. 2004.

Lemos-Dias T. Modelo de sistemas viáveis em organizações públicas: um estudo de caso da função de planejamento de informações estratégicas para informatização

da Secretaria Municipal de Saúde de Belo Horizonte. Dissertação (Mestrado) — Escola de Governo, Fundação João Pinheiro, Belo Horizonte. 1998.

Lemos-Dias T, Oliveira MPG, Câmara G, Carvalho MS. Problemas de escala e a relação área-indivíduo em análise espacial de dados censitários. Informática Pública. 2002. 4: 89-104

Levine N, Kim KE. The location of motor vehicle crashes in Honolulu: a methodology for geocoding intersections. Comput. Environ. And Urban Systems. 1998 6: 557-576.

MacEachren AM, Boscoe FP, Haug D, Pickle LW. Geographic Visualization: Designing manipulable maps for exploring temporally varying georeferenced statistics. Proceedings of the IEEE Information Visualization Symposium, Research Triangle Park, NC. 1998. 87-94.

Maguire DJ. An overview and definition of GIS. In: Maguire, D.J., Goodchild, M.F., Rhind, D.W. (eds), Geographical Informations Systems: Principles and Applications, v. 1, Longman. 1991. 9-20.

Marble D. An introduction to the structure design of GIS. USA. 1994.

Martins EB. Peso ao nascer e determinants ecológicos nos padrões nutricionais de crianças. Tese de Doutorado em Saúde Pública. ENSP/Fiocruz. 2007.

Martins EB, Carvalho MS. Associação entre peso ao nascer e o excesso de peso na infância: revisão sistemática. Cad. Saúde Pública. 2006. 22(11): 2281-2300.

McElroy JA, Remington PL, Trentham-Dietz A, Robert SA, Newcomb PA. Geocoding addresses from a large population-based study: lessons learned. Epidemiology. 2003. 4: 399-407.

Morad M. British standard 7666 as a framework for geocoding land and property information in UK. Computers, Environment and Urban Systems. 2002. 26: 483-492.

Morgenstern H. Ecologic studies. In: Rothman KJ, Greenland S, eds. Modern epidemiology., 2nd ed. Philadelphia: Lippincott. 1998. 459-480.

Navarro, G. A Guided Tour to Approximate String Matching. In: ACM Computing Surveys. 2001. 33(1): 31-88.

Newcombe, HB, Kennedy JM, Axford SJ, James AP. Automatic Linkage of Vital Records. Science. 1959. 130: 954-959.

Nunes FC. Análise Espacial da Leptospirose na Cidade de Savador-Bahia, no período de 1996-2006. Tese de Doutorado – ENSP/Fiocruz, Rio de Janeiro. 2007.

O Globo – Jornal dos Bairros. Quando a polêmica chega ao limite. Edição de 14/05/2009. Rio de Janeiro. 2009. 8-10.

Oliveira EXG, Travassos C, Carvalho MS. Territórios do Sistema Único de Saúde: mapeamento das redes de atenção hospitalar. Cad Saúde Pública. 2004. 20: 386-402.

Openshaw S, Charlton M, Wymer C, Craft A, Mark I. Geographical analysis machine for the automated analysis of point data sets. International Journal of Geographic Information Systems. 1987. 1: 35-51.

Oyana TJ, Rivers PA. Geographic variations of childhood asthma hospitalization and outpatient visits and proximity to ambient pollution sources at U.S.-Canada border crossing. International Journal of Health Geographics. 2005. 4: 14.

O'Reagan RT, Saalfeld AJ. Geocoding Theory and Practice at the Bureau of the Census. Washington, DC: Bureau of the Census Statistical Research Division Report Series. 1987.

Paull D. A geocoded national address file for Australia: the G-NAF what, why, who and when? 2003. Acessível em http://www.g-naf.com.au

Peleg K, Pliskin JS. A Geographic Information System simulation model of EMS: reducing ambulance response time. American Journal of Emergency Medicine. 2004. 22: 164-170.

Philips L. Hanging on the Metaphone. Computer Language. 1990. 7: 39-43.

Pickett KE, Pearl M. Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. J Epidemiol Community Health. 2001. 55(2): 111-122.

Pinheiro RS, Travassos C, Gamerman D, Carvalho MS. Mercados hospitalares em área urbana: uma abordagem metodológica. Cad Saúde Pública. 2001. 17: 1111-1121.

Pittman J, Andrews H, Struening E. The use of zip coded population data in social area studies of service utilization. Eval Program Plann. 1986. 9(4): 309-317.

Portugal JL. Integração SIAB e SIG: uma proposta para o funcionamento do programa de saúde da família. Tese de Doutorado – Centro de Pesquisas Aggeu Magalhães, Fundação Oswaldo Cruz, Recife. 2003.

Rabiner LR. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE. 1989. 77(2).

Ramos FR, Câmara G, Monteiro AMV. Territórios digitais urbanos. In: Almeida MA, Câmara G, Monteiro AMV, organizadores. Geoinformação em urbanismo: cidade real x cidade virtual. Oficina de Textos. São Paulo. 2007. 34-53.

Ratcliffe JH. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. Int J Geogr Inf Sci. 2001. 15: 473-485.

RIPSA (Rede Integrada de Informações Para a Saúde). Compatibilização de sistemas e base de dados (CBD) da Rede Integrada de Informações para a saúde (RIPSA) - Informe final, 1997. Informe Epidemiológico do SUS. 1997. 6: 35-41.

Rodrigues M, Quintanilha JA. A seleção de software SIG para gestão urbana. In: Anais do XV Congresso Brasileiro de Cartografia. São Paulo. SBC. 1991. 513–519.

Rose G. Sick individuals and sick populations — with discussion. Int J Epidemiol. 2001. 30: 427-432.

Rose KM, Wood JL, Knowles S, Pollitt RA, Whitsel EA, Die-Roux AV, Yoon D, Heiss S. Historical measures of social context in life course studies: retrospective linkage of addresses to decennial censuses. International Journal of Health GeograPhics. 2004. 3:27.

Rosero-Bixby L. Saptial access to health care in Costa Rica and its equity: a GIS-based study. Social Science & Medicine. 2004. 58: 1271-1284.

Rushton G, Armstrong MP, Gittler J. Geocoding in Cancer Research: A Review. American Journal of Preventive Medicine. 2006. 30(2): S16-S24.

Rutt CD, Coleman KJ. Examining the relationships among built environment, physical activity, and body mass index in El Paso, TX. Preventive Medicine. 2005. 40: 831-841.

Rytkonen M, Rusanen J, Nayha S. Small-area variation in mortality in the city of Oulu, Finland, during the period 1978-1995. Health Place. 2001. 7: 75-79.

Sabroza PC, Leal MC. Saúde, ambiente e desenvolvimento. Alguns conceitos fundamentais. In: Saúde, ambiente e desenvolvimento (Leal MC, Sabroza PC, Rodrigues R, Buss P, org.). Abrasco. Rio e Janeiro. 1992. 45-93.

Samantha C, Martin D. Zone design for environment and health studies using pre-aggregated data. Social Science & Medicine. 2005. 60: 2279-2742.

Santos M. A Natureza do Espaço. Editora Hucitec, São Paulo. 1996.

Santos SM, Barcellos C, Carvalho MS, Flores R. Detecção de aglomerados espaciais de óbitos por causas violentas em Porto Alegre, Rio Grande do Sul, Brasil, 1996. Cad Saúde Pública; 2001. 17: 1141-1151.

Santos SM. A importância do contexto social de moradia na auto-avaliação de saúde [tese]. Rio de Janeiro (RJ): Escola Nacional de Saúde Pública. 2008.

Scholten GI. Stillwell JH. Geographical information systems for urban and regional planning. London: Kluwer Academic. 1990.

Schwartz S. The fallacy of the ecological fallacy: the potencial misuse of a concept and the consequences. American Journal of Public Health; 1994. 84: 819-824.

Skaba DA, Terron SL. Mapas urbanos digitais do censo 2000: uma abordagem tecnológica. Revista IP – Informática Pública. 2003. 5: 205-219.

Skaba DA, Carvalho MS, Barcellos C, Martins PC, Terron SL. Geoprocessamento dos dados da saúde: o tratamento dos endereços. Cadernos de Saúde Pública 2004. 20: 1753-1756.

Skelly C, Wendy B, Hearnden M, Eyles R, Weinstein P. Disease surveillance in rural communities is compromised by address geocoding uncertainty: a case study of campylobacteriosis. Aust. J. Rural Health. 2002. 10: 87-93.

Smallman-Raynor MR, Cliff AD. Civil war and the spread of AIDS in Central Africa. Epidemiological Infections. 1991. 107: 69–80.

Smith TF, Waterman MS. Identification of common molecular subsequences. J. Mol. Biol. 1981. 147: 195–197.

Snow J. Sobre a Maneira de Transmissão do Cólera. Hucitec/Abrasco. São Paulo/Rio de Janeiro. 1990.

Souza WV, Barcellos C, Brito AM, Carvalho MS, Cruz OG, Albuquerque MFM, et al. Aplicação de modelo bayesiano empírico na análise espacial da ocorrência de hanseníase. Rev. Saúde Pública. 2001. 35: 474-480.0

Susser M. The logic of Ecological: I. The logic of analysis. American Journal of Epidemiology. 1994. 84(5): 825-829.

Tassinari WS. Modelagem espacial, temporal e longitudinal: diferentes abordagens do estudo de leptospirose. Tese de Doutorado – ENSP/Fiocruz, Rio de Janeiro. 2009.

Tassinari WS, Pellegrini D, As C, Reis R, Ko AI, Carvalho MS. Detection and modelling of case clusters for urban leptospirosis. Tropical Medicine and International Health. 2008. 13(4):1-10.

UNBC GIS Lab, Map Symbolization, 5. Labelling Features, (online). http://www.gis.unbc.ca/courses/geog205/labs/lab5/index.php. Acessado em 05/12/2008.

U.S. Census Bureau; 108th CD Census 2000 TIGER/Line Files Technical Documentation. Internet: http://www.census.gov/geo/www/tiger/tgrcd108/tgr108cd.pdf. Acessado em 05/02/2009.

USGS – US Geological Survey. Internet: www.usgs.gov. Acessado em 15/03/2009.

Waller LA, Gotway CA. Applied Spatial Statistics for Public Health Data. New York: John Wiley &Sons. 2004.

Wilmersdorf E. Geocoded information incorporated into urban online services – the approach of the City of Viena. Computers, Environemnt and Urban Systems. 2003. 27: 609-621.

Wood SN. Stable and efficient multiple smoothing parameter estimation for Generalized Additive Models. Journal of American Statistical Association. 2004. 99: 673-6896

Wrigley N, Hold T, Steel D, Tranmer, M. Analysing, modeling, and resolving the ecological fallacy In: Longley P, Batty M. Spatial Analysis: Modelling in a GIS Environment. John Wiley & Sons. 1996.

Zobel J, Dart P. Phonetic string matching: lessons from information retrieval, Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. Switzerland. 1996. 166-172.

ANEXO 1 – TABELAS AUXILIARES

TABELA DE TIPOS

Codigo	Nome	Codigo	Nome	Codigo	Nome
1	ACAMPAMENTO	51	ESTACAO	101	QUADRA
2	ACESSO	52	ESTACIONAMENTO	102	QUINTA
3	ADRO	53	ESTADIO	103	RAMAL
4	AEROPORTO	54	ESTANCIA	104	RAMPA
5	ALAMEDA	55	ESTRADA	105	RECANTO
6	ALTO	56	FAVELA	106	REDE ELETRICA
7	AREA	57	FAZENDA	107	RESIDENCIAL
8	ARTERIA	58	FEIRA	108	RETA
9	ATALHO	59	FERROVIA	109	RETIRO
10	ATERRO	60	FONTE	110	RETORNO
11	AUTODROMO	61	FORTE	111	RIO
12	AVENIDA	62	FREGUESIA	112	RODO ANEL
13	BAIA	63	GALERIA	113	RODOVIA
14	BAIRRO	64	GRANJA	114	ROTATORIA
15	BAIXA	65	HABITACIONAL	115	ROTULA
16	BALAO	66	HIPODROMO	116	RUA
17	BALNEARIO	67	ILHA	117	RUELA
18	BECO	68	JARDIM	118	SERRA
19	BELVEDERE	69	LADEIRA	119	SERTAO
20	BLOCO	70	LAGO	120	SERVIDAO
21	BOSQUE	71	LAGOA	121	SETOR
22	BOULEVARD	72	LARGO	122	SITIO
23	BURACO	73	LIMITE	123	SUBIDA
24	CAIS	74	LINHA DE TRANSMISSAO	124	SUPERQUADRA
25	CALCADA	75	LOTEAMENTO	125	TERMINAL
26	CAMINHO	76	MANGUE	126	TERRENO
27	CAMPO	77	MARGEM	127	TRANSVERSAL
28	CANAL	78	MARINA	128	TRAVESSA
29	CARTODROMO	79	MODULO	129	TRECHO
30	CHACARA	80	MONTE	130	TREVO
31	CHAPADAO	81	MORRO	131	TRINCHEIRA
32	CIDADE	82	NUCLEO	132	TUNEL
33	CIRCULAR	83	PARADA	133	UNIDADE
34	COLONIA	84	PARADOURO	134	VALA
35	COMPLEXO VIARIO	85	PARALELA	135	VALE
36	CONDOMINIO	86	PARQUE	136	VARGEM
37	CONJUNTO	87	PASSAGEM	137	VARIANTE
38	CORREDOR	88	PASSARELA	138	VELODROMO
39	CORREGO	89	PASSEIO	139	VEREDA
40	DESCIDA	90	PATIO	140	VIA
41	DESVIO	91	PLANALTO	141	VIA EXPRESSA
42	DISTRITO	92	PLATAFORMA	142	VIADUTO
43	EDIFICIO	93	PONTA	143	VIELA
44	ELEVADO	94	PONTE	144	VILA
45	ENTRADA PARTICULAR	95	PORTO	145	ZIGUE-ZAGUE
46	ENTREPOSTO	96	POSTO	146	CICLOVIA
47	ENTRONCAMENTO	97	PRACA	147	COMUNIDADE
48	ESCADARIA	98	PRACA DE ESPORTES	148	ENTRADA
49	ESCADINHA	99	PRAIA		
50	ESPLANADA	100	PROLONGAMENTO		

TABELA DE TÍTULOS

Codigo	Nome	Codigo	Nome	Codigo	Nome
1	ABADE	51	COMENDADOR	101	FILHAS
2	ACADEMICO	52	COMERCIANTE	102	FILHO
3	ADVOGADA	53	COMISSARIA	103	FILHOS
4	ADVOGADO	54	COMISSARIO	104	FISCAL
5	AJUDANTE	55	COMODORO	105	FISICO
6	ALFERES	56	COMPOSITOR	106	FOTOGRAFO
7	ALMIRANTE	57	COMPOSITORA	107	FRADE
8	ALUNA	58	CONDE	108	FREI
9	ALUNO	59	CONDESSA	109	FREIRA
10	ANCIAO	60	CONEGO	110	FUNCIONARIA
11	ANSPECADA	61	CONSELHEIRO	111	FUNCIONARIO
12	APOSTOLO	62	CONSTRUTOR	112	GENERAL
13	ARCEBISPO	63	CONSUL	113	GOVERNADOR
14	ARCIPRESTE	64	CONTABILISTA	114	GRAO
15	ARQUIDUQUE	65	CORONEL	115	GRUMETE
16	ARQUIDUQUESA	66	CORONEL-AVIADOR	116	GUARDA
17	ARQUITETA	67	CORRETOR	117	GUIA
18	ARQUITETO	68	DEFENSOR PUBLICO	118	HISTORIADOR
19	ASPIRANTE	69	DELEGADO	119	IMACULADA
20	ATENDENTE	70	DENTISTA	120	IMPERADOR
21	ATLETA	71	DEPUTADA	121	IMPERATRIZ
22	ATOR	72	DEPUTADO	122	INDUSTRIAL
23	ATRIZ	73	DESEMBARGADOR	123	INFANTE
24	AVIADOR	74	DESPACHANTE	124	INSPETOR
25	AVIADORA	75	DESPORTISTA	125	INTENDENTE
26	BACHAREL	76	DETETIVE	126	IRMA
27	BANCARIO	77	DIACONO	127	IRMAO
28	BANDEIRANTE	78	DOM	128	IRMAOS
29	BARAO	79	DONA	129	JARDINEIRO
30	BARONESA	80	DOUTOR	130	JESUITA
31	BISPO	81	DOUTORA	131	JORNALISTA
32	BOMBEIRO	82	DUQUE	132	JUIZ
33	BRIGADEIRO	83	DUQUESA	133	JUNIOR
34	CABO	84	ECONOMISTA	134	LEGIONARIO
35	CABOCLO	85	EMBAIXADOR	135	LEILOEIRA
36	CACIQUE	86	EMBAIXATRIZ	136	LEILOEIRO
37	CADETE	87	ENFERMEIRA	137	LIVREIRO
38	CANTOR	88	ENFERMEIRO	138	LOCUTOR
39	CAPELAO	89	ENGENHEIRA	139	LORDE
40	CAPITAO	90	ENGENHEIRO	140	MADAME
41	CAPITAO-AVIADOR	91	ESCOTEIRO	141	MADRE
42	CAPITAO-MOR	92	ESCRAVA	142	MAE
43	CARDEAL	93	ESCRITOR	143	MAESTRO
44	CARTEIRO	94	ESCRITORA	144	MAGISTRADO
45	CAVALHEIRO	95	ESCRIVAO	145	MAJOR
46	CHANCELER	96	ESTATISTICO	146	MAJOR BRIGADEIRO
47	CHEFE	97	ESTUDANTE	147	MAJOR-DOUTOR
48	CIENTISTA	98	EXPEDICIONARIO	148	MAQUINISTA
49	CINEASTA	99	FARMACEUTICO	149	MARECHAL
50	COMANDANTE	100	FERROVIARIO	150	MARINHEIRO

TABELA DE TÍTULOS (continuação)

Codigo	Nome	Codigo	Nome	Codigo	Nome
151	MARQUES	201	RADIALISTA	251	VISCONDESSA
152	MARQUESA	202	RAINHA	252	VIUVA
153	MEDICO	203	REGENTE	253	VIUVO
154	MENINO	204	REI	254	VOLUNTARIO
155	MESTRE	205	REITOR	255	VOVO
156	METALURGICO	206	REPORTER		
157	MINISTRO	207	REVERENDO		
158	MISS	208	SACRISTAO		
159	MISSIONARIO	209	SAGRADO		
160	MISTER	210	SANTA		
161	MONSENHOR	211	SANTO		
162	MOTORISTA	212	SAO		
163	MUSICO	213	SARGENTO		
164	NETO	214	SARGENTO-MOR		
165	NOSSA SENHORA	215	SECRETARIA		
166	NOSSO SENHOR	216	SEGUNDO-SARGENTO		
167	NUNCIO	217	SEGUNDO-TENENTE		
168	OPERARIA	218	SEMINARISTA		
169	OPERARIO	219	SENADOR		
170	ORGANISTA	220	SENHOR		
171	OUVIDOR	221	SENHORA		
172	PADRE	222	SENHORIA		
173	PAI	223	SENHORITA		
174	PAPA	224	SERTANISTA		
175	PARTEIRA	225	SEU		
176	PASTOR	226	SINDICALISTA		
177	PESCADOR	227	SINHA		
178	PILOTO	228	SOBRINHA		
179	PINTOR	229	SOBRINHO		
180	PINTORA	230	SOLDADO		
181	PIO	231	SOROR		
182	POETA	232	SUB-OFICIAL		
183	PRACINHA	233	SUB-TENENTE		
184	PREFEITO	234	TABELIAO		
185	PRESIDENTE	235	TENENTE		
186	PRIMEIRO-SARGENTO	236	TENENTE-AVIADOR		
187	PRIMEIRO-TENENTE	237	TENENTE-CORONEL		
188	PRINCESA	238	TERCEIRO-SARGENTO		
189	PRINCIPE	239	TIA		
190	PROCURADOR	240	TIO		
191	PROCURADORA	241	TIPOGRAFO		
192	PROFESSOR	242	TOPOGRAFO		
193	PROFESSORA	243	TROVADOR		
194	PROFETA	244	VEREADOR		
195	PROMOTOR	245	VICE		
196	PROMOTORA	246	VICE-GOVERNADOR		
197	PROVEDOR	247	VIGARIO		
198	QUIMICA	248	VIGILANTE		
199	QUIMICO	249	VIRGEM		
200	RABINO	250	VISCONDE		

ANEXO 2 – ARTIGO: GEOPROCESSAMENTO DOS DADOS DA SAÚDE: O TRATAMENTO DOS ENDEREÇOS

GEOPROCESSAMENTO DOS DADOS DA SAÚDE: O TRATAMENTO DOS ENDEREÇOS

Geoprocessing of health data: the addresses information treatment

Daniel Albert Skaba
Marilia Sá Carvalho
Christovam Barcellos
Paulo Cesar Martins
Sonia Luiza Terron

Abstract

This paper analyzes the actual stage of the address data in the Brazilian Health Information Systems, with a view to map large cities health events in Geographic Information Systems, for risk analysis and evaluation. Therefore it is necessary to perform the geocoding of these events to small geographic areas inside the urban limits. This study used a sample of the SINAN data base, and also proposes alternatives to work with this large amount of events.

Key words Health Information System; Address; Geographic Information Systems

Resumo

Este trabalho faz uma análise da situação atual das informações de endereços nos Sistemas de Informações em Saúde (SIS) no Brasil, visando sua utilização em Sistemas de Informações Geográficas (SIG), para a análise e avaliação de riscos dos eventos de saúde pública em grandes cidades, com localização destes eventos em áreas intra-urbanas. Utiliza como base de dados uma amostra dos cadastros do Sistema de Informação sobre Agravos de Notificação (SINAN) e tem como objetivo, propor alternativas para aproveitamento de grandes volumes de dados já existentes.

Palavras-chave Sistemas de Informação em Saúde; Endereço; Sistemas de Informações Geográficas

Introdução

O georreferenciamento dos eventos de saúde é o ponto de partida para a análise e avaliação de riscos, já que as causas dos problemas de saúde coletiva estão relacionadas com o meio ambiente e a população no entorno dos locais em que os eventos acontecem. Os Sistemas de Informações Geográficas (SIG), conjunto de ferramentas utilizadas para a manipulação de informações espacialmente apresentadas[1], permitem o mapeamento das doenças e contribuem na avaliação dos riscos[2,3]. Para estas análises é necessária a localização geográfica dos eventos, para associação destas informações (gráficas) a bases de dados (alfanuméricos). O georreferenciamento de um endereço, definido como o processo de associação deste a um mapa terrestre, pode ser efetuado de três formas básicas: associação a um ponto, a uma linha ou a uma área[4]. O elemento geométrico resultante, associado a uma base de dados, é a unidade utilizada nos SIG.

Diversos trabalhos utilizam como fonte de dados os registros dos diversos Sistemas de Informação em Saúde (SIS)[5], que contam com séries históricas de até 20 anos, em alguns casos. Nestas análises utiliza-se, na maioria dos casos, o município ou o bairro informado, para visualização da distribuição espacial dos eventos a serem estudados. Nos municípios mais densamente povoados esta escala geográfica já é insuficiente, sendo necessária a localização de áreas menores, principalmente nas áreas urbanizadas. Para aumentar a escala da análise, a utilização dos setores censitários vem sendo proposta em diversos trabalhos[6], utilizando como fonte de informação o endereço dos registros dos SIS, georreferenciados para setor censitário.

O objetivo deste trabalho é avaliar a qualidade dos dados, principalmente os endereços, armazenados nos Sistemas de Informação sobre Agravos Notificados (SINAN), a partir da análise de uma amostra destes, e propor alternativas para sua utilização, em grandes volumes, em SIG, de forma a viabilizar, no país, a análise espacial por micro-áreas.

Material e Metodologia

Bases de Dados de Referência

Para servir de base para localização dos endereços, foram utilizados o Cadastro de Segmentos de Logradouros (Cadlog) e o arquivo preliminar das Folhas de Coleta, ambos criados a partir do Censo 2000 do IBGE. O Cadlog apresenta uma relação de logradouros por Setor Censitário, com numeração de início e final do segmento, enquanto as Folhas de Coleta apresentam os endereços das unidades visitadas no Censo. Para a localização manual, além dos cadastros, foram utilizados os mapas de localidades e de setores do Censo 2000.

Bases da Saúde

Neste trabalho foram utilizados três arquivos de doenças transmissíveis, provenientes do SINAN, oriundos de: Campinas, com 410 registros de dengue; Macapá, com 1140 registros de hepatite e Rio de Janeiro com 2544 casos de leptospirose.

Metodologia

Este trabalho foi desenvolvido em quatro fases. A primeira fase consistiu na compatibilização das tabelas dos SIS com as tabelas bases de consultas (cadastro de segmentos de logradouros e folhas de coleta do IBGE), separando tipo, título e nome do logradouro e número da unidade de residência em sub-campos. Neste procedimento são utilizadas tabelas de tipos de logradouros (rua, avenida, travessa, etc.), e de títulos (doutor, governador, presidente, etc.), além das abreviações conhecidas destes, geradas pelo IBGE para o Censo 2000. Para o processo foi desenvolvido um aplicativo específico, adaptado às características de cada tabela fonte da informação.

Na segunda fase foi feita a busca automática dos endereços, que se deu em vários níveis, a partir da combinação entre os sub-campos da informação de endereço, a saber:

Tipo, título, nome e número;

Tipo, título e nome;

Tipo, nome e número;

Tipo e nome;

Título, nome e número;

Nome e número;

Nome

O resultado desta busca leva à localização de apenas um setor, mais de um setor ou nenhum setor. Nos dois últimos casos, passa-se às próximas fases de busca manual.

A terceira fase foi uma pesquisa manual dos endereços que apontaram para mais de um setor censitário. Foi uma operação relativamente rápida, facilitada pela relação de setores selecionados pela fase anterior e utilizando outras informações contidas na tabela, como os pontos de referência e complementos. A quarta e última fase consistiu na pesquisa manual dos endereços em que não há setores indicados automaticamente.

Resultados

Na tabela 1 são apresentados os resultados da pesquisa com a amostra selecionada. Na coluna de busca automática é apresentado o número de casos em que foi encontrado apenas um setor automaticamente. Na coluna seguinte, de busca manual, é apresentado o número de casos em que foi possível especificar o setor censitário a que

pertence o endereço em uma pesquisa manual aos cadastros e mapas. Na penúltima coluna (não encontrado) estão os casos em que os endereços estão incompatíveis com os cadastros e mapas ou poderiam estar em mais de um setor.

Tabela 1. Distribuição dos resultados do georreferenciamento e por arquivo

Arquivo	Em branco	Busca automática	Busca manual	Não encontrado	Total
Campinas (dengue)	19 (5%)	202 (49%)	139 (34%)	50 (12%)	410
Macapá (hepatite)	1 (0%)	821 (72%)	228 (20%)	90 (8%)	1.140
Rio de Janeiro (leptospirose)	95 (4%)	1.233 (48%)	421 (17%)	795 (31%)	2.544
TOTAL	115 (3%)	2.256 (55%)	788 (19%)	935 (23%)	4.094

Discussão

Em uma primeira abordagem verificou-se que os três arquivos pesquisados apresentam diferenças no modo de armazenamento do endereço. No arquivo de dengue de Campinas, o endereço está todo contido em um único campo, inclusive as informações complementares. No de leptospirose do Rio de Janeiro, o ponto de referência está em um campo separado e há um outro campo com a informação de bairro. Por último, no arquivo de hepatite de Macapá, o endereço está composto por quatro campos: nome do logradouro, número de porta, complemento e referência. Estas diferenças implicam em procedimentos específicos na primeira fase do desenvolvimento (compatibilização das tabelas).

Alguns fatores observados no processo são determinantes no resultado encontrado. Um deles é a qualidade dos dados coletados. Quando comparados os três exemplos, verifica-se que na amostra de Macapá não há praticamente informação de endereço em branco, nas outras duas cidades há em torno de 5%. Nestes não há qualquer possibilidade de se mapear o evento dentro do município. Outro fator importante é a natureza do evento. Algumas doenças, como é o caso, por exemplo, da leptospirose, têm freqüência muito maior em áreas sem um bom saneamento básico, como periferias e favelas, locais onde os endereços não apresentam regularidade, havendo grande perda de informação. O tratamento de endereços em favelas, acampamentos, invasões é um problema que merece estudo especial.

No Brasil não há um padrão único de endereços. Em algumas poucas cidades, como Belo Horizonte foi feito um trabalho[7] de cadastramento e compatibilização dos endereços da cidade. Mas, na maioria das cidades, as regras básicas de padronização de endereços não são obedecidas, principalmente nas periferias das cidades e nas áreas de ocupação irregular, como as favelas. Algumas cidades apresentam alto índice de numeração irregular e outras (como Brasília e Palmas) utilizam endereços por quadra e não por logradouro. Além disto, os mapas urbanos digitais, quando disponíveis, não seguem um padrão único.

Para o Censo 2000, a Coordenação de Estruturas Territoriais da Diretoria de Geociências do IBGE (DGC-CETE) desenvolveu um projeto, denominado Base Territorial do censo 2000[8], que resultou em produtos digitais padronizados para todo o país. Entre estes produtos estão 19.000 mapas de localidades urbanas, em formato CAD,

e as malhas de setores urbanos das 1048 maiores cidades brasileiras, além do Cadlog destas cidades.

Todos os produtos resultantes deste trabalho têm como unidade de referência o Setor Censitário, definido como "*a unidade de coleta de dados dos Censos formada por área contínua, situada em um único quadro urbano ou rural, com dimensão e número de domicílios ou de estabelecimentos que permitam, segundo cronograma estabelecido, o levantamento das informações por um único agente credenciado*"[9]. Seus limites respeitam os limites territoriais legalmente definidos (distritos, bairros, etc.), e os estabelecidos pelo IBGE para fins estatísticos (aglomerados rurais, aglomerados subnormais entre outras).

Para a redução da quantidade de endereços não compatíveis com os cadastros que servem de base para sua localização nos mapas, é fundamental aprimorarem-se os processos de captura de dados nos locais de atendimento. Neste sentido há necessidade de uma padronização desta entrada de dados, criando-se procedimentos assistidos por computador, através de cadastros que sirvam de fonte dirigida de informação. Toda esta ação requer uma união de esforços dos órgãos responsáveis nos diversos níveis e setores de governo envolvidos, além de produtores privados.

As melhorias resultantes das ações mencionadas serão vistas só nos eventos ocorridos após sua implantação. Para os dados já armazenados há uma série de trabalhos a serem desenvolvidos que possibilitam atingir um melhor percentual de localização dos eventos nas áreas intra-urbanas. Neste contexto, está inserida a compatibilização de cadastros de logradouros existentes em prefeituras, nos Correios, nas concessionárias de serviços públicos entre outros. Como exemplos de informações importantes para estas localizações, pode-se citar o CEP, nomes antigos de logradouros e tratamento fonético dos nomes.

Para os logradouros com numeração irregular, não sequencial, muito comuns na maioria das cidades do Brasil, é importante a existência de cadastros que contenham uma numeração individual de porta associada à unidade de pesquisa. Uma boa alternativa, testada nesta pesquisa, é a utilização das folhas de coleta do Censo 2000, que contêm os endereços com CEP e número de porta de todas as unidades visitadas no Censo, por setor censitário.

Uma outra alternativa, para uma melhor determinação da unidade de pesquisa em que está localizado o endereço, é a criação de unidades compostas por conjuntos de setores adjacentes e homogêneos segundo parâmetros pertinentes às análises desejadas. Deste modo, são facilitadas as localizações quando um endereço, sem o número de porta ou em logradouros com numeração irregular, pode estar contido em mais de um setor.

O problema aqui analisado tem impacto substantivo em diversas questões, que vão de estudos e pesquisas onde o mote é a análise geográfica, às atividades de vigilância epidemiológica, sendo particularmente relevante na perspectiva da construção de uma vigilância em saúde de base territorial, integrando a ocorrência dos agravos registradas nos diferentes SIS com aspectos ambientais relevantes.

Referências

1. ARONOFF, Stan, 1990. Geographic Information Systems: A Management Perspective. Canadá, WDL publications.

2. BARCELLOS, Christovam & RAMALHO, Walter, 2002. Situação Atual do Geoprocessamento e da Análise de Dados Espaciais em Saúde no Brasil. Revista Informática Pública, 4: 221-230.

3. RICHARDS, Thomas B.; RUSHTON, Gerard; BROWN, Carol K. & FOWLER, Littleton, 1999. Geographic Information and Public Health: Mapping the Future. Public Health Reports, 114: 359-373, Oxford University Press.

4. EICHELBERGER, P., 1993. The Importance of Adresses – The Locus of GIS. Proceedings of the URISA 1993 Annual Conference, 200-211. Atlanta GA.

5. Fundação Nacional de Saúde, Ministério da Saúde. Sistemas de Informação em Saúde. Internet: http://www.funasa.gov.br

6. CARVALHO, M. S. & CRUZ, O. G, 1998. Análise Espacial por Microáreas: Métodos e Experiências. In: Epidemiologia Contextos E Pluralidade (R. P. Veras, M. L. Barreto & N. Almeida Filho, org.), pp. 79-89, Rio de Janeiro: Editora Fiocruz.

7. OLIVEIRA, C. M., 2003. Lançamento de endereços no Geoprocessamento de Belo Horizonte. Anais do XXI Congresso Brasileiro de Cartografia, em CD.

8. SKABA, D. A. & TERRON, S. L., 2003. Mapas Urbanos Digitais do Censo 2000: Uma Abordagem Tecnológica. Informática Pública, 5 (2): 205-219.

9. Instituto Brasileiro de Geografia e Estatística (IBGE), 1997. XI Recenseamento Geral do Brasil: Manual de Delimitação de Setores. Rio e Janeiro.