PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Programa de Pós-Graduação em Informática

Métodos para Re-espacialização de Indicadores Socioeconômicos

Bruna Duarte Matias

Belo Horizonte

Abril de 2011

Bruna Duarte Matias

Métodos para Re-espacialização de Indicadores Socioeconômicos

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador: Profº. Drº. Clodoveu Augusto Davis Junior

Belo Horizonte
Abril de 2011

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Matias, Bruna Duarte
M433m Métodos para Re-espacialização de indicadores socioeconômicos / Bruna Duarte Matias. Belo Horizonte, 2011.
95 p.

Orientador: Clodoveu Augusto Davis Junior
Dissertação (Mestrado) – Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática

1. Indicadores sociais. 2. Indicadores econômicos. 3. Kernel, Funções de. 4. Sistemas de informação geográfica. 5. Processamento de dados. I. Davis Junior, Clodoveu Augusto. II. Pontifícia Universidade Católica de Minas Gerais. Pós-Graduação em Informática. III. Título.

CDU: 91:681.3

PUC Minas
Programa de Pós-graduação em Informática

ATA DE DEFESA DE DISSERTAÇÃO DO ALUNO
BRUNA DUARTE MATIAS

Realizou-se, no dia 15 de abril de dois mil e onze, às 09 horas e 30 minutos, na sala de Multimeios 30 - Bloco I da PUC Minas, Unidade São Gabriel, a 63ª defesa de dissertação do Programa de Pós-graduação em Informática, com título *"Métodos para Re-espacialização de Indicadores Socioeconômicos"* apresentada por Bruna Duarte Matias.

A banca examinadora foi composta pelos seguintes professores:

Prof. Clodoveu Augusto Davis Junior - Orientador (UFMG)
Profª. Karla Albuquerque de Vasconcelos Borges (PRODABEL)
Prof. Luis Enrique Zárate Gálvez (PUC Minas)

A banca examinadora considerou a dissertação:

(X) Aprovada (o candidato terá até dez dias para entregar o texto final da dissertação).
( ) Aprovada de forma condicional (o candidato terá até quarenta e cinco dias para entregar o texto final da dissertação).
( ) Reprovada.

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da comissão.

Belo Horizonte, 15 de abril de 2011.

Giovana Cássia da Silva – Secretária

Prof. Clodoveu Augusto Davis Júnior - Orientador (UFMG)
Doutor em Ciência da Computação, UFMG

Profª. Karla Albuquerque de Vasconcelos Borges (PRODABEL)
Doutora em Ciência da Computação, UFMG

Prof. Luis Enrique Zárate Gálvez (PUC Minas)
Doutor em Engenharia Metalúrgica e de Minas, UFMG

AGRADECIMENTOS

A Deus que sempre me amparou mesmo quando eu fraquejava em minha fé.

Aos meus pais, Dora e Sebastião, por todo amor e dedicação sempre me apoiando e me fazendo sentir que sou capaz de seguir em frente.

À minha irmã Bia e ao meu irmão João Emílio pelo carinho.

Ao Roberto, pela motivação e paciência.

Às minhas amigas Silvinha, Simy e Flavinha pela torcida.

Aos companheiros da Fundação João Pinheiro: Mônica Galupo, Maria Luiza Marques, Fátima Fortes, Priscilla de Souza, Vera Scarpelli, Olinto Nogueira, Rosânia e Fernando Prates, pelo incentivo e apoio.

Ao meu orientador, professor Clodoveu Davis, pelo ensinamento e profissionalismo admiráveis.

RESUMO

A geração de indicadores socioeconômicos é uma ferramenta importante para análises de dados. Indicadores são informação produzida especificamente para auxiliar nas tomadas de decisões, uma vez que sintetizam de modo consistente variáveis de diversas áreas de conhecimento, transformando-as em informações de rápida visualização e interpretação. No entanto, essas variáveis são coletadas e organizadas de diversas formas, com diferentes granularidades espaciais e temporais. Portanto, utilizar esses dados para a formação de indicadores complexos implica em desafios para a sua transformação e adaptação a cada situação de uso. Sendo assim, esta dissertação apresenta uma coletânea de metodologias, métodos e técnicas utilizadas para compatibilização de dados complexos, segundo as dimensões geográfica, temporal e temática. Apresentamos também um modelo conceitual como proposta para trabalhar com informações com granularidade espacial distinta, através da função *Kernel*. O modelo foi aplicado em um estudo de caso utilizando dados das mesorregiões, microrregiões e municípios de Minas Gerais.

Palavras-chave: Indicadores Socioeconômicos. Compatibilização. *Kernel*.

ABSTRACT

The generation of socio-economic indicators is an important tool for data analysis. Indicators constitute information that is specifically produced to support the decision-making process, since they are designed to consistently synthesize variables from various fields of knowledge. However, such variables are collected in several different ways, with varying spatial and temporal granularities. Therefore, in order to be able to use such data in the construction of complex indicators, it is necessary to face the challenge of transforming and adapting them to each situation of use. Therefore, this dissertation presents a collection of methodologies, methods and techniques that can be employed to combine complex data, considering their geographic, temporal and thematic dimensions. A conceptual model on how to work with variables of diverse spatial granularity using Kernel functions is also presented. The model has been verified through a case study that involves population data from Minas Gerais state's mesoregions, microregions and municipalities.

Key-Words: Socio-economic indicators. Kernel

LISTA DE FIGURAS

Figura 1: Busca pela informação enriquecida para auxiliar na tomada.....18
Figura 2: Processo de elaboração da base geográfica compatibilizada.....33
Figura 3: Processo de concepção de um *Data Warehouse*.....52
Figura 4: Uma visão geral das etapas que compõem o processo de KDD.....55
Figura 5: Comparação entre os métodos.....88

LISTA DE TABELAS

Tabela 1: População urbana setorial do IBGE e população urbana estimada através das duas fórmulas.....42
Tabela 2: Estrutura de composição de um índice 67

LISTA DE GRÁFICOS

Gráfico 1: Populações das Microrregiões Estimada *versus* Populações das Microrregiões IBGE – Minas Gerais, 2000..... 73
Gáfico 2: Populações das Microrregiões Estimada *versus* Populações das Microrregiões IBGE (sem população da Microrregião mais populosa) – Minas Gerais, 2000.....73
Gráfico 3: Populações dos Municípios Estimada *versus* Populações das Mesorregiões IBGE – Minas Gerais, 2000.....75
Gráfico 4: Populações dos Municípios Estimada *versus* Populações das Mesorregiões IBGE (sem população do município mais populoso) – Minas Gerais, 2000.....76
Gráfico 5: População Estimada com *Kernel* ponderado pelas populações das sedes municipais *versus* População dos municípios IBGE – Minas Gerais, 2000.....84
Gráfico 6: População Estimada com *Kernel* ponderado pelas populações das sedes municipais *versus* População dos municípios IBGE (sem município mais populoso) – Minas Gerais, 2000.....85
Gráfico 7: População Estimada com *Kernel* ponderado pelas populações dos distritos municipais *versus* População dos municípios.....86
Gráfico 8: População Estimada com *Kernel* ponderado pelas populações dos distritos municipais *versus* População dos municípios IBGE (sem município mais populoso) – Minas Gerais, 2000.....87

LISTA DE MAPAS

Mapa 1: Mesorregião dividida por Microrregião – Minas Gerais.....71
Mapa 2: Mesorregião dividida por Municípios – Minas Gerais.....74
Mapa 3: Estimador de *Kernel* ponderado pela população das sedes municipais.....83
Mapa4: Estimador de *Kernel* ponderado pela população dos distritos municipais 86

LISTA DE SIGLAS

DW - *Data Warehouses*

DWE - *Data Warehouse* Espacial

GOLAPA - *Geographic On-Line Analytical Processing Architecture*

SIG - Sistemas de Informação Geográfica

OLAP - *On-line Analytical Processing*

PNUD - Programa das Nações Unidas para o Desenvolvimento

ONU - Organização das Nações Unidas

IDH - Índice de Desenvolvimento Humano

IBGE - Instituto Brasileiro de Geografia e Estatística

IPEA - Instituto de Pesquisas Econômicas Aplicadas

IQVU - Índice de Qualidade de Vida Urbana

IVS - Índice de Vulnerabilidade Social

UP - Unidades de Planejamento

IMRS - Índice de Mineiro de Responsabilidade Social

KDD - *Knowledge Discovery in Databases*

DATASUS - Sistema Único de Saúde

INEP - Instituto de Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

GeoMDQL - *Geographic Multidimensional Query Language*

SUMÁRIO

1 INTRODUÇÃO 14
1.1 Motivação 16
1.2 Objetivos 16
1.2.1 Objetivo Geral 16
1.2.2 Objetivos Específicos 17
1.3 Justificativa 17
1.4 Estrutura da Dissertação 19

2 TRABALHOS RELACIONADOS 20
2.1 Estudos em bancos de dados geográficos 20
2.2 Sistemas de Informação Geográfica em ambientes de organização pública 22
2.3 Importância e Formulação de Indicadores Socioeconômicos 23
2.4 Construção de Indicadores e Índices utilizando dados geográficos 25

3 CONCEPÇÃO DE UM CONJUNTO DE MODELOS DE COMPATIBILIZAÇÃO DE DADOS COMPLEXOS 29
3.1 Seleção de Metodologias, Métodos e Técnicas 29
3.2 Formalização dos Principais Conceitos Aplicados 29
3.2.1 Nível 1: Agregação Espacial 30
3.2.2 Nível 2: Granularidade de Tempo 30
3.2.3 Nível 3: Temas 31
3.3 Conjunto de Modelos de Compatibilização de Dados Complexos 31
3.3.1 Compatibilização Segundo a Agregação Espacial 31
3.3.1.1 Compatibilização de Setores Censitários *31*
3.3.1.2 Sobreposição de Mapas *38*
3.3.2 Compatibilização Segundo a Granularidade de tempo 45
3.3.3 Compatibilização Segundo o Tema 46

4 DESCOBERTA DE CONHECIMENTO NO PROCESSO DE GERAÇÃO DE INDICADORES SOCIOECONÔMICOS COMPLEXOS USANDO BANCOS DE DADOS ESPACIAIS 49
4.1 Modelagem dos dados 50
4.2 Armazenamento dos dados 51
4.3 Descoberta de Conhecimento em Banco de Dados 53
4.4 Relação entre a Construção de Índices Complexos e o processo de KDD 56
4.4.1 Descoberta de Conhecimento em Banco de Dados Geográficos *57*
4.4.2 Seleção dos dados *57*
4.4.3 Pré-processamento *58*
4.4.4 Transformação *61*
4.4.5 Mineração de Dados *64*
4.4.6 Interpretação e análise dos resultados *67*

5 MODELO CONCEITUAL DE GERAÇÃO DE INDICADORES SOCIOECONÔMICOS COMPLEXOS USANDO DADOS ESPACIAIS 69
5.1 Seleção dos dados: Formulação do banco de dados geográficos 69
5.1.1 Modelagem de aplicações: Estudos de casos 70
5.2 Transformação: Estratégias para estimar informações que envolvem diferentes níveis de granularidade espacial 70
5.2.1 Estimação por Ponderação de Áreas 71
5.2.2 Variáveis Sintomáticas 77
***5.2.2.1* Considerações iniciais** *77*
5.2.2.2 Modelagem proposta *78*
5.2.3 Função Kernel 79
***5.2.3.1* Considerações iniciais** *79*
***5.2.3.2* Modelagem proposta** *80*
5.2.4 Aplicações das modelagens propostas: Estudos de casos 81

6 CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS 89

REFERÊNCIAS 91

1 INTRODUÇÃO

Os Sistemas de Informação Geográfica (SIG) são uma importante ferramenta para trabalhar com grandes volumes de dados espaciais, podendo ser aplicados a diversos campos da ciência, facilitando sua integração e a operacionalização de estudos e investigações científicas. As áreas e aplicações dos SIG são bastante variadas, como, por exemplo, planejamento urbano e gerenciamento de serviços públicos.

Diversas áreas de conhecimento envolvem um conjunto muito amplo de variáveis que são usadas para a produção de resultados e análises. Essas variáveis frequentemente são coletadas e organizadas de forma diferenciada a partir da necessidade, interesses ou limitações de cada área de estudo. No Brasil, o setor público é um importante fornecedor de bases de dados, nas quais uma diversidade de temas é apresentada. Isso é claramente percebido quando acessamos a página do Instituto Brasileiro de Geografia e Estática (IBGE) na Web e encontramos informações sobre economia, educação, atividades agrícola, industrial e mineral, renda, emprego, etc. Outro exemplo é o banco de dados do Sistema Único de Saúde, o DATASUS, que disponibiliza informações de estatísticas vitais, epidemiológicas e outras. Podem ser reconhecidos também outros esforços em coletar e organizar informações, tais como o Instituto de Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e o Instituto de Pesquisa Econômica Aplicada (IPEA).

A questão que surge é o que fazer quando precisamos construir indicadores a partir de informações que são provenientes de fontes distintas, observando que a granularidade espacial e temporal da coleta de dados varia entre as fontes. A capacidade de aplicação multidisciplinar da geotecnologia possibilita a integração e operacionalização dessas informações para a produção de resultados, porém as informações sobre uma mesma realidade geográfica freqüentemente são organizadas em bancos de dados com diferentes representações. Quando as respectivas informações possuem uma identificação georreferenciada, os dados referentes a uma mesma unidade geográfica podem ser compatibilizados, possibilitando a extração de informações que auxiliarão nas análises.

A diversidade espacial com que as informações são representadas varia conforme interesses ou restrições metodológicas no planejamento e construção dos bancos de dados. A complexidade envolvida nas diferentes representações espaciais entre esses bancos impõe desafios para a seleção e construção de indicadores relevantes a cada dimensão estudada, como, por exemplo, nas ciências sociais, ciências econômicas e ciências ambientais, assim como nos programas sociais desenvolvidos pelo setor público e/ou privado. Portanto, transformar esses dados em informações prontas para auxiliar nas tomadas de decisões implica em desafios para a geração de indicadores socioeconômicos complexos usando dados espaciais.

O problema levantado possibilitou a elaboração de uma proposta de trabalho em duas etapas. A primeira é referente a uma coletânea de metodologias, métodos e técnicas utilizadas para compatibilização de dados complexos, conforme sua complexidade, segundo as dimensões geográfica, temporal e temática. No entanto, muitas dessas abordagens levam em consideração somente uma das dimensões apresentadas. Portanto a importância desta proposta se deve ao fato de reunir mecanismos das três complexidades em um único ambiente, formando um conjunto de recursos para compatibilização de dados complexos, que possa auxiliar no desenvolvimento de apoio à tomada de decisão. A segunda etapa consiste na construção de um modelo conceitual como proposta para trabalhar com informações com distinta granularidade espacial.

1.1 Motivação

Em busca de propostas para integração de informações de bases de dados geográficos provenientes de fontes diversas em um ambiente único que possibilite a geração de indicadores socioeconômicos complexos usando dados espaciais, este trabalho apresenta a concepção de um conjunto de modelos, técnicas e metodologias de compatibilização de informações, assim como uma nova proposta através de um modelo conceitual que pretende contribuir para a construção de índices complexos, sendo uma base teórica que auxilie na solução de problemas associados às dificuldades com agregação e principalmente com a desagregação de dados com diferentes granularidades espaciais. As etapas envolvidas na construção de índices complexos podem ser consideradas similares às aplicadas no processo de descoberta de conhecimento em um *Data Warehouse (DW)*, visto a necessidade de compatibilizar as informações em um mesma realidade geográfica transformando-as em variáveis prontas para serem utilizadas nos respectivos cálculos.

1.2 Objetivos

1.2.1 Objetivo Geral

Desenvolver um mecanismo de apoio que auxilie nas soluções de problemas associados às dificuldades de agregação e/ou desagregação de dados com diferentes granularidades espaciais e temporais que dificultam a construção de indicadores socioeconômicos.

1.2.2 Objetivos Específicos

a) Avaliar métodos, técnicas e metodologias que possam ser aplicados na compatibilização de dados;
b) definir métodos e técnicas para a seleção de variáveis convencionais e geográficas;
c) conceber um conjunto de modelos com métodos, técnicas e metodologias que possibilitam a integração de dados espaciais em tempos distintos, dados temporais com periodicidades distintas e dados em que o mesmo tema apresenta metodologias distintas;
d) definir um modelo conceitual que possa ser aplicado na construção de indicadores socioeconômicos a partir de dados convencionais e espaciais, organizados em um *data warehouse*, visando deixar a informação pronta para auxiliar na tomada de decisões.

1.3 Justificativa

A construção de indicadores socioeconômicos complexos promove a informação pronta para tomada de decisões, possibilitando rápida avaliação da situação social nas diversas áreas focalizadas e identificando as características sociais e econômicas de cada região. Permite a aplicabilidade de análises mais completas em projetos em que existem fontes de dados variadas e a necessidade de fixar metas e avaliar o impacto das ações governamentais ou não. Quando há compatibilização espacial e/ou temporal entre os dados provenientes de fontes diversas, tem-se uma informação completa e enriquecida, porém, quando essa compatibilização não existe, tem-se uma informação incompleta e limitada, conforme figura 1.

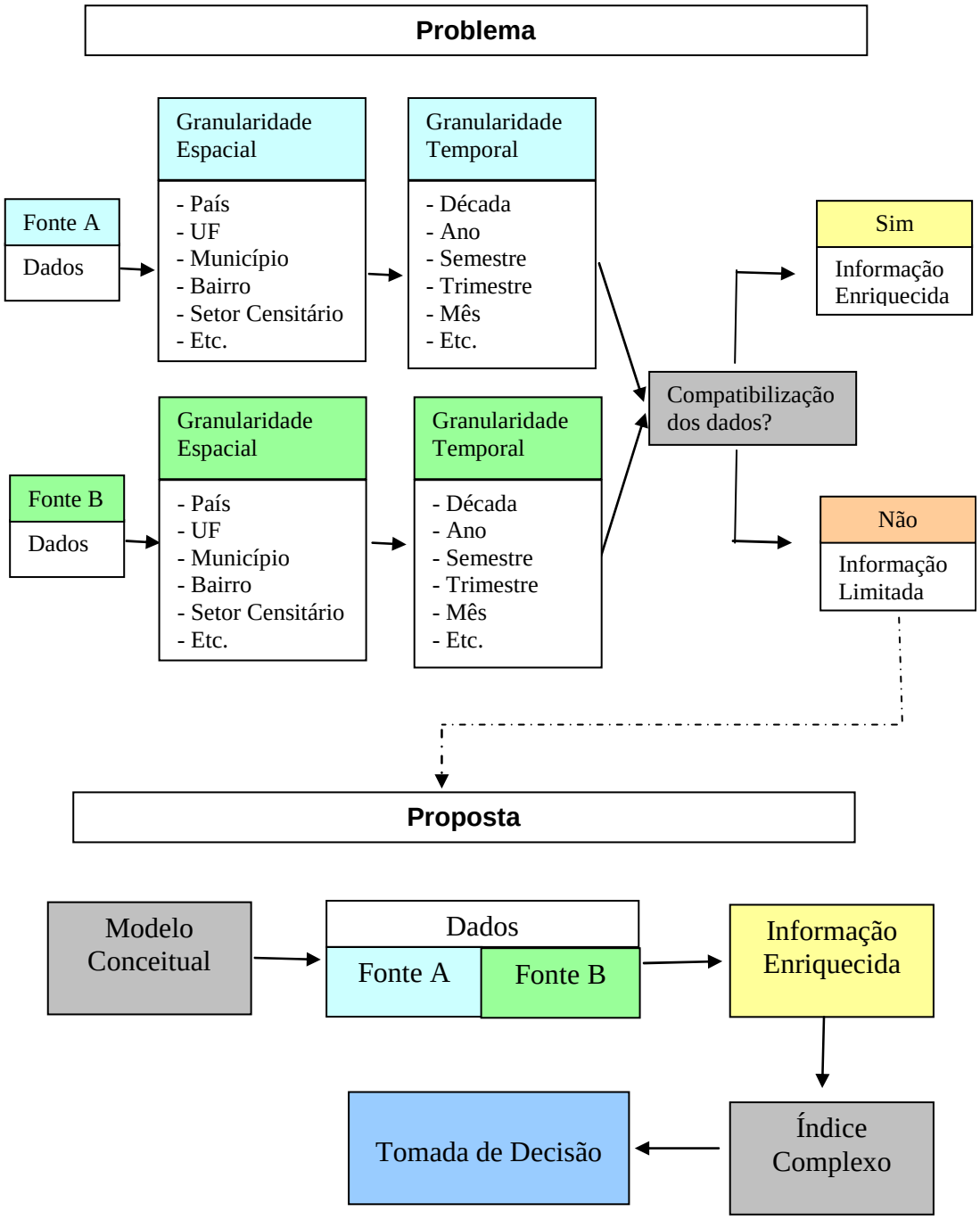


Figura 1: Busca pela informação enriquecida para auxiliar na tomada de decisão.

Fonte: Elaboração própria.

Verificamos também que o processo de construção de indicadores e índices se assemelha a várias etapas do processo de descoberta de conhecimento em *Data Warehouse (DW)*. Pois, dados de fontes distintas implicam muitas vezes em diferentes granularidades espaciais e temporais, portanto temos que realizar

tratamento adequado à informação visando compatibilizá-la transformando-a em informação consistente para a etapa final de cálculo do indicador complexo.

Assim, este trabalho propõe desenvolver recursos para o tratamento de dados organizados originalmente segundo um conjunto de unidades espaciais de referência, de modo a permitir sua espacialização segundo outro critério. Por exemplo, considere-se o problema de determinar a população residente em uma bacia hidrográfica, a partir de dados demográficos publicados por município, e sabendo-se que os contornos da bacia podem, por vezes, dividir municípios. A estimativa da população da bacia pode ser feita de diferentes formas, porém é razoável imaginar que as transformações mais simples possam levar a maiores erros, e que estimativas apoiadas por dados complementares e desenvolvidas com algoritmos adequados possam aproximar-se mais da situação real.

1.4 Estrutura da Dissertação

Esta dissertação está organizada da seguinte maneira. O Capítulo 2 apresenta trabalhos relacionados. O Capítulo 3 traz a conceituação dos modelos de compatibilização trabalhados na dissertação. O Capítulo 4 apresenta um modelo conceitual para compatibilização de dados geográficos. O capítulo 5 contem a conclusão e sugestões para trabalhos futuros e no capítulo 6 é apresentada a referência bibliográfica.

2 TRABALHOS RELACIONADOS

Neste capítulo, são discutidos trabalhos referentes aos estudos em bancos de dados geográficos, aplicações de Sistemas de Informação Geográfica em ambientes de organização pública, importância e formulação de indicadores socioeconômicos, construção de indicadores e índices utilizando dados geográficos e descoberta de conhecimento no processo de geração de indicadores socioeconômicos complexos.

2.1 Estudos em bancos de dados geográficos

Sistemas de Informação Geográfica (SIG), conforme Davis Junior (2000) são muitas vezes considerados sistemas de apoio à decisão, pois são capazes de coletar, modelar, gerenciar, manipular, analisar e integrar dados referenciados geograficamente, propiciando uma consistente análise espacial das informações, que pode ser aplicada em uma grande diversidade de áreas.

A disponibilização da informação usando seu componente espacial facilita o entendimento do comportamento da mesma através da visualização de sua localização. O uso do componente espacial dos dados permite determinar relacionamentos entre vários fenômenos geográficos, como, por exemplo, a distribuição espacial das condições socioeconômicas de uma determinada região, detectando as localidades mais carentes. No entanto, existe a necessidade de uma abordagem que integre as características multidimensionais com as espaciais de cada unidade geográfica estudada.

Data Warehouses (DW), ou armazém de dados, é um sistema de computação que armazena informações referentes às atividades de uma organização em bancos de dados voltados para a sumarização e análise, cuja estrutura fornece relatórios de informações que auxiliam na tomada de decisão. Conforme Sousa (2007) a integração entre DW (características multidimensionais) e SIG (características espaciais) gera um novo conceito, chamado de *Data Warehouse* Espacial (DWE) que cria um ambiente único de auxílio à tomada de decisão possibilitando o processamento eficiente de consultas considerando conjuntamente os dados

espaciais e convencionais, segundo diferentes critérios de agregação e em diferentes níveis de sumarização. Sendo assim, pode-se perceber que a construção de um DWE caracteriza-se como um fator importante no suporte à realização de consultas multidimensionais e geográficas. Fidalgo, Times e Souza (2001) propõem uma arquitetura denominada GOLAPA (*Geographic On-Line Analytical Processing Architecture*) para a integração de Sistemas de Informação Geográfica (SIG) a recursos especializados para realizar consultas em *data warehouse* (DW), denominados *On-line Analytical Processing* (OLAP). Essa proposta viabiliza a integração das tecnologias de DW, OLAP e SIG em um ambiente único, cujo contexto reúne um modelo de dado específico para suporte decisório (DW) à agilidade de análise sobre um grande volume de dados (OLAP) e a capacidade de análise e visualização espacial (SIG).

A possibilidade de aplicação em múltiplos campos da ciência faz do SIG uma importante ferramenta facilitando a operacionalização de estudos e investigações científicas envolvendo dados geográficos. Nesse caminho, Granero e Polidori (2002) apresentam elementos de um modelo de crescimento urbano que reúne saberes das ciências sociais (particularmente no campo da morfologia urbana), das ciências ambientais (no campo da ecologia de paisagem) e da ciência da computação (no campo da programação orientada a objetos), integradas num ambiente de SIG – sistema de informações geográficas.

A busca por um único ambiente com capacidade de processamento geográfico multidimensional para dar suporte ao processo de tomada de decisões tem uma importante preocupação, a consulta aos dados. Diante dessa necessidade, Fidalgo et al.(2004) propõem uma linguagem de consulta geográfica intitulada GeoMDQL (*Geographic Multidimensional Query Language*) que possibilita a utilização simultânea entre os operadores analítico-multidimensionais e geográficos. A linguagem proposta permite a consulta em um *data warehouse* geográfico, visto que o mesmo é composto por dados convencionais e geográficos.

A integração entre os Sistemas de Informação Geográfica (SIG) e a Análise Espacial é um assunto bastante discutido. Reis (2005a) descreve a importância da análise espacial, em especial da Análise Estatística de Dados Espaciais, cada vez mais difundida na comunidade de SIG, embora ainda não seja a mais significante comercialmente. Para Frozza e Mello (2006) a troca de informações entre sistemas distintos está entre as dificuldades enfrentadas pelos Sistemas de Informação

Geográficos, visto que uma mesma realidade geográfica frequentemente é representada por diferentes formas em diferentes SIG. Portanto, a troca de informações entre SIG torna-se importante, considerando os ganhos nas análises das informações com cruzamentos de dados de diversas fontes.

2.2 Sistemas de Informação Geográfica em ambientes de organização pública

Em Davis Junior (2002) é possível perceber a necessidade de debater os desafios tecnológicos e organizacionais existentes para os setores de administrações públicas que ambicionam utilizar Sistemas de Informação Geográficos (SIG). A acelerada transformação de novas possibilidades de aplicação dessa tecnologia reduz custos, ocasionando uma crescente popularização da informação. No entanto, existem ainda alguns insucessos, devidos à complexidade dos ambientes atuais, associada às deficiências na coleta da informação e na formação de pessoal especializado em muitas administrações públicas brasileiras. Nessa direção, Souza e Torres (2003) afirmam a importância do geoprocessamento para as políticas públicas, uma vez que permite que a informação seja relacionada com a sua localização no espaço de forma a identificar as características específicas de uma região para que seus problemas sejam detectados individualmente, ou seja, a necessidade da informação não ser limitada ao total municipal, mas disponível por diferentes unidades intra-urbanas permitindo, por exemplo, avaliar se os serviços públicos estão distribuídos seguindo alguma democratização de acesso.

Um trabalho realizado pela Secretaria Estado da Educação e a Universidade Federal do Paraná, Lobo e Amato (2003) propôs um sistema de apoio à tomada de decisões para o microplanejamento da rede estadual de ensino, os componentes desse sistema permitem realizar análises espaciais referentes à ampliação da rede existente compatibilizando informações de oferta de vagas escolares à demanda da população nas respectivas faixas etárias. Percebe-se claramente a restrição quanto à espacialização das informações trabalhadas, visto que, o trabalho limitou-se a utilizar as malhas digitais urbanas e rurais de setores censitários em função da agregação das informações socioeconômicas disponíveis no banco de dados do IBGE, que também estavam disponíveis por setor censitário. Torres (2005) também

descreve a importância da utilização do SIG para a implantação de sistemas de informação sócio-demográficos em sistemas públicos de informação no auxilio das análises necessárias para a tomada de decisões nas políticas públicas nas escalas regional e local.

O Brasil possui grande diversidade de fontes cadastrais nas quais as informações disponibilizadas possuem diferentes níveis de agregações e periodicidades, por exemplo, os dados censitários fornecem ricas informações, mas o problema está na periodicidade, visto que os dados são atualizados geralmente em intervalos de tempo maiores. No caso da utilização de amostras, tem-se a atualização das informações em intervalos menores, porém a desagregação espacial é limitada em função do tamanho e da representatividade da amostra nas diferentes áreas. Para as informações do registro civil, podem ocorrer tanto problemas na desagregação espacial quanto na ocorrência de sub-registro e/ou de sobrerregistro da informação. Assim, problemas de compatibilidade entre as informações provenientes de fontes diversas constituem um desafio para a integração dos respectivos dados em um ambiente único para consultas e análises.

Um trabalho realizado pela Universidade de São Paulo, Almeida, Quintanilha e Ho (2007) propôs a criação de um banco de dados geográfico compatibilizando variados formatos de dados espaciais cujas informações provinham de diversos órgãos públicos e privados que estavam representadas em diferentes formatos e extensões de *software*. Os dados trabalhados possuíam diferentes sistemas de projeção cartográfica com escalas e precisões diferenciadas. Assim, tem-se a indicação de que as informações provenientes de fontes diversas podem ser integradas em um único ambiente que possibilite a geração de indicadores socioeconômicos complexos usando dados espaciais. Portanto, este estudo bibliográfico segue em busca de trabalhos e propostas que possam dar continuidade à construção do estado da arte referente à geração de indicadores socioeconômicos complexos usando dados espaciais.

2.3 Importância e Formulação de Indicadores Socioeconômicos

A construção de indicadores requer processos metodológicos que permitem sua elaboração de forma confiável minimizando possíveis distorções que possam surgir na análise dos resultados, uma vez que, o principal objetivo dos indicadores é informar de forma concisa os fenômenos estudados.

Segundo Jannuzzi (2008) as atividades de planejamento do setor público são as principais definidoras do surgimento dos indicadores sociais. Esse processo de concepção ocorreu ao longo do século XX, tendo como marco conceitual os anos 1920 e 1930. Porém, somente em meados dos anos 1960 ganhou destaque científico, tendo em vista a crescente busca por sistemas de monitoramento das transformações sociais e dos impactos das políticas públicas que estavam sendo implantadas. Em meados da década de 1980, além de centros de pesquisa vinculados ao sistema de planejamento público, universidades e sindicatos passaram a investir em definições metodológicas para a construção de medidas que informassem de forma qualitativa e quantitativa as condições de vida da sociedade, dando origem aos indicadores sociais.

Os indicadores constituem informação pronta que auxilia nas tomadas de decisões, pois sintetizam de modo consistente os resultados encontrados. Por exemplo, com informações de nível de escolaridade e população por faixa etária ou população total de determinada região, é possível avaliar o grau de analfabetismo da população jovem e adulta, a quantidade de crianças em idade escolar e assim decidir quais investimentos serão necessários para atender a respectiva região de forma a melhorar as condições referentes à escolaridade. Portanto, o processo de sistematização de uma metodologia de análise e monitoramento de uma determinada realidade, deve se completar com a construção de indicadores que sintetizam os resultados encontrados, transformando-os em informações de rápida visualização e interpretação.

Esforços internacionais para a construção de indicadores sociais podem ser reconhecidos através de trabalhos como o realizado pelo Programa das Nações Unidas para o Desenvolvimento (PNUD) que é um órgão da Organização das Nações Unidas (ONU), que tem como objetivos promover o desenvolvimento humano e erradicar a pobreza. O Relatório do Desenvolvimento Humano do PNUD publica anualmente, desde 1990, o Índice de Desenvolvimento Humano (IDH), criado pelo professor Amartya Sen, ganhador do Prêmio Nobel de Economia em

1998, para medir o grau de desenvolvimento humano sustentável de uma sociedade.

No Brasil, pode-se verificar que grande parte dos indicadores sociais é formulada por órgãos de pesquisa vinculados ao setor público. A demanda por esses indicadores é também, na maioria das vezes, originada pelo setor público com o objetivo de analisar as condições de vida da população para a formulação de políticas públicas e avaliação de seus impactos.

Uma parceria entre o PNUD, a Fundação João Pinheiro, o Instituto Brasileiro de Geografia e Estatística (IBGE) e o Instituto de Pesquisas Econômicas Aplicadas (IPEA) lançou em 1998 o primeiro Atlas de Desenvolvimento Humano no Brasil, com o IDH de todos os municípios brasileiros, e em 2001 começou a produzir o Atlas de Desenvolvimento Humano para algumas regiões metropolitanas no país.

2.4 Construção de Indicadores e Índices utilizando dados geográficos

Dentre as diversas propostas para construção de indicadores tem-se o trabalho “Mapeando a exclusão social em Belo Horizonte” apresentado por Nahas (2000) demonstrando preocupação com o georreferenciamento das informações, uma vez que propõe uma análise realizada por índices, dados brutos e taxas referentes a cinco dimensões: ambiental, cultural, econômica, jurídica, e de segurança de sobrevivência. Esse estudo multidimensional se torna possível devido à compatibilização regional das diferentes dimensões trabalhadas, bem como a limitação de trabalhar somente com informações confiáveis, consistentes e com o mesmo georreferenciamento.

Entre vários outros trabalhos destinados à construção de indicadores sociais que possam auxiliar nas tomadas de decisões referentes ao planejamento público brasileiro, tem-se o Índice de Qualidade de Vida Urbana (IQVU) e o Índice de Vulnerabilidade Social (IVS). Conforme Nahas (2001) o IQVU e o IVS são compostos por indicadores georreferenciados em unidades espaciais intraurbanas (Unidades de Planejamento – UP), porém avaliam as condições de vida na cidade com abordagens distintas. O IQVU consiste em uma medida de acesso aos

serviços, ou seja, utiliza indicadores com informações sobre habitação que retratam o lugar em que vive a população. O IVS utiliza indicadores referentes a dados populacionais e domiciliares, tais como idade e escolaridade, entre outros, caracterizando o indivíduo. Quando se trabalha com informações georreferenciadas, os indicadores calculados com objetivos diferentes podem ser regionalmente compatibilizados e assim aumentar o ganho das informações. Como o IQVU e o IVS são georreferenciados em UP, pode-se compatibilizá-los e avaliar as características urbanísticas de uma mesma Unidade de Planejamento com as características dos indivíduos que a compõem. Por exemplo, avaliar se uma determinada UP que possui alto índice urbanístico irá apresentar baixa vulnerabilidade social, ou avaliar quais UPs com alto índice urbanístico apresentam alta vulnerabilidade social.

Outro trabalho científico que demonstra a preocupação nas análises espaciais envolvidas na granularidade da subdivisão espacial é apresentado por Dias et al. (2002) que discutem os problemas enfrentados na construção de indicadores sociais utilizando dados agregados por área e nos cuidados para sua interpretação, uma vez que podem ocorrer conclusões impróprias sobre o fenômeno em estudo. Como exemplo, o referido trabalho cita os efeitos associados a pequenas populações, uma vez que a estimação de taxas em áreas com pequenas populações sofre grande influência dos valores muito distintos.

Palheta da Silva et al.(2004) ressaltam a importância dos dados estarem georreferenciados para a definição das unidades espaciais para o Índice de Qualidade de Vida Urbana do município de Barcarena (PA), assim como para acompanhar as futuras mudanças urbanísticas. O banco de dados citado pelos autores foi gerado a partir de um conjunto de informações recolhidas por várias instituições, porém com o mesmo georreferenciados, fato que possibilita a integração de bases de dados distintas.

Em Nahas et al.(2006) encontramos a metodologia do Índice de Qualidade de Vida Urbana estendida para os municípios brasileiros, o IQVU-BR, que dimensiona a oferta de recursos e serviços urbanos, considerando na mensuração a possibilidade espacial de acesso da população a tal oferta. Segundo os autores, foram utilizadas diversas fontes de informações, órgãos dos governos estadual e federal, universidades, centros de pesquisas, entre outras, com o objetivo de identificar estatísticas georreferenciadas nos municípios brasileiros que permitissem a agregação espacial nesse nível possibilitando o cálculo dos indicadores. Outras

preocupações foram referentes aos anos em que as informações estavam disponíveis e a facilidade ou não em se acessá-las, seja em CD ou pela Internet. Os dados selecionados para o cálculo dos indicadores tinham que possuir características conceituais e territoriais que atendessem o objetivo do índice, porém quando alguma condição não era satisfeita, foi criado um indicador alternativo que fosse o mais próximo possível ao proposto inicialmente. Antes de concluir os cálculos do índice, procedimentos matemáticos para preparar as bases de dados tais como atribuição de valores dos indicadores para a compatibilização de bases de dados nos municípios criados após a pesquisa do Censo Demográfico, isso foi necessário uma vez que foram utilizados dados de diferentes fontes, anos-base e georreferenciamento distintos.

A Fundação João Pinheiro elaborou em 2005 o Índice Mineiro de Responsabilidade Social (IMRS) para os municípios do estado de Minas Gerais, cujos resultados estão organizados em uma base de dados juntamente com indicadores relacionados às dimensões saúde, educação, renda, segurança pública, habitação e meio ambiente, cultura e esporte, e finanças municipais. No final de 2009, foi lançando uma segunda versão que atualiza e amplia a respectiva base de dados contemplando os anos de 2000 a 2007. As fontes de dados utilizadas são basicamente de registros administrativos, que apresentam a vantagem de ter periodicidade curta, apesar de algumas deficiências, possibilitando a construção de séries anuais. Essa proposta organiza em uma mesma base, informações de diferentes instituições com diferentes formatos facilitando a sua disseminação tanto no ambiente público quanto para a sociedade. A base oferece informações para os anos de 2000 a 2007 sendo que alguns indicadores possuem dados para todos os anos. Porém, isso nem sempre ocorre para todos os indicadores devido à falta de atualização ou de compatibilidade geográfica para determinados dados em alguns anos. Os índices sintéticos referentes a cada dimensão estudada, assim como o índice final, estão disponíveis somente para os anos de 2000, 2002, 2004 e 2006. No processo de construção dos indicadores, foram arbitrados pesos e padrões de referência para os indicadores. Nos casos em que os indicadores não satisfaziam alguns pré-requisitos como abrangência temporal e geográfica, grau de aproximação entre o indicador e o conceito ou fenômeno medido, confiabilidade, grau de variabilidade no curto prazo, periodicidade adequada e entre outros, foi necessária a utilização de *proxies* para preencher as lacunas resultantes nas séries de dados.

Portanto, os autores decidiram por calcular os índices como médias de 2 ou 3 anos, por exemplo, o IMRS do ano de 2004 foi calculado pela média dos indicadores referentes aos anos de 2003, 2004 e 2005, sendo uma alteração em relação à metodologia aplicada na primeira versão do IMRS.

O IVS, IQVU-BH, IQVU-BR e o IMRS têm em comum o fato de serem índices que foram calculados priorizando fontes de dados que produzem estatísticas que são atualizadas em curto ou médio prazo. Assim, esses índices se tornam úteis para acompanhar a evolução das informações fornecidas, pois podem ser atualizados numa periodicidade mais próxima ao das modificações ocorridas nas características urbanas, sociais e econômicas da população em estudo. Podemos perceber também que ambos utilizaram dados vindos de diferentes fontes, o que levou a cada proposta fazer os ajustes mais adequados de forma a compatibilizar tais informações possibilitando a construção dos indicadores e índices

Diante o exposto, verifica-se claramente a importância dos dados e indicadores estarem georreferenciados para o ganho da informação, uma vez que a localização geográfica possibilita a construção de várias análises de uma mesma região, auxiliando na tomada de decisões.

Apresentamos a seguir uma seleção de metodologias, métodos e técnicas que tem por finalidade transformar dados de diferentes granularidades, compatibilizando suas informações com o intuito de contribuir para os estudos de construção de índices e indicadores cujas informações não possuem a mesma granularidade espacial e temporal.

3 CONCEPÇÃO DE UM CONJUNTO DE MODELOS DE COMPATIBILIZAÇÃO DE DADOS COMPLEXOS

Neste trabalho consideramos dados complexos, aqueles cujas informações estão disponíveis em diferentes formas: dados espaciais em tempos distintos, dados temporais com periodicidades distintas e dados em que o mesmo tema apresenta metodologias distintas. A construção da base teórica e conceitual para a seleção de modelos de compatibilização de dados complexos será detalhada a seguir.

3.1 Seleção de Metodologias, Métodos e Técnicas

Para a concepção, construção e expressão de um modelo que integre dados em diferentes níveis de granularidade: geográfico, temporal e temático; foram selecionadas metodologias, métodos e técnicas através dos seguintes critérios:

a) Abordagem científica;
b) Abordagem tecnológica;
c) Eficiência na definição dos elementos estudados na proposta;
d) Coerência conceitual durante a elaboração da proposta.

3.2 Formalização dos Principais Conceitos Aplicados

Para concepção de um conjunto de modelos de compatibilização de dados complexos, foram definidos os principais conceitos, tais como agregação espacial, granularidade de tempo e tema. Os mesmos definem o grau de complexidade dos dados, como apresentado a seguir.

3.2.1 Nível 1: Agregação Espacial

O nível de agregação espacial é um dos mais importantes aspectos considerados neste trabalho. Quando uma informação está disponível de forma georreferenciada, é possível analisá-la e compará-la com outras informações, podendo-se assim construir índices complexos que possam auxiliar na tomada de decisões.

De modo geral, os dados são agregados conforme o nível espacial de referência, porém em diversas situações podemos encontrar outros tipos de agregações. Por exemplo, as informações que uma empresa possui de seus clientes podem ser agregadas de forma a traçar um perfil de seus consumidores, no qual o cliente é a unidade de referência que será agregada. Quando uma universidade divulga a nota final dos candidatos ao vestibular, essa nota foi obtida através das notas das provas das diferentes matérias, sendo o aluno o nível de agregação utilizado.

Para compor o conjunto de modelos de compatibilização de dados complexos, quaisquer níveis de agregação espacial podem ser considerados. Em nossos testes e estudos de caso, os principais níveis de agregação espacial considerados foram: Unidade da Federação, Município, Região Metropolitana, Bairro, e Setor Censitário.

3.2.2 Nível 2: Granularidade de Tempo

Para acompanhar a evolução de determinada informação é necessário que as variáveis em estudo estejam em unidades de tempo compatíveis com o período analisado. Por exemplo, se queremos analisar a taxa de analfabetismo em determinada localidade nos últimos 5 anos, precisamos de variáveis referentes ao número de pessoas analfabetas e à população total, ambas com a mesma periodicidade. Assim, este trabalho priorizou as informações disponíveis nos seguintes tempos: decenal e anual. Existem também grandes pesquisas

relacionadas a aspectos econômicos como inflação, taxa de emprego e desemprego e outros, que estão disponíveis por semestre, trimestre e mês.

3.2.3 Nível 3: Temas

Cada pesquisa utiliza uma metodologia de acordo com seus objetivos, recursos financeiros, ou outras limitações. Por isso, um mesmo tema pode ser abordado de diferentes formas, como por exemplo, renda domiciliar e renda familiar. Os principais temas escolhidos estão relacionados a aspectos socioeconômicos e as variáveis foram trabalhadas conforme as dimensões a qual pertencem.

3.3 Conjunto de Modelos de Compatibilização de Dados Complexos

A seguir estão organizadas as metodologias, métodos e técnicas que possibilitam a integração de dados complexos conforme cada nível de complexidade: agregação espacial, granularidade de tempo e tema.

3.3.1 Compatibilização Segundo a Agregação Espacial

Uma unidade espacial pode possuir distintas representações geométricas e as análises das informações nelas contidas podem sofrer distorções resultantes dos diferentes tipos de agregação. A seguir serão apresentados metodologias, métodos e técnicas existentes que compatibilizam os dados conforme a agregação espacial da localidade em estudo.

3.3.1.1 Compatibilização de Setores Censitários

Os limites das unidades espaciais são definidos conforme a necessidade do estudo e interesses ou restrições metodológicas. A busca por uma área mínima que represente de forma homogênea e confidencial as informações relativas aos indivíduos que a compõem faz com que os setores censitários sejam trabalhados com grande freqüência. Os setores censitários, segundo o Instituto Brasileiro de Geografia e Estatística (2000), constituem uma unidade territorial urbana ou rural onde apenas um recenseador coleta as informações.

Por ser uma unidade espacial definida com o propósito de delimitar o espaço de trabalho de um recenseador, o traçado dos setores censitários nem sempre obedece à lógica urbana. Ou seja, não se consegue garantir a homogeneidade de um setor censitário segundo nenhum critério. Além disso, existe a possibilidade de divisão irregular, entre setores censitários, dos vazios urbanos. Por exemplo, a área de uma praça pode ser incorporada a qualquer um dos setores censitários que a circundam, em vez de ser dividida entre eles, ocasionando distorções quando a área do setor é usada para algum tipo de ponderação, conforme Oliveira et al (1996).

Análises temporais envolvendo os setores censitários ou outras unidades espaciais geram problemas de natureza espaço-temporal, pois a cada período de tempo a unidade espacial em estudo pode sofrer modificações quanto à abrangência de seus limites. Isso pode acarretar em distorções nas análises das informações trabalhadas, pois as variações encontradas podem ser referentes à mudança de limites da unidade geográfica e não à mudança de comportamento dos dados. Portanto, para a realização de análises temporais é necessário que as áreas das unidades espaciais permaneçam as mesmas ao longo do período estudado.

A compatibilização entre os dados censitários ameniza possíveis distorções geradas sobre dados agregados em áreas de geometrias distintas. Na literatura existem várias propostas para compatibilização de setores censitários em tempos distintos. A seguir serão apresentados alguns procedimentos, ressaltando suas contribuições, diferenças e limitações.

A proposta de Câmara et al.(2005) oferece um procedimento que compatibiliza a geometria dos setores censitários e de seus respectivos dados para a realização de análises temporais com o auxílio de imagens *Landsat*. O procedimento propõe a agregação de áreas e a desagregação de alguns setores. Imagens de satélite são utilizadas na identificação e quantificação das áreas

ocupadas por usos urbanos, possibilitando a integração de dados populacionais a dados do meio físico. Foram compatibilizados setores urbanos de São José dos Campos entre os anos de 1991 e 2000, conforme mostra a figura 2, onde os setores resultantes da base de 1991 estão destacados em vermelho e, os setores resultantes da base de 2000 estão em amarelo.

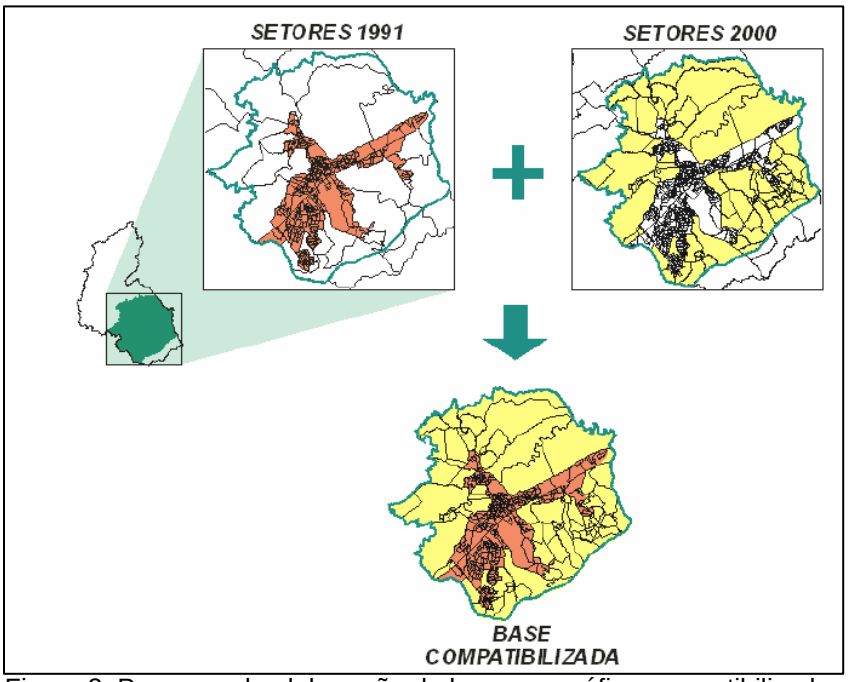


Figura 2: Processo de elaboração da base geográfica compatibilizada
Fonte: Câmara et al (2005).

Para a elaboração da base geográfica compatibilizada, primeiramente foram registradas as imagens *Landsat* dos anos de 1990 e 2000 utilizando o mosaico digital do ano 2000 (os autores utilizaram uma imagem do ano de 1990, ao invés de 1991, pois no ano do recenseamento ocorreu a presença de nuvens nas imagens). As áreas onde existia indicação de ocupação urbana foram extraídas das imagens após uma classificação realizada com o apoio de mapas digitais do sistema viário e das quadras sobrepostas às imagens. Segundo os autores, este procedimento permitiu uma delimitação mais precisa das áreas urbanas.

A sobreposição das bases cartográficas gerou alguns polígonos resultantes da interseção de linhas que representam as mesmas formas, porém com representações digitais diferentes. Para solucionar esse problema, foi realizada uma análise sobre os dados originais comparando-os com mapa do sistema viário e escolheu-se a geometria dos setores de 2000, por ser considerada mais confiável, como base para a digitalização.

Os autores utilizaram tabelas de comparabilidade dos setores 1991-1996 e 1996-2000 e das áreas urbanas (delimitadas sobre as imagens *Landsat*) existentes nos anos de 1990 e 2000, foi elaborada uma tabela de equivalência entre os polígonos da base compatibilizada e os setores censitários de 1991 e 2000. Como resultado final obteve-se tabelas contendo a base compatibilizada associada aos dados censitários de 1991 e mesma base compatibilizada, porém associada aos dados censitários de 2000. Os autores, reconhecendo as limitações da proposta, pois a desagregação leva em consideração somente a área ocupada sem se preocupar com a existência de densidades diferenciadas, propõem como alternativa para amenizar tais limitações a compatibilização das bases censitárias utilizando dados de outras fontes, como mapas de uso do solo (identificação de áreas residenciais) ou cadastro de imóveis (identificação de domicílios).

Em Umbelino e Barbieri (2008) encontra-se uma proposta mais simplificada, porém eficaz. Os autores descrevem detalhadamente os procedimentos operacionais envolvidos no trabalho, segue uma breve apresentação desses passos.

Através da associação dos *softwares* MapInfo e Excel, foi apresentada uma metodologia semi-automatizada para a compatibilização de setores censitários permitindo a criação de áreas iguais possibilitando assim, análises temporais das regiões em estudo. A metodologia foi aplicada utilizando os setores censitários dos censos demográficos de 1991 e 2000, e como unidade territorial a bacia hidrográfica do Córrego do Onça, situada entre os municípios de Belo Horizonte e Contagem. Os polígonos dos setores censitários de 1991 e 2000 tiveram seus respectivos códigos armazenados no *software* MapInfo, e para cada ano foi criada uma chave do processo de unificação. Essa chave foi obtida através do cálculo de área pelo método esférico, pois os autores acreditam que esse método fornece uma medida mais precisa do terreno estudado, já que leva em consideração a curvatura da superfície do planeta. Assim, cada ano teve sua chave determinada pela área em m^2 de cada polígono. O banco de dados resultante, em ambos os anos, foi composto por uma coluna com códigos do setor e outra com a área em m^2. Para evitar problemas com a existência de valores iguais de área para os setores censitários em cada período, trabalharam-se os valores com 12 casas decimais. Para a construção das áreas mínimas de agregação, foram apresentadas quatro possibilidades de configurações realizadas nos setores censitários:

a) Compatibilização quando o setor censitário não sofreu alteração de 1991 para 2000.

Nesse caso o objetivo é encontrar todos os setores que em ambos os anos apresentam mesma área. Isso foi feito no *software* Excel, utilizando um comando de comparação para testar a igualdade ou não das áreas dos polígonos. Os polígonos que satisfizeram a condição de igualdade em suas áreas foram testados no MapInfo, para verificar se houve uma perfeita sobreposição dos mesmos. Os setores que não foram selecionados nessa primeira etapa serão reavaliados nas etapas seguintes até que todos sejam compatibilizados.

b) Compatibilização quando dois ou mais setores em 2000 precisam ser agregados para chegar a uma área comum em 1991.

A preocupação agora é em agregar dois ou mais setores em 2000 para reproduzir a mesma área em 1991. Para isso, foi necessária a separação de um polígono em 1991 em função de outros que o sobrepõem em 2000 utilizando o MapInfo. Para a verificação da igualdade de valores e sobreposição de área foi aplicado o mesmo procedimento utilizado na primeira etapa da compatibilização.

c) Compatibilização quando dois ou mais setores em 1991 precisam ser agregados para chegar a uma área comum em 2000.

Essa compatibilização segue as mesmas premissas da segunda etapa, porém agora um polígono de 2000 será dividido em função de outros polígonos que o sobrepõem em 1991.

d) Compatibilização quando os limites dos setores nos dois anos não apresentaram semelhanças implicando na agregação de dois ou mais setores em 1991 e em 2000 para determinar uma área comum.

Utilizando os setores que ainda não foram compatibilizados nas etapas anteriores, pois as áreas nos dois períodos não possuíam qualquer semelhança, foi realizada uma análise visual no MapInfo através da sobreposição dos polígonos em

1991 e 2000 para determinar quais setores seriam agregados manualmente, determinando uma área comum.

Em cada uma das 4 etapas apontadas anteriormente, os polígonos que foram satisfazendo os respectivos critérios de compatibilização tiveram seus setores organizados em um novo arquivo formando uma base compatibilizada.

Quando os setores censitários estão em uma bacia hidrográfica, podem ocorrer problemas metodológicos, pois seus limites são definidos com base em curvas de nível segundo critérios técnicos e não por limites políticos das localidades. Portanto, a alternativa sugerida pelos autores considera que os setores censitários sejam homogêneos e que somente a porcentagem do setor inserida dentro da bacia seja avaliada. Assim, são analisados somente os dados dos habitantes ou domicílios correspondentes ao respectivo percentual.

Apesar das limitações, a proposta apresentada contribui para os avanços metodológicos em busca de uma completa compatibilização das unidades espaciais em tempos distintos. Ao trabalhar com a agregação e desagregação de unidades espaciais com geometrias distintas, o método possibilitou a compatibilização de setores censitários, o que nos leva a áreas mais homogêneas e análises menos distorcidas.

Confirmando a importância dos estudos em compatibilização de informações geográficas, em 2009 foi apresentado por Lobô (2009), um método para compatibilizar dados estatísticos de setores censitários urbanos de 1991 e 2000 da Região Metropolitana de Belém (RMB) utilizando técnicas de geoprocessamento e o programa Terraview, desenvolvido pelo Instituto Nacional de Pesquisas Espaciais (INPE).

Devida a possibilidade de alterações dos limites, uma região terá em anos distintos malhas geográficas também distintas. Assim, torna-se necessário desenvolver um método que compatibilize essas malhas para que suas respectivas informações possam ser comparadas e analisadas corretamente. O objetivo é transferir as estimativas dos dados de população dos setores censitários de 1991 para os de 2000 para que a malha de setores de 2000 tenha a estimativa de população de 1991 já a malha de 2000 permanecerá com seus dados originais. Isso possibilitará que as duas malhas sejam comparadas sem que ocorram distorções nas análises.

Para a concepção do método, o autor inseriu no programa Terraview os mapas vetoriais dos setores censitários de 1991 e 2000 juntamente com os respectivos dados estatísticos, formando assim um sistema de informações georreferenciadas. As bases trabalhadas foram disponibilizadas pelo IBGE. Como o interesse da proposta era trabalhar somente a área urbana da RMB, foi aplicado um ajuste na população desses setores, pois as informações contidas nas áreas não urbanizadas seriam descartadas. Assim, foi considerado que 90% dos habitantes residiam na área urbanizada do setor e 10% na área não urbanizada. Conseqüentemente obteve-se as estimativas dos totais de população da área urbanizada para os anos de 1991 e 2000. Ainda no Terraview, foi possível extrair a área dos setores e a densidade populacional bruta de cada setor censitário para o ano de 1991. Esse cálculo baseia-se na razão entre a população e a área do setor. O passo seguinte foi gerar uma malha de pontos ortogonais, distantes 100 m entre si. A cada ponto gerado, foi atribuída a densidade demográfica do setor censitário onde estava situado em 1991 e seus respectivos valores foram aplicados a cada setor censitário correspondente em 2000. Nos casos em que os setores continham um único ponto, seu valor foi igual ao do ponto. Porém, quando o setor apresentava dois ou mais pontos, o próprio Terraview calculou a média dos valores dos pontos respectivos. Alguns setores censitários de 2000 ficaram sem atribuição de dados e foram desconsiderados. Porém, as conclusões não foram prejudicadas, pois isso ocorreu em poucos casos em relação ao total de setores.

Para estimar o número de habitantes de cada setor, a densidade de 1991 foi multiplicada pela área do setor censitário de 2000. Assim, gerou-se a malha de setores censitários de 2000 contendo dados referentes ao ano de 1991. O autor ressalta que uma desvantagem do procedimento é que, ao se atribuir valores de 1991 à malha de setores de 2000, os valores da malha resultante não são mais originais, e sim estimados. Outra fragilidade da proposta está relacionada ao fato de se trabalhar com a densidade média e, portanto nos casos em que os habitantes não estão distribuídos homogeneamente pelo seu território, será acrescentada uma distorção nas respectivas informações.

3.3.1.2 Sobreposição de Mapas

A sobreposição de mapas é uma técnica muito utilizada para compatibilizar unidades geográficas às respectivas informações. Uma proposta interessante é a de Oliveira et al.(1996), onde foi apresentado uma metodologia que adequou as Unidades de Planejamento (UPs) propostas pela Secretaria Municipal de Planejamento de Belo Horizonte à malha de setores censitários do IBGE. Esse procedimento avaliou a malha viária e as edificações do espaço urbano produzindo um conjunto de unidades de planejamento, onde se preocupou em minimizar o número de possíveis conflitos com a malha de setores utilizando critérios de homogeneidade estrutural e demográfica. Assim, foram definidas unidades de planejamento com maior qualidade em suas respectivas informações demográficas.

Dentre as unidades geográficas avaliadas, primeiramente foram rejeitadas as que apresentaram níveis elevados ou muito baixos de agregação de dados e as em que os limites seguem critérios que não consideram as grandes diferenças demográficas, sociais, urbanas e econômicas dentro de uma mesma região. Foram selecionadas para o estudo unidades geográficas que não apresentassem informações muito heterogêneas e que suas áreas não fossem muito pequenas ou muito grandes, ocasionando distorções na leitura dos dados. Foi justificado que a escolha das UPs se deve ao fato de terem sido criadas considerando as unidades físico-territoriais de referência para a compatibilização das diferentes bases de dados existentes, podendo ser subdivididas e adequadas conforme a necessidade de demandas futuras.

Para realizar uma compatibilização entre várias camadas de informação, foi utilizado um procedimento ao nível de geoprocessamento entre dois ou mais *layers* denominada *matching*. Conforme os autores, as linhas que formam o contorno dos setores censitários devem ocupar o centro do arruamento, tornando-as compatíveis com as linhas que formam o contorno das respectivas UPs. Para avaliar se houve compatibilização das camadas de dados, foram empregados recursos de processamento de polígonos disponíveis no sistema de geoprocessamento da PRODABEL, constatando que alguns setores censitários estavam contidos em mais de uma UP, o que sugeria uma ausência de compatibilização entre os *layers* dos setores censitários e UP. As UPs foram digitalizadas sem considerar o contorno dos

setores censitários contribuindo assim, para que algumas camadas de dados não fossem compatíveis.

Em virtude dos problemas de compatibilidade, os autores realizaram uma revisão cartográfica das UPs com o objetivo de compatibilizá-las aos setores censitários. Esta proposta proporciona uma alternativa para compatibilização das informações que minimiza os problemas oriundos de dados onde a população se distribui de forma heterogênea no espaço urbano. O que não ocorre com o recurso da interpolação simples dos dados, uma vez que o mesmo considera que os dados populacionais se distribuem de forma homogênea no espaço urbano. Para a construção da nova base cartográfica, conforme os autores foi necessário ajustar a digitalização das UPs que possuíam setores censitários nos quais mais de 95% de sua área total estavam em uma única UP. Com a utilização do *matching* modelou-se os contorno das UPs seguindo os limites dos setores censitários. No segundo passo foram redigitalizadas as UPs que terminavam em praças ou vazios populacionais, seus limites foram adaptados acompanhando os contornos dos setores censitários em que a praça ou vazio urbano se encontravam por completo. Nos casos em que os setores censitários apareciam em duas UPs, mas não estavam divididos por estrutura urbana expressiva e não envolviam áreas de favela, foram igualmente adaptados em uma única UP. Nos grandes aglomerados de favelas da cidade, foi necessário separá-los em UPs adequadas, pois quando estas áreas eram avaliadas em conjunto com os bairros vizinhos, incidiram distorções em suas características.

Nos casos em que mais de uma UP continha o mesmo setor censitário, foi aplicada uma interpolação contingenciada dos respectivos setores com o apoio do banco de dados do IPTU de 1º de janeiro de 1994 da Secretaria Municipal da Fazenda. Foi considerado que a população se distribui de forma homogênea em relação aos endereços residenciais do cadastro imobiliário urbano. Apesar de algumas UPs resultantes da revisão cartográfica ainda possuir algum grau de heterogeneidade, os autores ressaltam que essa diferença é menor do que se tomarmos, por exemplo, uma região administrativa cujos limites seguem critérios administrativos ou políticos, ocasionando, em muitos casos, um maior nível de agregação de dados.

Alguns trabalhos apresentam metodologias que dividem por igual as informações dos setores, porém esse procedimento possui fragilidade nos casos em que na realidade a população do setor original não estiver distribuída de forma

homogênea. Por exemplo, nos casos em que a população estiver concentrada em um ponto do setor, ao dividi-lo por igual, as informações também ficaram dividas por igual fazendo com que ajam distorções em sua análise, uma vez que não retratará a realidade do local.

Problemas de estimativas confiáveis de dados inter-censitários são discutidos em diversos trabalhos, como por exemplo, o proposto por Souza (2004), que apresenta uma metodologia para estimar a população intra-urbana utilizando dados de sensoriamento remoto e informações do espaço residencial urbano construído. O estudo integrou no ambiente SPRING imagem de Satélite e dados do IBGE para o ano 2000 dos setores censitários da malha urbana do município de São José dos Campos determinando quais áreas apresentavam características de ocupação residencial semelhantes. Para a construção da nova base digital, foi gerado um mosaico cujas imagens tiveram a geometria corrigida através do método de correção polinomial simples, sendo que para coletar os pontos de controle foram utilizadas as ortofotos digitais obtidas em 2000.

Através da interpretação visual de fotografias aéreas, foram mapeados os diferentes usos do solo urbano e as texturas relativas ao uso residencial foram separadas das texturas referentes aos demais usos: comercial, industrial áreas de lazer, rede viária, áreas e áreas com vegetação. Para determinar os setores residenciais de textura homogênea presentes na classe de uso residencial, foram analisadas as seguintes variáveis físicas: Tamanho do lote (forte indicador de renda familiar); Ocupação do lote: Organização da ocupação do lotes e das quadras; Arborização das ruas e lotes; Traçado e tratamento do sistema viário; Densidade de residências. Com o uso das imagens foi possível analisar qualitativamente a densidade de ocupação do setor homogêneo, utilizando como critério o tamanho das habitações e o padrão de ocupação dos terrenos. A análise da homogeneidade dos telhados forneceu a identificação de conjuntos habitacionais que geralmente são relacionados à população de baixa renda, enquanto o acabamento das construções forneceu um padrão construtivo das residências. Assim, os setores residenciais que apresentaram texturas homogêneas, foram integrados à base geográfica digital do IBGE contendo os limites dos setores censitários no ano 2000 da região trabalhada. O processo de integração foi realizado por procedimentos básicos de ajuste de dados vetoriais.

O autor definiu a variável “numero médio de habitantes por domicilio” utilizando os dados populacionais censitários do IBGE para ano 2000 que foi integrado à base digital geográfica. Através da delimitação de polígonos dentro de cada setor homogêneo, foi possível encontrar o número de habitantes por hectare que define a “densidade habitacional” que é responsável por determinar o “numero de unidades habitacionais dos setores”. Esse procedimento seguiu os alguns critérios. Para facilitar a operacionalização do processo, foi trabalhado um conjunto de amostras representando aproximadamente 3% da área de cada um dos setores homogêneos amostrais; A partir da análise visual utilizando as opções de zoom e escala, foram verificadas a delimitação e a distribuição dos polígonos, o que possibilitou determinar as unidades residenciais dentro de cada polígono.

A taxa de ocupação do setor homogêneo e a estimativa populacional foram determinadas através de equações (descritas detalhadamente no artigo), utilizando informações de domicílios ocupados fornecidos pela Unidade do IBGE de São José dos Campos. Porém, em alguns casos não se teve acesso a todas as informações necessárias para a realização dos cálculos. Para solucionar esse problema, a autora sugere que algumas dessas informações possam ser levantadas a partir de estimativas baseadas em amostras e visitas ao campo nas áreas que contêm os polígonos de interesse. Para realizar as estimativas populacionais, as áreas de mesma textura homogênea e os setores censitários do IBGE para o ano de 2000 foram compatibilizados e incorporados à base geográfica da área de estudo. No SPRING, foram realizados ajustes das linhas dos setores através dos procedimentos de edição vetorial disponíveis. Para que um mesmo setor do IBGE não fosse usado como amostra de duas texturas homogêneas distintas, foi verificado se os respectivos setores estavam contidos completamente e dentro de uma única textura.

Para analisar a variância da variável “número médio de habitantes por domicílio” e verificar se as áreas selecionadas poderiam ser avaliadas como homogêneas em relação a essa variável, foi aplicado um procedimento estatístico conhecido como teste de hipóteses. Após a aplicação do teste estatístico ficou comprovado que a homogeneidade da textura e a homogeneidade do número médio de habitantes por domicílio estão relacionadas. Para verificar a “densidade habitacional” dos setores selecionados, foram criados alguns polígonos dentro desses setores e para o levantamento das unidades residenciais dentro desses polígonos foi realizada uma interpretação na tela do computador identificando

através da resposta espectral dos diferentes materiais, os telhados das edificações. Através da multiplicação da densidade obtida para os polígonos amostrais pela área dos setores equivalentes, foi estimado o número de habitações nos respectivos setores.

A tabela 1 apresenta a estimativa da população dos setores amostrais seguindo dois métodos: primeiro multiplicou-se o número total de habitações do setor pelo número médio de habitantes por domicílio, denominada de população (1); no segundo método, multiplicou-se o número total de habitações do setor pelo número médio de habitantes por domicílio e pela a taxa de ocupação do setor denominada de população (2). Comparando os resultados das estimativas populacionais dos dois métodos com os dados oficiais do IBGE pode ser verificado que, para os setores das texturas homogêneas (3 e 9) selecionadas, houve um pequeno aumento no número de unidades residenciais no setor da textura homogênea 3 e uma pequena redução no setor 9. Essas diferenças podem estar relacionadas às particularidades das características de cada setor escolhido, porém podemos verificar que os resultados estão próximos com os dados obtidos no levantamento realizado pelo IBGE. Verifica-se também que as estimativas obtidas através do segundo método geraram resultados mais próximos com os publicados pelo IBGE, pois esse procedimento considerou a taxa de ocupação refinando a estimativa.

TABELA 1: População urbana setorial do IBGE e população urbana estimada através das duas

SETORES	Número de unidades residências IBGE	Número de unidades habitacionais Estimadas por fotointerpretação (D_h)	IBGE população	ESTIMATIVA População(1) $P(sth) = D_h . I.$	ESTIMATIVA População (2) $P(sth) = D_h . I. \ t_h$
Textura homogênea 3	11255	11797	39708	48132	39949
Textura homogênea 9	4277	3878	13049	14349	11622
Total	15532	15675	52757	62481	51571

Fonte: Souza, 2004.

Essa proposta além de possibilitar a estimativa populacional intercensitária, mostrou que ao trabalhar com setores homogêneos foi possível agregar informações e realizar estimativas sem que ocorram grandes distorções nas análises. Demonstrou também que com aproximadamente 3% da área total dos setores amostrais foi possível realizar uma estimativa populacional com resultados válidos em torno de 90%, quando comparados com os dados oficiais do IBGE para o mesmo período

Outra proposta utilizando imagens de satélite e fotografias aéreas como alternativa para auxiliar nas estimativas de dados populacionais pode ser encontrada em Reis (2005b), que, através das técnicas propostas por J. T Harvey, estima as populações para os setores censitários de Belo Horizonte em 1996, com exceção dos setores especiais - asilos, orfanatos e presídios - e os que continham menos de 1 habitante por 10 m^2, utilizando imagens de satélite e a contagem populacional também no ano de 1996.

Com o auxílio dos *softwares* SPRING 4.0 e R 1.8.1 (R *Development Core Team*), os bancos de dados foram construídos obtendo-se a média da reflectância em cada banda para todos os setores censitários. Foi aplicado o método da máxima verossimilhança para classificar quais eram os pixels urbanos e cada um deles foi associado a um dos setores censitários urbanos.

Para a modelagem do banco de dados de setores e de pixels, foi utilizado o modelo de regressão:

$$p_i = \beta_o + \sum_{j=1}^{k} \beta_j r_{ij} + \varepsilon_i$$

Onde segundo a autora, J. T Harvey propõe duas abordagens: quando a abordagem é feita pelos setores censitários, p_i representa a densidade populacional do setor censitário *i*, r_{ij} é a média da reflectância dos *pixels* do setor *i* na *j*-ésima banda do sensor, β_0 *e* β_j com *j* = 1, 2, ... *k são* os parâmetros a serem estimados e ε_i representa a parte da densidade populacional dos setores que não é explicada pelo modelo de regressão; nos casos em que a na abordagem é realizada pelos *pixels*, r_{ij} é a reflectância do *pixel i* na *j*-ésima banda do sensor e p_i representa a população do pixel. Porém, as contagens populacionais se referem aos setores censitários. Portanto, a modelagem dos dados ao nível dos *pixels* apresenta uma

dificuldade. Para resolver este problema utilizou-se um procedimento que foi denominado como regressão iterada. Esse procedimento consistiu em, após classificar os *pixels* como residenciais e não-residenciais, foi feita uma estimativa inicial da população em cada *pixel* que posteriormente foram refinadas iterativamente. A estimativa inicial de população em cada *pixel*, *pi*, é obtida através da divisão da população do setor pelo número de *pixels* do setor.

Após a equação de regressão ser determinada e os valores de p_i estimados e ajustados, o total da população do setor censitário depois do ajuste foi considerado como o total populacional conhecido do setor. A população ajustada do *pixel i* foi determinada pela soma da população estimada para o *pixel i* e da média dos resíduos do setor censitário do respectivo *pixel*. As estimativas iniciais de população na variável dependente são substituídas pelos os $p_{i's}$ ajustados na iteração anterior e a equação de regressão é novamente estimada. Assim foram ajustados novos $p_{i's}$ que substituem os $p_{i's}$ na próxima iteração. As iterações prosseguem até que um critério de parada seja definido. Esse critério pode ser com base no coeficiente de determinação (R^2) ou no quadrado médio dos resíduos ou em alguma outra medida de qualidade de ajuste. Quando o valor da medida escolhida como critério de parada não apresentar variações significativas de uma iteração para outra, podem-se finalizar as iterações. A proposta chama a atenção para possíveis problemas de multicolinearidade entre as reflectâncias nas bandas que podem afetar a convergência do processo de regressão iterada e problemas com estimativas negativas para as populações associadas aos *pixels*, pois o modelo de regressão linear utilizado não possui restrições. Assim, foi sugerido que a cada iteração, as estimativas negativas sejam transformadas em zero e ajustes nas estimativas dos outros pixels sejam feitos de modo a manter constante o total populacional da região trabalhada.

Para avaliar os resultados, a autora analisou as medidas de erro onde o erro relativo para cada setor foi calculado pela diferença absoluta entre o valor observado da população e o valor estimado pelo modelo, dividida pelo valor observado e para encontrar o erro relativo total foi realizada a diferença entre o total populacional estimado, considerando todos os setores, e o total populacional observado, dividida pelo total observado.

O banco de dados de setores foi construído com as informações geradas pela regressão do modelo mais indicado, foram feitas várias tentativas até a escolha do

modelo mais adequado e seus resultados foram analisados e validados através de técnicas estatísticas apropriadas.

No caso do banco de dados de *pixels*, foi realizada uma amostragem onde foram selecionados 25% de cada setor. Segundo a autora, a escolha desta fração amostral foi arbitrária, contudo a vantagem em se trabalhar com a amostragem dos pixels se deve ao fato de diminuir o esforço computacional nas análises uma vez que, os bancos de dados de pixels podem ser grandes. Vários modelos de regressão foram ajustados e após a escolha do melhor, foram aplicados procedimentos estatísticos para avaliar os resultados. Foi apontado que os setores com menor densidade tendem a ter sua população superestimada, enquanto os setores com maior densidade tendem a apresentar subestimação da respectiva população. No modelo baseado nos pixels a densidade populacional não é considerada, pois os *pixels* têm a mesma área. Uma solução sugerida foi incorporar uma variável auxiliar para indicar a densidade populacional, possibilitando que os modelos fossem ajustados separadamente para setores muito densos e pouco densos. De modo geral, ficou claro que o mais adequado é construir um modelo para cada região estudada, incorporando a ele os dados disponíveis e atualizando as informações sempre que necessário.

3.3.2 Compatibilização Segundo a Granularidade de tempo

Na maioria dos casos em que as variáveis estudadas apresentam elevado grau de confiabilidade, as mesmas geralmente não estão disponíveis na periodicidade desejada. Alguns estudos propõem que para as variáveis que por natureza não apresentam grandes oscilações no período avaliado, seus valores sejam replicados preenchendo as lacunas da série, porém esse procedimento pode ser muito frágil o que requer muito cuidado ao aplicá-lo. Outra solução é a realização de médias, como foi proposto na segunda versão do Índice Mineiro de Responsabilidade Social (IMRS), onde para preencher as lacunas da série de dados os índices foram calculados através das médias de 2 ou 3 anos. Por exemplo, o IMRS do ano de 2004 foi obtido através da média dos indicadores referentes aos anos de 2003, 2004 e 2005.

Em Tassinar (2004) tem-se um exemplo da aplicação da técnica de suavização dos dados em estudos de leptospirose no município do Rio de Janeiro no período de 1996 a 1999. O método escolhido foi o de o *kernel* de intensidade que estima a quantidade de eventos por unidade de área. Essa técnica não-paramétrica estimou a intensidade da ocorrência de casos de leptospirose em toda a superfície analisada. No trabalho proposto por Barbosa (2005), foi aplicado métodos de suavização à demanda do álcool para o período de janeiro de 1980 a dezembro de 1995. Os métodos testados foram: média simples, média móvel simples, suavização exponencial simples, suavização exponencial de *Holt,* suavização exponencial de *Holt-Winters*. Através desses diferentes trabalhos aplicados a diferentes áreas do conhecimento, pode-se perceber a aplicação multidisciplinar da técnica de suavização de dados em estudos de análise temporal.

3.3.3 Compatibilização Segundo o Tema

A escolha do tema que será retratado pelo indicador dependerá da disponibilidade do respectivo dado. Uma vez que sua contribuição é o fato de permitir comparações entre anos distintos, as variáveis envolvidas em seu cálculo devem ser equivalentes em todo o período analisado. Essa equivalência pode ser afetada por problemas relacionados com as diferenças metodológicas que uma mesma variável pode sofrer em cada coleta, ou os dados dessa variável não estão disponíveis para todo o período em estudo. Uma alternativa para minimizar esses problemas é utilizar *proxies* dessas variáveis completando assim as lacunas deixadas pela falta de informação e/ou incompatibilidade de coleta em todo período. O problema relacionado com a periodicidade em que o dado está disponível se assemelha com as questões apontadas na seção anterior referente à compatibilização segundo a granularidade de tempo, assim observa-se que a utilização de *proxies* também é uma boa alternativa para ser aplicada nesses casos.

Podemos perceber que a variável renda é uma das variáveis mais trabalhadas dentre as várias aplicações de *proxies,* fato que pode ser visto nos inúmeros trabalhos propostos. Essa preferência se deve ao fato de que a renda é considerada por muitos pesquisadores como uma excelente medida de condição

social e econômica de uma pessoa ou de sua família. Além dos problemas decorrentes de restrições metodológicas apontadas anteriormente, estudiosos ressaltam que a variável renda pode apresentar distorções referentes às declarações, ou seja, subdeclaração da renda por parte das pessoas com maior poder aquisitivo e em alguns casos, pessoas que declaram ter uma renda maior do que realmente possuem.

O procedimento de *proxy* consiste em identificar e avaliar um padrão no comportamento de uma variável conhecida que atenda aos requisitos metodológicos e de periodicidade do respectivo estudo e supor que esse padrão se assemelha com o da variável desejada. Tafner e Ferreira (2005) propõem uma escala denominada como Escala de Capacidade de Consumo Domiciliar (ECD) que estima indiretamente os rendimentos domiciliares utilizando informações da posse de bens de consumo duráveis. Apesar do prévio conhecimento de que uma família pode ter elevada renda sem ter necessariamente muitos bens duráveis, foi verificado que isso não impossibilita a utilização dos dados de bens duráveis como *proxy* da renda. Foram utilizados dados domiciliares urbanos que possuem energia elétrica e informações sobre a posse de bens duráveis.

Em Rocha Junior (2007) foi utilizado o consumo de energia elétrica residencial como indicador socioeconômico na área urbana dos municípios de Vitória, Vila Velha, Cariacica e Serra, mapeando a condição econômica dessa população. O procedimento consistiu em utilizar o consumo de energia elétrica residencial para estimar a renda domiciliar através da regressão linear simples, sendo necessário utilizar bases cartográficas digitais de limites de setores censitários, limites estaduais e municipais, juntamente com os microdados da amostra do Censo Demográfico 2000 do IBGE, agregados por área de ponderação. Os dados de consumo de energia elétrica residenciais possuíam identificação georreferenciada permitindo sua localização na área de estudo. Para a etapa relacionada ao geoprocessamento das bases cartográficas, foi utilizado o sistema computacional ArcGIS 9.0, a etapa de análise e tratamento dos dados censitários e de consumo de energia elétrica contou com o auxilio do Excel, e por fim, o *Statgraphics Centurion XV*, na versão 15.2.05 auxiliou na realização dos testes dos modelos de regressão e cálculos estatísticos.

Diante do exposto nesta seção, verificamos que a complexidade da técnica envolvida no tratamento da informação com a finalidade de compatibilizar os dados

com granularidades diferentes depende de recursos e conhecimento tecnológico, tempo e principalmente do objetivo que se pretende alcançar. Nessa direção, apresentamos a seguir a necessidade de inserção de técnicas para compatibilizar diferentes granularidades espaciais e temporais no contexto de descoberta de conhecimento em bancos de dados espaciais provenientes de diferentes fontes.

4 DESCOBERTA DE CONHECIMENTO NO PROCESSO DE GERAÇÃO DE INDICADORES SOCIOECONÔMICOS COMPLEXOS USANDO BANCOS DE DADOS ESPACIAIS

As diversas fontes (bancos de dados) podem apresentar informações em diferentes níveis de granularidade, o que impede muitas vezes que essas informações se completem, sendo necessário aplicar transformações aos dados compatibilizando-os de forma a possibilitar os cálculos de indicadores e índices. Sob essa premissa, acreditamos que as etapas do processo de geração de índices complexos constituem em importantes passos para a descoberta de conhecimento em banco de dados espaciais. A metodologia de construção de indicadores varia conforme as características dos dados e/ou limitações em seus cálculos. Um indicador que utiliza dados que são confiáveis e de fácil acesso, seu cálculo pode se tornar simples, contudo não significa que sua contribuição para o estudo seja menos importante. O que define o grau de importância de um indicador são os critérios adotados para a concepção do mesmo, preocupando-se em garantir que os dados utilizados sejam confiáveis, atualizados periodicamente e georeferenciados. Sendo assim, tanto indicadores de fácil cálculo quanto indicadores cujo resultado final só é obtido depois de aplicar técnicas de tratamento da informação, são importantes para estudos que envolvem avaliações socioeconômicas. O problema surge quando o indicador de interesse não pode ser calculado diretamente de seus dados na forma como estão, pois os mesmos se encontram com várias limitações que prejudicarão os resultados finais.

Portanto, este trabalho propõe que as etapas para a elaboração de indicadores socioeconômicos complexos representam passos importantes para a descoberta de conhecimento em bancos de dados espaciais provenientes de diferentes fontes. Pois, antes que o cálculo dos índices possa ser concretizado, na maioria das vezes é necessário aplicar técnicas de tratamento da informação para que os dados possam atender aos critérios necessários para a realização dos cálculos dando origem a indicadores confiáveis que serão utilizados para a concepção do índice. Considerando a variedade das fontes de dados utilizadas, foram realizadas análises de ferramentas disponíveis que auxiliam no suporte e

desenvolvimento de um banco de dados geográfico cujas informações sejam gerenciadas e processadas por diferentes fontes. Assim, um estudo detalhado sobre a modelagem de dados socioeconômicos e apresentado a seguir.

4.1 Modelagem dos dados

Limitações associadas à coleta e disponibilização dos dados, estão presentes a qualquer fonte de informação. Essas limitações podem ser referentes à metodologia aplicada, periodicidade e unidade geográfica em que os dados são coletados e disponibilizados. Por isso, ao calcular qualquer indicador, é preciso observar se tais limitações, ou outras, estão presentes nos dados utilizados. No Brasil, existem várias instituições cujas informações levantadas e disponibilizadas, possuem um elevado grau de confiabilidade, porém não significa que todas as suas informações satisfazem as condições que necessitamos para a construção de índices e indicadores. O IBGE, INEP, DATASUS e etc., são exemplos de onde se encontrar informações confiáveis, apesar de muitas vezes possuírem algumas limitações metodológicas.

A necessidade em obter informações provindas de grandes bases de dados associada ao avanço tecnológico, impulsiona o aprimoramento e o surgimento de ferramentas destinadas à análise de dados, como por exemplo, a ferramenta *On-Line Analytical Processing (OLAP)*, ou Processamento Analítico *On-Line* em português, que possibilita o processamento de grandes quantidades de dados através de análises com distintas agregações dos dados e apresentação dos resultados em forma de tabelas ou de gráficos. Neste contexto, esta seção apresenta a importância que as metodologias - *Data Warehouse (DW)*, *Knowledge Discovery in Databases (KDD)* e *Data Mining* - destinadas à análise e processamento de bases de dados distintas, possuem no processo de construção de indicadores. Em especial, indicadores complexos georreferenciados gerados a partir de bases de dados processadas em instituições públicas.

4.2 Armazenamento dos dados

No Brasil, e em várias partes do mundo, as instituições de poder público possuem grandes registros de informações acumuladas durante décadas. Pesquisas importantes são realizadas todos os anos e ou em determinados, com o objetivo de obter dados referentes aos aspectos populacionais. Uma vez coletadas essas informações, tem-se a preocupação em como armazenar as respectivas bases de dados, cada uma com sua particularidade, em um ambiente único para que dele seja extraído o conhecimento desejado de forma rápida e confiável auxiliando à tomada de decisões.

Começamos com o conceito de *Data Warehouse (DW)*, que em português pode ser traduzido como Armazém de Dados, tem como principal objetivo armazenar um grande conjunto de bases de dados distintas em um único ambiente. Uma busca na literatura mostra que um número grande de autores pondera que a melhor definição de *Data Warehouse* é a sugerida por Inmon e Hackathorn (1997), onde DW é considerado um suporte ao processo de tomada de decisão reunindo um conjunto de informações que estão agregadas e orientadas por assunto sendo não voláteis variando com o tempo. Possui alta capacidade de processamento e armazenamento possibilitando trabalhar com um volume grande de dados disponíveis em séries históricas em vários anos, possibilitando analisar informações de forma rápida visando à melhoria dos processos para a tomada de decisões. A utilização do DW geralmente é demandada por instituições privadas ou públicas que possuem grande volume de dados armazenados e trabalhados. Para que esses dados sejam processados de forma confiável, é necessário que pessoas habilitadas estejam envolvidas no seu processo de elaboração e construção.

Várias sugestões de fluxogramas apresentam de forma esquemática o processo utilizado para a concepção de um DW, que de maneira geral, estão estruturados da seguinte forma:

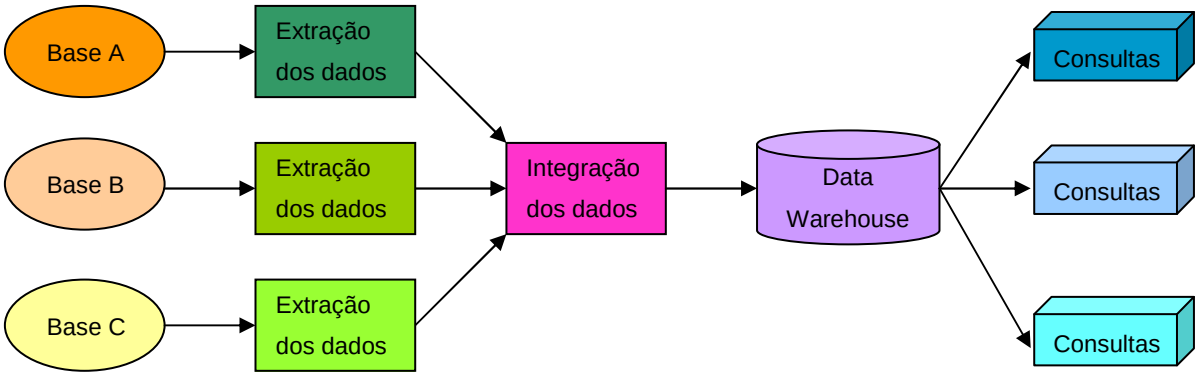


Figura 3: Processo de concepção de um Data Warehouse.

Fonte: Elaboração própria.

A figura 3 apresenta as principais etapas presentes na concepção de um DW. A primeira etapa consiste em verificar os dados de cada uma das bases trabalhadas, essas bases podem ser ou não de fontes distintas, extraindo aqueles que serão utilizados. A segunda etapa se destina à integração dos dados selecionados e no armazenamento dos mesmos em um ambiente único, o DW. Por fim, as consultas são realizadas conforme a necessidade do usuário.

Ao compararmos a figura 3 com a realidade envolvida na coleta e distribuição das informações presentes nas diversas bases de dados utilizadas pelos órgãos públicos, podemos perceber quanto o DW esta presente no processo de armazenamento dessas fontes. De certa forma, podemos dizer que o processo aplicado na concepção do DW é exatamente igual à complexidade envolvida na necessidade de armazenar em um ambiente único, os diversos bancos de dados produzidos pelas instituições governamentais. Pois, para a construção de um indicador ou índice, precisamos selecionar e sintetizar informações que retratem adequadamente a realidade da área estudada, o problema é que essas informações constantemente vêm de fontes distintas onde cada uma está armazenada em um banco de dados distinto, fato que se assemelha a primeira etapa do DW. Posteriormente temos que compatibilizar de forma temporal, geográfica e temática essas informações para que os cálculos dos indicares sejam realizados adequadamente, necessidade semelhante à segunda etapa do DW. Por último, deseja-se obter a informação pronta para análises, ou seja, um índice ou indicador que sintetize as informações de determinada localidade possibilitando uma avaliação

rápida e confiável das condições socioeconômicas com objetivo de aplicar e ou monitorar políticas públicas adequadas com a realidade da população presente.

Nesse caminho, os estudos acerca do DW são de interesse à construção de indicadores, visto que impulsiona o surgimento de técnicas que auxiliam o estudo e tratamento das bases de dados. Em especial, este trabalho se dedica à extração de conhecimento a partir do estudo de padrões, tendências e correlações que podem estar presentes nas bases de dados geográficos e temporais disponíveis pelos órgãos públicos, possibilitando a construção de índices e indicadores que possam auxiliar na tomada de decisão de maneira rápida e eficiente, com a finalidade de avaliar as condições da população em estudo.

4.3 Descoberta de Conhecimento em Banco de Dados

Esse processo é denominado por muitos autores como KDD – *Knowledge Discovery in Databases* (descoberta de conhecimento em bases de dados). Conforme Collazos e Barreto (1999) o termo KDD foi criado em 1989, porém segundo Steiner et al. (2006) o processo de busca e extração de conhecimento, até o ano de 1996 era classificado por alguns autores tanto como KDD como por *Data Mining*, sendo considerados sinônimos. No contexto atual, encontram-se definições que diferenciam os dois termos, vários autores consideram o procedimento KDD sendo todo o processo de descoberta de conhecimento nos dados, enquanto *Data Mining* (Mineração de Dados) se baseia na aplicação de algoritmos e técnicas para extrair possíveis correlações em grandes volumes de dados. De acordo com Collazos, Barreto e Pellegrini (2000) *Data Mining* é uma etapa do KDD que se preocupa em encontrar padrões no comportamento dos dados através de técnicas estatísticas, agrupamentos, programação entre outras.

Seguindo o contexto atual, neste trabalho também se optou por utilizar o termo KDD para representar o conjunto de técnicas utilizadas em todo o processo de construção do conhecimento desde a coleta de dados, tratamento da informação, interpretação, análise e apresentação dos resultados obtidos. O processo de *Data Mining* também foi considerado como uma etapa do KDD, sendo um processo que

explora grandes bases de dados através do uso de algoritmos e técnicas específicas na procura por padrões e correlações consistentes entre as variáveis e posteriormente a validação desses padrões.

A característica multidisciplinar do KDD pode ser vista em estudos e aplicações em trabalhos desenvolvidos em diversas áreas. Na medicina temos o artigo de Collazos, Barreto, Roisenberg (2002) que relatam a utilização de KDD em bancos de dados referentes à mal formação congênita de recém nascidos . Santos (2006) apresenta um trabalho em que KDD foi aplicado ao setor finaceiro mostrando que a descoberta de conhecimento em bases de dados auxilia na tomada de decisões nos problemas de concessão de crédito. Assim, este projeto visa apresentar essa característica multidisciplinar do KDD aplicada a problemas de construção de índices e indicadores sociais e econômicos, utilizando bases de dados distintas tendo como principal chave o georreferenciamento das informações trabalhadas.

Fayad et al. (1996) definem KDD como um método que trabalha grandes conjuntos de dados permitindo identificar padrões confiáveis sendo considerado não trivial, interativo (permite a interferência do usuário), iterativo (pode ser efetuado até alcançar o resultado esperado), e pode ser aplicado em várias etapas dependendo da necessidade do pesquisador. Essa definição está representada na figura 4, onde é apresentada uma visão geral das principais etapas que compõem o processo de KDD.

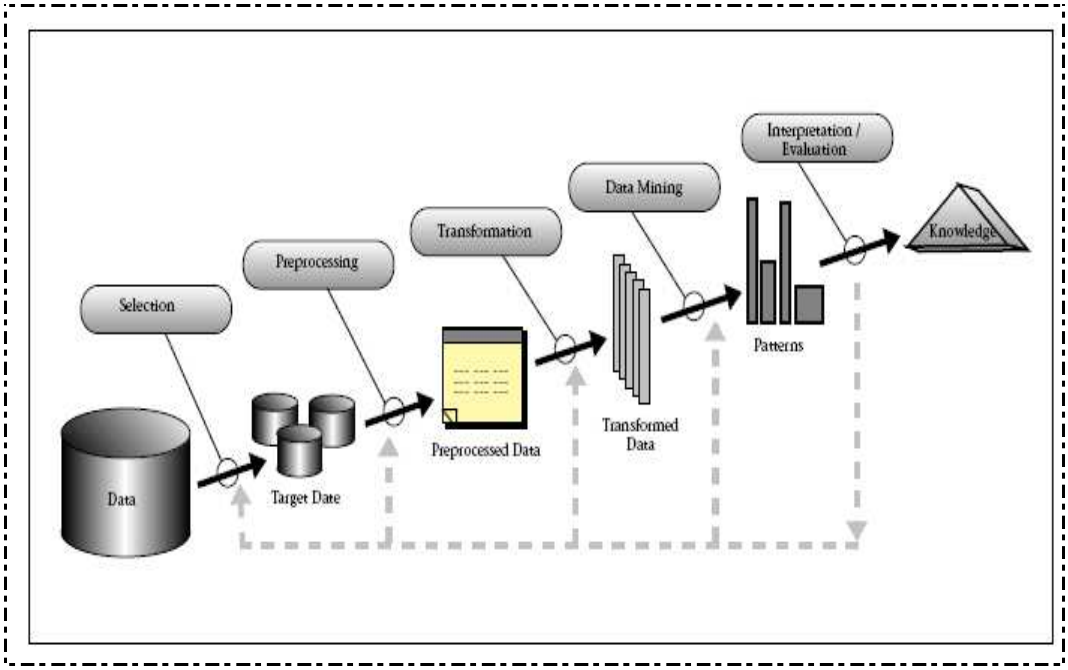


Figura 4: Uma visão geral das etapas que compõem o processo de KDD.
Fonte: Fayyad et al. , 1996

Conforme sugeridas na literatura, as principais etapas do processo de KDD são:

a) Seleção: nessa etapa os dados das várias bases são analisados para posterior seleção daqueles que tendem a atender as condições necessárias para serem utilizados na busca por padrões que possam auxiliar no estudo.

b) Pré-processamento: uma vez selecionados, os dados precisam ser tratados e preparados para serem usados pelos algoritmos, pois esses dados podem apresentar valores inválidos, inconsistentes ou redundantes causando ruídos na base. Portanto tais problemas devem ser identificados e os respectivos dados retirados da base.

c) Transformação: muitas vezes a etapa de pré-processamento não é suficiente para contornar as limitações e ou problemas provindos da forma como os dados estão disponíveis. Assim, quando necessário, é aplicada alguma transformação que busca tornar esses dados compatíveis entre si e entre as condições necessárias para a realização do estudo. Dentre as diversas propostas de transformações,

podemos destacar a lineares, não linear, logarítmicas, exponenciais e etc., além de técnicas de redução de dimensionalidade e de projeção dos dados.

d) Mineração: essa etapa é conhecida como Data Mining e é responsável pela aplicação de algoritmos e técnicas específicas para realizar a busca por padrões, tendências e correlações entre os dados analisados. É considerada por muitos, como a principal etapa aplicada ao processo de extração de conhecimento de bases de dados.

e) Interpretação: por fim, é realizada a análise e interpretação dos resultados adquiridos através da mineração dos dados para a extração do conhecimento provindo de bases de dados. Os principais recursos para a realização dessa etapa são as análises numéricas: equações de correlações, regressão, previsões e etc.; e análises visuais tais como gráficos e mapas.

A seguir é demonstrado como as principais etapas da descoberta do conhecimento auxiliam no tratamento da informação com o objetivo de dar suporte ao processo de construção de indicadores complexos utilizando banco de dados georeferenciados.

4.4 Relação entre a Construção de Índices Complexos e o processo de KDD

A metodologia de construção de indicadores e índices complexos envolve procedimentos necessários ao tratamento das informações, pois as mesmas constantemente estão disponíveis em formatos diferenciados não podendo ser aplicadas diretamente nos cálculos. Portanto, nesses casos é necessário realizar vários procedimentos que tornem essas informações compatíveis entre si e compatíveis com a metodologia dos cálculos que darão origem aos indicadores e índices. Esses procedimentos fazem com que a construção de indicadores e índices transforme os dados em conhecimento. Em especial estamos em busca do conhecimento adquirido a partir do estudo em bases de dados geográficos cujas informações se referem aos aspectos sociais e econômicos da população segundo

sua localização no espaço. Esse conhecimento possibilita identificar problemas específicos de cada localidade e consequentemente auxilia na tomada de decisão de quais políticas serão destinadas a resolver os problemas identificados, assim como posterior monitoramento das mesmas permitido acompanhar sua evolução.

4.4.1 Descoberta de Conhecimento em Banco de Dados Geográficos

Existe grande interesse por meio de órgãos públicos na extração do conhecimento em bases de dados contendo informações sociais e econômicas georreferenciadas, pois essas informações possibilitam analisar e acompanhar as políticas aplicadas pelos governos conforme a área em estudo. Essas análises geralmente ocorrem com o auxilio de informações sintéticas através de indicadores e índices. O problema é como obter informações confiáveis que possam auxiliar nas análises de maneira rápida e eficiente, considerando a diversidade de fontes que as disponibilizam.

Conforme a complexidade do indicador procurado ou o grau de limitações que os dados apresentam, são realizados tratamentos nas informações com o objetivo de adequá-las e viabilizar o estudo. Esses tratamentos variam de acordo com as necessidades ou dificuldades presentes no processo de construção dos indicadores.

4.4.2 Seleção dos dados

Antes de iniciar a seleção dos dados para a construção dos indicadores e índices, é necessário definir as dimensões que serão avaliadas. Por exemplo, se o objetivo for calcular uma medida que represente de forma sintetizada a condição econômica da área estudada, precisamos procurar as fontes de dados que disponibilizam as variáveis candidatas a compor os respectivos cálculos.

Outro exemplo, o IDH é um índice que avalia o desenvolvimento humano através de 3 dimensões: renda (fator que mede se um indivíduo vive em condições de vida digna), educação (avalia se o indivíduo possui acesso ao conhecimento) e longevidade (retrata se o indivíduo apresenta um vida longa e saudável). Após a localização das respectivas fontes, delas são extraídas somente as variáveis que satisfazem as condições temáticas de cada dimensão.

Esta etapa requer que os pesquisadores envolvidos no processo de construção de indicadores, sejam especialistas nos temas abordados, pois são eles que decidem quais variáveis serão selecionadas para a composição dos respectivos cálculos. Podemos perceber aqui a característica interativa do processo de construção de indicadores, uma vez que o pesquisador interfere diretamente na seleção das variáveis decidindo quais serão selecionadas para seguir em busca da descoberta do conhecimento em banco de dados socioeconômicos.

Contudo, é neste momento, que se iniciam os problemas que teremos que contornar nas fases seguintes, pois na grande maioria das vezes os dados selecionados não estão no formato desejado. Isso ocorre devido à diversidade metodológica aplicada no processo de coleta e distribuição das informações por cada fonte de dados.

4.4.3 Pré-processamento

Após a seleção das variáveis de interesse, inicia-se a limpeza dos dados através de análises que identificam a consistência das informações. Nesta etapa o pesquisador avalia cada variável em busca do refinamento necessário que as mesmas precisam passar de modo que possam atender as expectativas dos cálculos. Dentre as diversas formas de análises, são aplicados vários procedimentos que possam auxiliar no pré-processamento das informações. Podem ser aplicados algoritmos apropriados que identificam problemas no conjunto de dados selecionados como: repetição, erros e ausência de informação. Além de identificar os problemas, muitos algoritmos estão ajustados para solucioná-los. Podem ser aplicadas também técnicas mais simples como uma clássica equação de subtração,

cujo resultado pode se tornar tão confiável quanto uma técnica mais sofisticada. No processo de construção de indicadores socioeconômicos a etapa de pré-processamento dos dados além de procedimentos de caráter quantitativo, podem também ser aplicados procedimentos de análises qualitativas, subjetivas. Consequentemente, o pesquisador é o principal responsável pela execução dessa etapa, uma vez que, é de responsabilidade dele a escolha do melhor procedimento a ser aplicado em cada caso analisado. Muitas vezes para a escolha de tais procedimentos, é necessária a realização de vários testes em que diversos métodos são testados até que se decida qual será o mais adequado. Portanto ficam mais evidentes as características interativas e iterativas presentes no processo de construção de índices e indicadores complexos. Assim, para a execução do pré-processamento alguns aspectos relacionados com a forma em que os dados estão disponíveis precisam ser avaliados e decidir qual tratamento será aplicado:

a) Interpretação e padronização dos valores dos atributos

É muito comum que uma mesma variável apresente valores distintos conforme a fonte de origem. Por exemplo, a variável população pode ser encontrada no IBGE, conforme o ano de pesquisa. Isso significa que nos anos em que não ocorrem o Censo Demográfico Brasileiro, os dados de população são estimados através da Pesquisa Nacional por Domicílios. Porém, para esses mesmos anos existem outros institutos de pesquisas que apresentam suas próprias projeções populacionais, como por exemplo, o Cedeplar - Centro de Desenvolvimento e Planejamento Regional de Minas Gerais. Portanto, diferenças metodológicas entre as pesquisas poderão gerar diferentes valores para a mesma variável. Sendo assim, é necessário interpretar cada variável e cada fonte, escolhendo as que contem teor metodológico compatível com estudo em questão.

b) Tratamento nos formatos

Cada fonte disponibiliza suas informações em um ambiente específico, sejam em tabelas, textos, imagens e etc. Nesse momento o pré-processamento se preocupa em organizar essas informações em um mesmo formato, preferencialmente em tabelas, pois esse formato possibilita aplicar operações matemáticas para a realização dos cálculos dos indicadores. Essas tabelas poderão

ser trabalhadas em diversos *softwares* estatísticos como SPSS, Minitab e R e em *softwares* de análise espacial como MapInfo, entre outros. A escolha do *software* a ser utilizado depende de uma série de fatores como o grau de complexidade em que os dados estão disponíveis e o domínio que o pesquisador tem em utilizá-lo.

c) Tratamento de dados incompletos

Algumas variáveis selecionadas podem apresentar dados incompletos por razões técnicas como dificuldade na coleta dos dados ou por questões operacionais como erro ao cadastrar as informações no banco de dados. Dentre as soluções possíveis, podemos destacar as técnicas de interpolação, utilização da média e mediana e análises em séries temporais. Essas técnicas baseiam-se nos valores dos dados existentes para calcular estimativas dos valores ausentes. Uma outra opção, quando possível, é ir a campo tentar coletar as informações que estão faltando. Contudo, essa alternativa apesar de ser extremamente eficaz, é também a mais difícil de ser aplicada na prática. Pois o retorno a campo necessita de investimentos financeiros e de tempo que muitas vezes não estão disponíveis, outro obstáculo surge quando a pesquisa é temporal e o dado já sofreu modificações desde a aplicação da primeira pesquisa.

d) Retirada de informações repetidas

Após a definição das dimensões a serem analisadas, é necessário que cada tema tenha sua variável ou conjunto de variáveis correspondentes ao respectivo estudo e não duas ou mais variáveis que reflitam a mesma ocorrência. Para evitar ambigüidades nos dados que serão utilizados nos cálculos dos indicadores, são retiradas as informações duplicadas escolhendo para permanecer na base as que apresentam características metodológicas mais adequadas ao estudo.

e) Exclusão dos dados

Por fim, aqueles dados que não puderam ser tratados e preparados nas etapas anteriores, são identificados e retirados da base. Portanto, torna-se necessária uma nova seleção de dados. A experiência adquirida possibilita a adoção

e identificação de critérios que possam ser úteis na escolha dos dados promovendo o seu pré-processamento de maneira a minimizar a perda de informações.

Diante do exposto, podemos concluir que a etapa de pré-processamento dos dados, não é uma tarefa muito fácil e que demanda tempo e participação direta do pesquisador para decidir quais estratégias serão aplicadas diante de cada limitação e dos problemas decorrentes dos dados selecionadas para compor a base de dados final. Percebemos também que a iteração e interação dos agentes responsáveis pelo pré-processamento das informações auxiliam no processo de descoberta do conhecimento em bases geográficas de dados socioeconômicos possibilitando adequar esses dados à metodologia aplicada na concepção de indicadores.

4.4.4 Transformação

Uma vez pré-processados, os dados ainda podem passar por transformações, visto que cada fonte aplica uma metodologia de coleta e disponibilização das informações podendo acarretar em diferentes estruturas geográfica, temporal e temática. Portanto, antes de haver uma integração entre essas bases, é necessário avaliar os dados para que cada particularidade seja analisada e compatível com as demais. As técnicas de transformações variam de acordo com o objetivo, pode ser linear, não linear, projeção dos dados, entre outras. No caso dos indicadores, é necessário que os dados envolvidos no seu cálculo, possuam compatibilidade temporal, espacial e temática para evitar possíveis distorções em sua análise. A seguir estão apresentadas algumas das técnicas utilizadas na etapa de transformação de dados socioeconômicos.

a) Suavização dos dados

A suavização dos dados é muito utilizada nos estudos de análises temporais. Isso ocorre devido ao fato que dados pertencentes a períodos distintos podem sofrer alterações metodológicas em sua coleta ao longo do tempo. Ao analisar esse conjunto de informações e verificar diferenças nos valores observados de um

período para outro, temos a tendência a concluir que essas alterações foram decorrentes do próprio dado, porém isso nem sempre pode ser verdade sendo essas mudanças motivadas pela diferentes metodologias aplicadas.

A utilização ou não da suavização dos dados poderá depender do comportamento das variáveis em estudo. Essas variáveis podem ser consideradas estruturais ou conjunturais. Entendem-se como variáveis estruturais aquelas que demoram a sofrer alterações significativas, como por exemplo, a escolaridade, pois o nível educacional da população não se altera rapidamente, é preciso um intervalo maior de tempo para que as mudanças sejam percebidas. Já as variáveis conjunturais sofrem alterações significativas com maior velocidade. Por exemplo, a renda per capita oscila conforme as circunstâncias econômicas e sociais presentes no período analisado. Por isso, é muito comum a aplicação de suavizações em dados econômicos, uma vez que, os mesmos estão mais sujeitos às oscilações decorrentes de fatores externos que afetam sua estrutura. Isso significa que se não levarmos em consideração a possibilidade da ocorrência de tal interferência, podemos concluir que houve um crescimento ou decrescimento econômico em determinada região quando na realidade o que ocorreu foi um caso isolado decorrente de algum acontecimento externo. Temos como exemplos, técnicas de suavizações utilizando métodos conhecidos como exponencial, *kernel* e através de cálculos de médias aritméticas. Além dessas aplicações, podemos citar as transformações lineares, não lineares, projeção dos dados, entre outras.

b) Compatibilização por tema

A seleção das variáveis que comporão cada dimensão estudada, pode se tornar uma tarefa trabalhosa quando o indicador calculado precisa ser avaliado e acompanhado por um período de tempo. O que isso significa? Significa que umas das condições para que uma variável seja selecionada para compor um indicador é o fato de a mesma estar disponível em todo o período analisado e os seus dados sejam coletados seguindo uma mesma metodologia. Como essa condição nem sempre é satisfeita, não podemos simplesmente retirar a variável da base sem antes tentar encontrar solução para o problema, pois essa variável pode ser extremamente importante ao estudo e/ou muitas vezes não podemos substituí-la por outra.

Quando não temos a informação de uma variável durante todo o período analisado, seja por problemas de incompatibilidade da abrangência temporal, geográfica e entre outros, é muito comum a utilização de *proxies* para preencher as lacunas nas séries de dados. Neste trabalho, a utilização de *proxies* é tratada como uma etapa de transformação nos dados em busca do conhecimento, uma vez que, auxilia na construção do indicador para todo o período analisado permitindo a realização de estudos que acompanham, por exemplo, a evolução dos programas de inclusão social implantados em áreas mais carentes. A construção do IMRS, apresentado anteriormente, confirma a aplicabilidade que as técnicas de *proxies* possuem no processo de extração de conhecimento em banco geográficos para a construção de indicadores complexos.

c) Compatibilização de áreas geográficas

Essa pode ser considerada uma das principais transformações que os dados utilizados na construção de indicadores podem sofrer. É muito comum que o limite de uma mesma área seja representado por diferentes contornos. Por exemplo, ao buscarmos informações de uma cidade, podemos encontrar dados referentes aos bairros, unidades de planejamento, unidades políticas. Se tentarmos unir as informações de um determinado bairro com uma determinada unidade de planejamento, mesmos que apresentem uma aproximação de área, os resultados obtidos estarão incompatíveis com a realidade da região estudada, pois geralmente o limite dos contornos dessas duas unidades geográficas não são exatamente os mesmos, o que poderá acarretar em distorções nas análises.

Para que as análises sejam feitas de forma apropriada, ou seja, próximas da realidade, as informações precisam retratar a mesma unidade geográfica. Contudo, para que essa condição seja satisfeita muitas vezes é necessário aplicar transformações que possam compatibilizar os dados dessas diferentes formas geográficas que se referem a uma mesma área.

Uma transformação bastante aplicada e que tem demonstrado resultados bastante satisfatórios, é a compatibilização dos setores censitários para formar áreas homogêneas. Essa transformação possibilita que as informações coletadas e disponíveis por unidades geográficas distintas possam ser compatibilizadas e posteriormente analisadas sem que aja influência de dados pertencentes a uma área

diferente da estudada. O uso dos setores censitários é indicado para a solução de problemas de compatibilização de áreas devido ao rigor metodológico que é empregado na definição de seus contornos, além de se tratar de áreas menores que possibilitam o aumento da homogeneidade das informações. Outra vantagem dessa transformação é a possibilidade que os setores censitários tem de ser agregados em áreas maiores ou quando áreas maiores podem ser desagregadas em setores, dependendo da necessidade do estudo. Essa flexibilidade que os setores censitários proporcionam, auxilia também quando necessitamos compatibilizar informações de uma mesma área, porém em períodos distintos. Por exemplo, um município que em um ano possui determinada área, porém em um intervalo de tempo seu território pode ser expandindo ou divido criando outros municípios ou até mesmos parte de seu território cedido a um município vizinho. Se compararmos as informações entre os períodos sem considerar as mudanças de território, parte das informações que agora pertence a outra localidade será analisada juntamente com os dados reais do município acarretando em conclusões que se distanciam da realidade atual.

4.4.5 Mineração de Dados

A etapa de *Data Mining* é considerada por muitos pesquisadores como a principal etapa do KDD, seu procedimento consiste em utilizar recursos computacionais para aplicar técnicas estatísticas identificando padrões ainda desconhecidos nos dados das diversas bases presentes em um *Data Warehouse*, dentre as técnicas aplicadas na mineração de dados podemos destacar:

a) Classificação

Preocupa-se em buscar uma função que consiga classificar um item a uma ou várias classes pré-definidas, sendo necessário pressupor o conhecimento prévio das possíveis classes e a correta classificação dos itens usados na modelagem. Isso torna a classificação uma técnica preditiva, pois como as classes são pré-definidas ela prevê automaticamente a classe de um novo dado. A análise de discriminante é uma técnica estatística muito utilizada no processo de classificação.

b) Agregação

Essa técnica também é conhecida como *clustering* e dedica-se a buscar similaridades entre as informações de forma a agrupá-las compondo grupos homogêneos onde as classes ou categorias são determinadas pelos próprios dados. Esse método se difere da classificação, pois o estudo é descritivo e as classes não são pré-definidas.

c) Associação

Destina-se em identificar fatos que possam ser direta ou indiretamente associados. Busca encontrar informações que apesar de serem diferentes possuem algum grau de associação fazendo com que possam ser mantidas juntas. É considerada uma técnica descritiva, pois é usada para avaliar padrões em dados históricos com o objetivo de identificar se existe alguma relação temporal entre as informações de uma base de dados.

d) Regressão

É uma equação construída a partir dos dados das variáveis presentes na base de dados. Essa equação explica de forma aproximada o comportamento do conjunto de dados analisado e, portanto considerada uma técnica preditiva.

f) Predição

Através da equação encontrada da análise de regressão, podem-se fazer previsões de um valor futuro para as variáveis presentes na base de dados com base nos estudos dos valores anteriores dessas variáveis.

As técnicas aplicadas na etapa de mineração se assemelham com muitos recursos aplicados em bancos de dados geográficos para extração de informação referente a cada variável utilizada no processo de construção de indicadores. índices. Para a construção de índices que abordam mais de uma dimensão é necessário estudar o padrão de relacionamento entre essas dimensões, nesse

momento podemos perceber a semelhança com a técnica de associação, pois para que um índice sintetize uma realidade, as dimensões retratadas devem estar relacionadas, porém não necessariamente precisam ser semelhantes. Por exemplo, o IMRS é contemplado pelas dimensões: saúde, educação, renda, segurança pública, meio ambiente e saneamento, cultura, esporte e lazer e finanças municipais.

A definição dos temas abordados por cada dimensão retratada no índice se assemelha a técnica de classificação, pois os temas são definidos conforme a dimensão a que pertencem. As características de cada tema são avaliadas para em seguida ser atribuída uma dimensão. Uma vez definidos os temas, agora é a vez de escolher os indicadores que o compõem, lembrando sempre de avaliar as condições geográficas, temporais e de confiabilidades dos dados disponíveis. Mais uma vez pode-se perceber a influência da classificação no processo.

Para a escolha das variáveis que darão origem aos indicadores, várias análises podem ser feitas para estudar o comportamento dos dados. Podem ser aplicadas técnicas de análise descritiva como estudo de médias, desvio-padrão, correlações, análises visuais através de gráficos, regressões e etc.

Nos casos em que se tem interesse em construir índices que retratam áreas homogêneas, é necessário dividir a região em unidades com características semelhantes. Por exemplo, para que um município seja dividido em unidades menores e homogêneas, primeiramente precisamos decidir qual será a característica similar que agrupará os conjuntos. Uma variável bastante utilizada nesse procedimento é a distribuição de renda, assim são formados grupos homogêneos de regiões mais pobres separados das regiões mais ricas. A técnica de agrupar os dados em conjuntos homogêneos é conhecida como *clusterização* e permite que sejam criadas novas unidades cujas características possam ser similares.

Em resumo, a tabela 2 apresenta a estrutura de um índice que leva em seu cálculo vários níveis de informação. Claro que essa estrutura não é única, podendo variar conforme a necessidade e restrições do estudo, se tornando mais ou menos complexa que o proposto. Podemos observar que independente da complexidade do índice, é necessário adotar critérios para a seleção das informações trabalhadas, pois o índice sintetiza uma determinada realidade e uma escolha errada poderá causar distorções no resultado final prejudicando as análises.

TABELA 2: Estrutura de composição de um índice

Índice	Dimensões	Temas	Indicadores	Variáveis
Índice X	Dimensão A	Tema 1	Indicador 1	Váriável 1
				Variável 2
			Indicador 2	Váriável 1
				Variável 2
		Tema 2	Indicador 1	Váriável 1
				Variável 2
			Indicador 2	Váriável 1
				Variável 2
	Dimensão B	Tema 1	Indicador 1	Váriável 1
				Variável 2
			Indicador 2	Váriável 1
				Variável 2
		Tema 2	Indicador 1	Váriável 1
				Variável 2
			Indicador 2	Váriável 1
				Variável 2

Fonte: Elaboração própria.

Diante do exposto, fica clara a característica interativa e iterativa presentes no processo de construção de indicadores, uma vez que em todas as etapas existe a interferência do pesquisador que decide quais procedimentos utilizar, realiza as interpretações dos resultados e toma a decisão final podendo repetir o processo quantas vezes forem necessárias até os dados estarem na modelagem adequada para compor os índices e indicadores.

4.4.6 Interpretação e análise dos resultados

Por fim é realizada a análise e interpretação dos resultados adquiridos através dos procedimentos apontados nas etapas anteriores para a extração do

conhecimento obtido na construção de indicadores complexos envolvendo bases de dados geográficas. Essa etapa do estudo pode ser realizada de diversas formas, desde uma análise numérica dos resultados até uma análise visual da informação.

Os pesquisadores envolvidos no processo de construção dos indicadores avaliam os resultados obtidos através de gráficos, modelos matemáticos e estatísticos ou realizam análises mais qualitativas levando em consideração o seu conhecimento prévio sobre o assunto. Por isso, é importante que os principais atores do estudo sejam especialistas para decidir se o que foi descoberto é um conhecimento relevante que poderá ser aplicado, caso contrário, é preciso recomeçar o estudo melhorando ou refazendo as etapas anteriores até que o objetivo seja alcançado. Portanto, esta iteração no processo de construção dos indicadores permite que sejam feitas várias tentativas até a extração do conhecimento nas bases de dados em estudo, ou concluir que não será possível obter resultados satisfatórios das mesmas.

As etapas envolvidas durante a construção de índices complexos levam ao conhecimento das informações trabalhadas juntamente com o descobrimento de novas técnicas que irão auxiliar outros trabalhos. Assim, apresentamos a seguir a contribuição principal deste trabalho, um modelo que avalia o comportamento da distribuição espacial da população pelo território em análise.

5 MODELO CONCEITUAL DE GERAÇÃO DE INDICADORES SOCIOECONÔMICOS COMPLEXOS USANDO DADOS ESPACIAIS

Este capítulo descreve a elaboração de um modelo conceitual que visa a integração de dados de fontes distintas para geração de indicadores complexos. Este capítulo consolida os demais com a finalidade de demonstrar os conceitos desenvolvidos ao longo deste trabalho, sendo realizado um estudo de caso. Os procedimentos utilizados para a construção de indicadores complexos são propostos visando sua incorporação às técnicas de tratamento dos dados com a finalidade de tornar compatíveis as informações temporais incompatibilidades geográficas. Portanto, pretende-se mostrar que a aplicação proposta aqui fornece uma alternativa para a descoberta do conhecimento em bases de dados geográficas através de procedimentos que viabilizam a transformação dos dados de forma a possibilitar a construção de indicadores e índices complexos a partir de bases cujas informações apresentam alguma incompatibilidade com as demais e que não podem simplesmente ser retiradas do conjunto uma vez que são necessárias ao estudo e não poderão ser substituídas. Sendo assim, a seguir são apresentadas as etapas do procedimento proposto como etapas de pré-processamento e enriquecimento de dados para estudos com indicadores georreferenciados.

5.1 Seleção dos dados: Formulação do banco de dados geográficos

A principal preocupação ao formular um banco de dados geográficos é a seleção de fontes cujas variáveis apresentam confiabilidade, qualidade e periodicidade de atualização suficientes para indicar de forma contínua a evolução das condições socioeconômicas em cada unidade espacial. No entanto, isso representa um desafio para a construção de indicadores complexos usando dados espaciais, pois apesar de uma determinada fonte ser reconhecida pelo seu rigor metodológico, suas informações podem não atender a todas as necessidades de determinado estudo. Devido ao rigor metodológico e alto grau de confiabilidade das informações disponibilizadas, o IBGE foi escolhido como a principal fonte utilizada

nesta proposta, sendo trabalhados os dados do Censo Demográfico Brasileiro. Para a aplicação do estudo, também foram escolhidas dados de outras fontes confiáveis: GEOMINAS e INPE. Portanto, temos aqui a etapa de seleção como o primeiro passo para a descoberta de conhecimento

5.1.1 Modelagem de aplicações: Estudos de casos

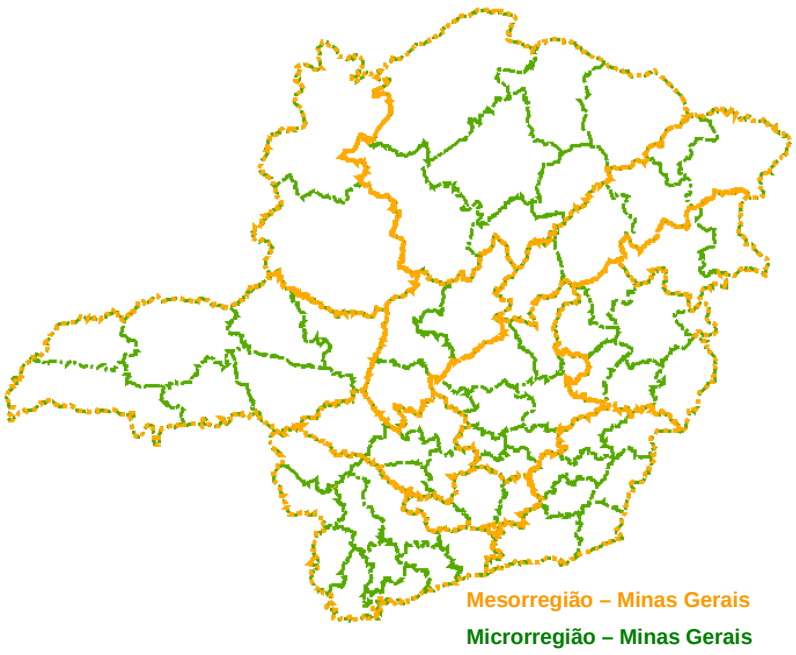
Em todas as aplicações iremos supor que não conhecemos as populações reais das localidades em estudo. No entanto, todas as populações reais são conhecidas e fornecidas pelo IBGE; o objetivo de tais suposições é testarmos os métodos e posteriormente validar as estimativas, comparando-as com os valores reais. As unidades espaciais utilizadas são as microrregiões, mesorregiões e municípios do estado de Minas Gerais.

5.2 Transformação: Estratégias para estimar informações que envolvem diferentes níveis de granularidade espacial

Este estudo descreve a seguir a aplicação de alguns interpoladores espaciais para a estimação de variáveis populacionais para o desenvolvimento de indicadores socioeconômicos. Esta etapa corresponde ao passo de transformação das informações para o enriquecimento dos dados no processo de descoberta de conhecimento. Primeiramente são apresentadas aplicações de um procedimento univariado, a estimação por ponderação de área, e discutidos os principais problemas ou limitações de sua aplicabilidade. Posteriormente são utilizados métodos bivariados para trabalhar e estimar informações com diferentes níveis de granularidades espaciais, variáveis sintomáticas e estimador de *Kernel*, demonstrando os principais ganhos nos resultados em relação aos métodos univariados.

5.2.1 Estimação por Ponderação de Áreas

Considerando a geometria de uma região, foi aplicado o método de ponderação de áreas para estimar a população das microrregiões do estado de Minas Gerais para o ano de 2000 utilizando a informação populacional das mesorregiões para o mesmo ano. Foi escolhido o ano de 2000 por se tratar de um ano censitário cujas informações reais possibilitam avaliar e validar as estimações encontradas. As malhas digitais trabalhadas foram obtidas no site do Geominas e os dados de população no CD do universo do Censo Demográfico Brasileiro divulgado pelo IBGE. O mapa 1 mostra a divisão de mesorregiões (linha laranja) em microrregiões (linha verde). Observe que as respectivas áreas respeitam os contornos umas das outras, ou seja, trata-se de uma hierarquia espacial: estado > mesorregião > microrregião. Este processo assume que a distribuição espacial da população é uniforme em toda a região, o que não é muito realista. Para demonstrar este fato, apresentamos a seguir um exemplo, em que as populações das microrregiões são estimadas por proporção de área a partir da população de cada mesorregião.



Mapa 1: Mesorregião dividida por Microrregião – Minas Gerais
Fonte: Geominas.

Utilizando o *software* TerraView 3.5.0, a malha das mesorregiões foi sobreposta pela malha da microrregião, delimitando as áreas comuns. Utilizando o mesmo *software* foram calculadas as áreas das microrregiões e identificada a mesorregião de origem. As informações obtidas foram exportadas para uma planilha eletrônica, onde foi realizado o cálculo das estimativas nos seguintes passos:

1º passo: Através do *software* TerraView 3.5.0, as áreas das 66 microrregiões foram agregadas segundo as 12 mesorregiões e posteriormente replicadas na mesorregião correspondente.

2º passo: A partir das áreas encontradas no passo anterior, e com o auxílio da planilha eletrônica, aplicou-se o cálculo da fórmula a seguir:

$$P_{Micro_i} = \frac{A(Micro_i)}{A(Meso_j)} * P_{Meso_i} \quad ; \quad i = 1, 2, \ldots n \; ; \; j = 1, 2, \ldots m \qquad (1)$$

Onde:

P_{Micro_i} = estimativa da população da Microrregião i,

$A(Micro_i)$ = área da Microrregião i,

$A(Meso_j)$ = área da Mesorregião j,

P_{Meso_j} = população da Mesorregião j.

A tentativa de estimar as populações das microrregiões corresponde à etapa de pré-processamento das informações com o objetivo de tratar dados incompletos baseando-se em um dado existente (população da mesorregião), no processo de descoberta de conhecimento.

Os resultados encontrados estão dispostos no gráfico 1, verifica-se que houve uma correlação de 0,38 entre a população estimada e a população publicada pelo IBGE, sendo a maior estimativa a população da microrregião de Belo Horizonte. Temos aqui um baixo padrão de correlação, indicando grande fragilidade da estimativa da população levando em consideração a proporção de áreas.

No gráfico 2, o gráfico de dispersão foi refeito retirando a população da microrregião de Belo Horizonte por apresentar maior estimativa e ser a mais populosa, essa decisão partiu pelo fato da microrregião de Belo Horizonte apresentar um “salto” em relação as demais. Esperava-se que assim, ocorresse um aumento da correlação dos resultados encontrados, no entanto houve uma pequena redução.

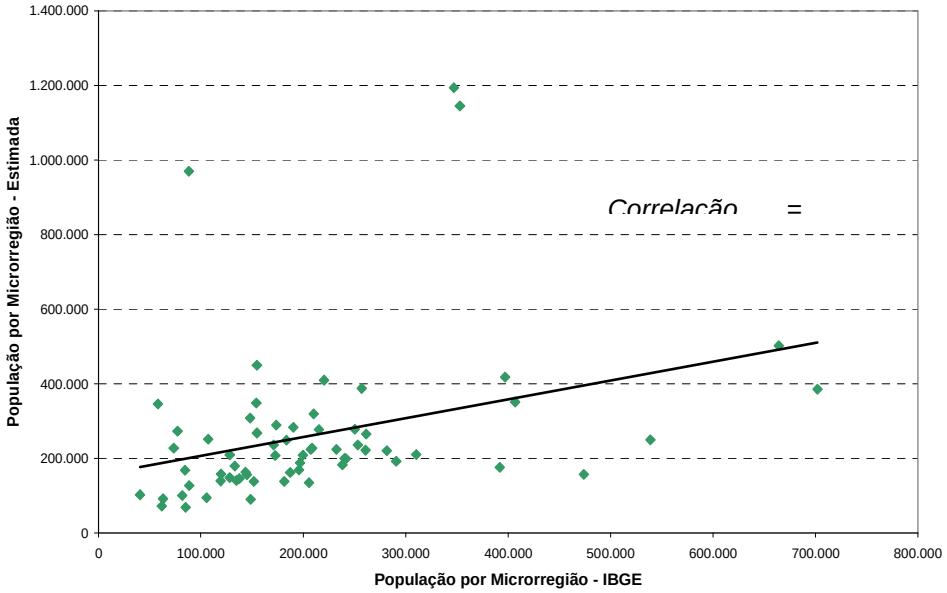


Gráfico 1: Populações das Microrregiões Estimada *versus* Populações das Microrregiões IBGE – Minas Gerais, 2000.
Fonte: Censo Demográfico Brasileiro, 2000.

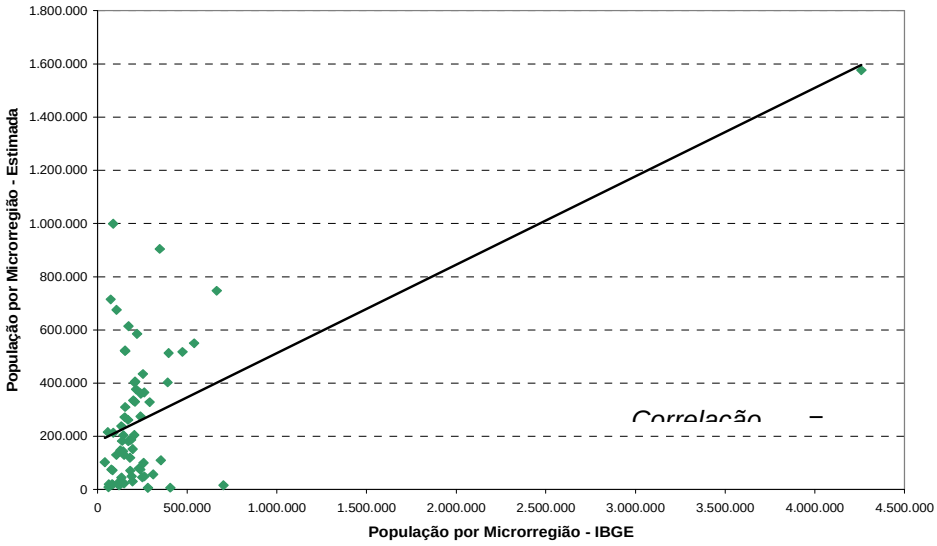
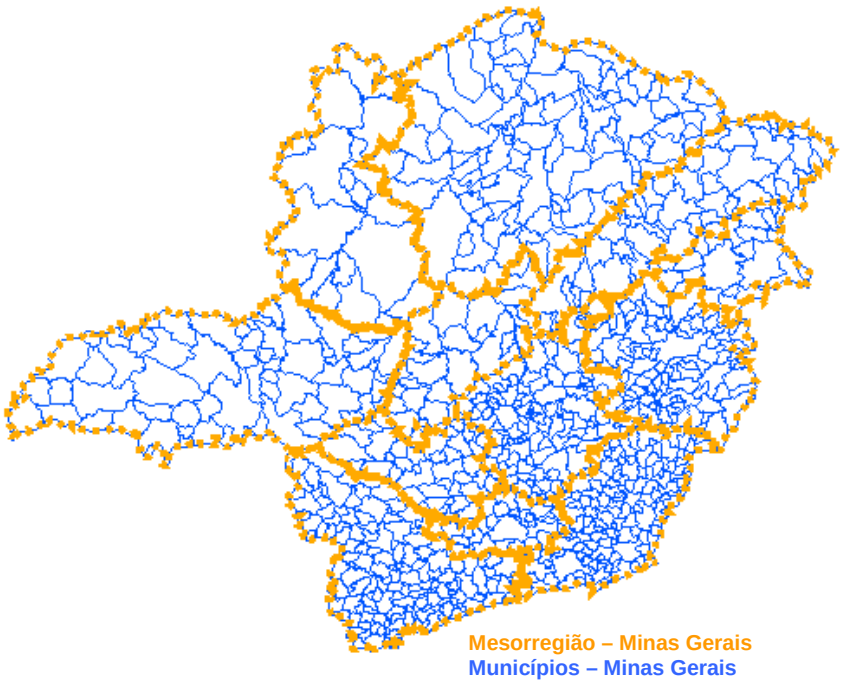


Gráfico 2: Populações das Microrregiões Estimada *versus* Populações das Microrregiões IBGE (sem população da Microrregião mais populosa) – Minas Gerais, 2000
Fonte: Censo Demográfico Brasileiro, 2000.

Utilizando os mesmos *softwares* e passos apontados anteriormente, repetiu-se o exercício, porém agora, utilizando as informações da mesorregião para estimar a população dos 853 municípios em Minas Gerais para o ano 2000. O mapa 2 demonstra as 12 mesorregiões (linha laranja) dividas pelos 853 municípios (linha azul) respeitando seus limites.



Mapa 2: Mesorregião dividida por Municípios – Minas Gerais
Fonte: Geominas.

1º passo: Através do *software* TerraView 3.5.0 as áreas das 66 microrregiões foram agregadas segundo as 12 mesorregiões e posteriormente replicadas na mesorregião correspondente.

2º passo: A partir das áreas encontradas no passo anterior e com o auxílio do *software* Excel, aplicou-se o cálculo da fórmula a seguir:

$$P_{Mun_i} = \frac{A(Mun_i)}{A(Meso_j)} * P_{Meso_i} \quad ; \quad i = 1, 2, \ldots n \; ; \; j = 1, 2, \ldots m \qquad (1)$$

Onde:

P_{Mun_i} = estimativa da população do Município i,

$A(Mun_i)$ = área do Município i,

$A(Meso_j)$ = área da Mesorregião j,

P_{Meso_j} = população da Mesorregião j.

Os resultados encontrados estão dispostos no gráfico 3, a correlação entre a população estimada e a população publicada pelo IBGE foi de 0,17, neste caso, não há indícios de um padrão significativo entre os valores estimados e os valores reais.

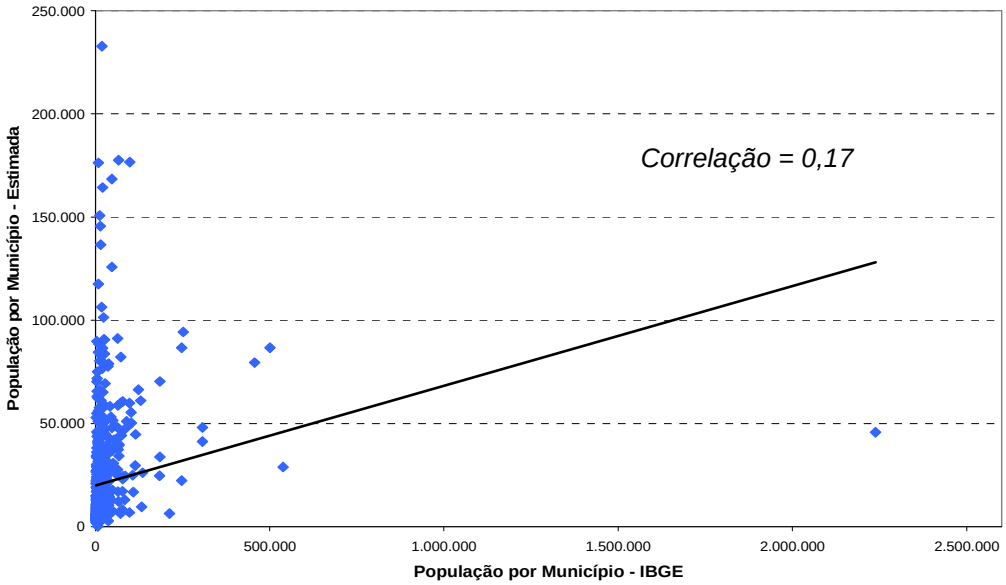


Gráfico 3: Populações dos Municípios Estimada *versus* Populações das Mesorregiões IBGE – Minas Gerais, 2000.
Fonte: Censo Demográfico Brasileiro, 2000.

No gráfico 4 temos o gráfico de dispersão se a população do município de Belo Horizonte, neste caso houve um pequeno acréscimo na correlação das estimativas encontradas passando a ser 0,30, porém os resultados ainda não satisfatórios não apresentam alto padrão de correlação com a realidade. Comparando esses resultados com os apresentados anteriormente, podemos

concluir que a divisão da mesorregião em áreas menores implicará em maiores distorções nas estimativas populacionais. Sendo assim, percebemos que o método de ponderação de áreas não é o mais adequado para obter estimativas populacionais quando a informação trabalhada for dividida em áreas muito pequenas.

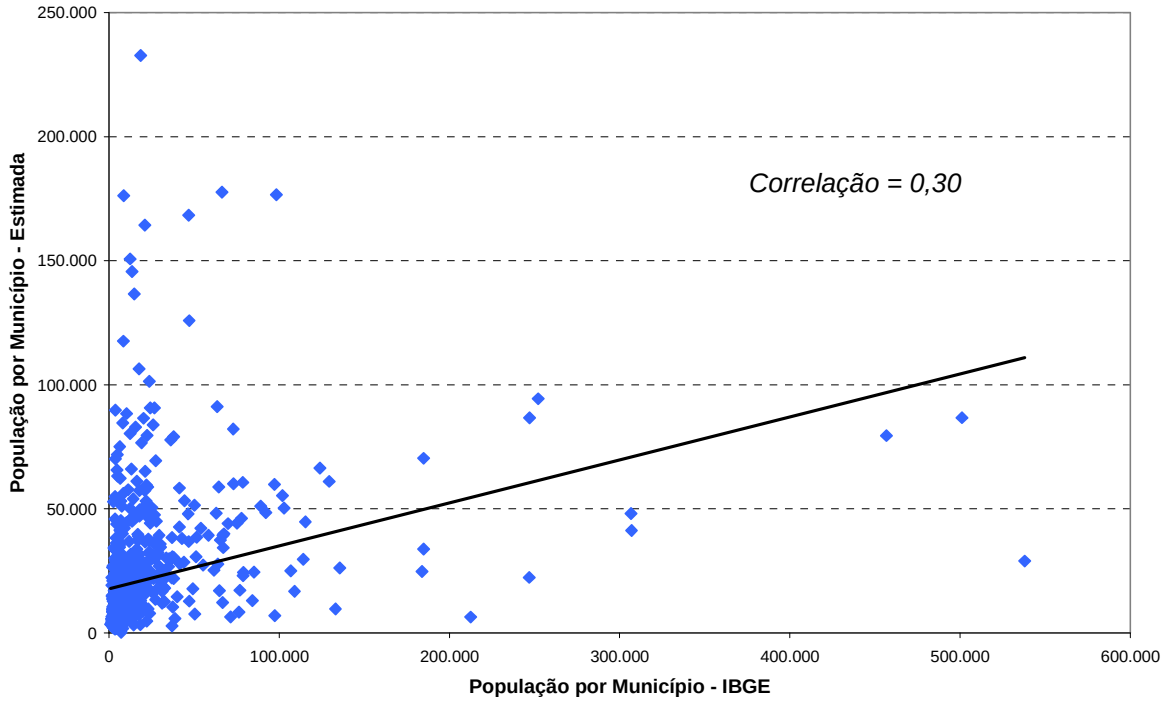


Gráfico 4: Populações dos Municípios Estimada *versus* Populações das Mesorregiões IBGE (sem população do município mais populoso) – Minas Gerais, 2000.
Fonte: Censo Demográfico Brasileiro, 2000.

Os exemplos demonstram que a hipótese de distribuição uniforme da população não é aceitável nos casos tratados. Eventualmente este método pode ser usado em situações em que há uma maior homogeneidade em termos de densidade demográfica, como por exemplo na estimação da população de partes bem definidas de um município. Assim, é necessário estudar métodos alternativos para melhorar a probabilidade de acerto na elaboração de estimativas populacionais quando há variação da granularidade espacial. As subseções a seguir apresentam outros métodos.

5.2.2 Variáveis Sintomáticas

Na literatura, encontramos diversos trabalhos utilizando variáveis sintomáticas para obter uma estimativa de uma subárea. A seguir temos uma descrição de aplicação desse método e posteriormente é apresentada uma proposta de modelagem do mesmo. Temos aqui, além de etapas de pré-processamento de informações com o objetivo de tratar dados incompletos baseando-se nos valores dos dados existentes para calcular estimativas dos valores ausentes, etapas de transformação das informações através de interpoladores com o objetivo de estabelecer compatibilidade espacial dos dados.

5.2.2.1 Considerações iniciais

O trabalho desenvolvido por Jardim (1992) utilizou a distribuição *pro-rata* supondo que a razão entre a população de cada município e o estado do Rio Grande do Sul é igual à razão das variáveis sintomáticas: nascido vivo, óbito, consumo de energia elétrica, eleitor e matrícula escolar. Logo, a população do estado foi distribuída entre os municípios na mesma proporção da participação da variável sintomática como demonstra a fórmula a seguir:

$$P_{h,t} = \frac{S_{h,t}}{S_{T,t}} * P_{T,t} \qquad (3)$$

Sendo:

$P_{h,t}$ a estimativa da população do município h no ano t,

$S_{h,t}$ o valor da variável sintomática S para o município h no ano t,

$S_{T,t}$ o valor da variável sintomática S para o total do Estado no ano t,

$P_{T,t}$ a população total do Estado no ano t.

Sendo assim, foi considerado que a razão entre as populações de cada área menor (município) e a população da área maior (estado) é igual à razão das variáveis sintomáticas

5.2.2.2 Modelagem proposta

Baseando-se na distribuição *pro-data*, este trabalho propõe uma modelagem no método apresentado anteriormente com o objetivo de estimar a população de uma região A cujas áreas sejam uma desagregação de outra região B. Considerando que a área total da região A pode ser dividida em subáreas a_i cujo contorno espacial é delimitado pela sua interseção com a área total da região B, espera-se estimar a população da região A assumindo como variável sintomática os valores da estimativa de *Kernel* da região maior para desagregar sua população para cada região menor. Diante de tais suposições, apresentamos a seguinte fórmula:

$$P_{a_i} = \frac{S_{a_i}}{S_T} * P_T \qquad (4)$$

Onde:

P_{a_i} = estimativa da população da região ai,

S_{a_i} = valor da estimativa de *Kernel* para cada sub-região a_i pertencente à região A,

S_T = valor da estimativa de *Kernel* para o total da região B,

P_T = população total da região B,

Propomos como variáveis sintomáticas o valor da intensidade de eventos por meio da estimativa de *kernel* e o valor dos *pixels* das imagens para a aplicação do

procedimento. No entanto, primeiramente é necessária a apresentação da função de *Kernel* na subseção a seguir.

5.2.3 Função *Kernel*

Com o objetivo de estimar a distribuição da população pelo território espacial, estudou-se o método bivariado de *Kernel,* que possibilita obter uma estimativa de intensidade do padrão de pontos indicando a ocorrência de eventos por unidade de área. A estimativa de *Kernel* é usada em dois momentos: primeiramente como variável sintomática para o procedimento proposto na subseção anterior e posteriormente para desenvolver uma outra proposta para estimar dados com diferentes granularidades espaciais.

5.2.3.1 Considerações iniciais

Nos casos em que o interesse é identificar um padrão em que cada ponto corresponda à ocorrência do evento analisado, podemos utilizar um estimador de intensidade ou eventos por unidade de área. A função de *Kernel* é considerada um processo bivariado, pois permite estimar processos pontuais através de sua densidade de probabilidade que reflete a concentração de eventos da área estudada. As distâncias de cada ponto central em relação aos eventos observados contribuem para o cálculo da intensidade estimada no ponto.

A função de *Kernel* permite estimar processos pontuais através de sua densidade de probabilidade que reflete a concentração de eventos da área estudada. As distâncias *h* de cada ponto central s em relação aos eventos observados contribuem para o cálculo da intensidade estimada no ponto s. Podem também ser atribuídos centróides em cada unidade de área e ponderá-los com os atributos em análise. Assim, uma grade regular é sobreposta à área em estudo para que uma medida, o estimador de *Kernel*, seja redistribuído sobre os centróides. Segundo Amaral (2002) primeiramente a largura do *Kernel* é ajustada conforme a

densidade local dos centróides e posteriormente adapta-se à estrutura da população local, sendo que, em regiões onde a população é pequena, a largura do *Kernel* não se altera, enquanto que nos casos em que a região apresenta alta densidade, a largura do *Kernel* é reduzida proporcionalmente, seguindo uma função de decaimento com a distância afetando a distribuição da população dentro destas áreas. Conforme a autora, deste modo os pesos são atribuídos a cada célula compreendida pelo *Kernel* e utilizados para redistribuição da contagem total da variável de interesse a partir da localização do centróide para as células adjacentes, em que a extensão das áreas com população no modelo final sofrerá influencia da largura do *Kernel*.

Tendo em vista o estudo dos valores associados a cada ponto, tem-se a fórmula a seguir:

$$\hat{\lambda}(s)_h = \sum_{i=1}^{n} \frac{1}{h^2} \left(\frac{s - s_i}{h} \right) y_i \qquad (5)$$

Onde:

s = centro da área estimada
s_i = localização do ponto i
h = tamanho do raio
y_i = valor do atributo associado a cada ponto i
$\hat{\lambda}(s)_h$ = estimativa da quantidade do atributo por unidade de área

5.2.3.2 Modelagem proposta

Primeiramente propomos uma modelagem em que o estimador de *Kernel* é gerado para uma grade sobre uma região B considerando sua área como evento. O mapa de *Kernel* encontrado é sobreposto a uma subdivisão dessa mesma área para auxiliar na desagregação do atributo para as respectivas subáreas.

Em outra proposta, apresentamos a estimação de um atributo de uma região A através da função *Kernel* utilizando como ponto central os centróides de cada subárea a_i resultante da interseção da área da região A com a região B. Assim a área da região B será modelada e uma re-distribuição com *Kernel* encontrado é atribuído aos centróides das subáreas b_i.

Com o objetivo de estimar a população de uma região A cujo formato sobrepõe outra região B. Considerando que a área total da região A é composta por subáreas a_i cujo contorno espacial se difere pela interseção com cada subárea b_i pertencente à área total da região B. Seguindo essas suposições propomos a fórmula abaixo:

$$P_A = \sum_{i=1}^{n} \frac{S_i}{S_T} * P_T \qquad (6)$$

$$S_i = a_i \cap B_i$$

Onde:

P_A = estimativa da população da região A,

S_i = valor do *pixel* da região de interseção da sub-região a_i pertencente à região A com a região de B_i,

S_T = valor do *pixel* para a região B_i,

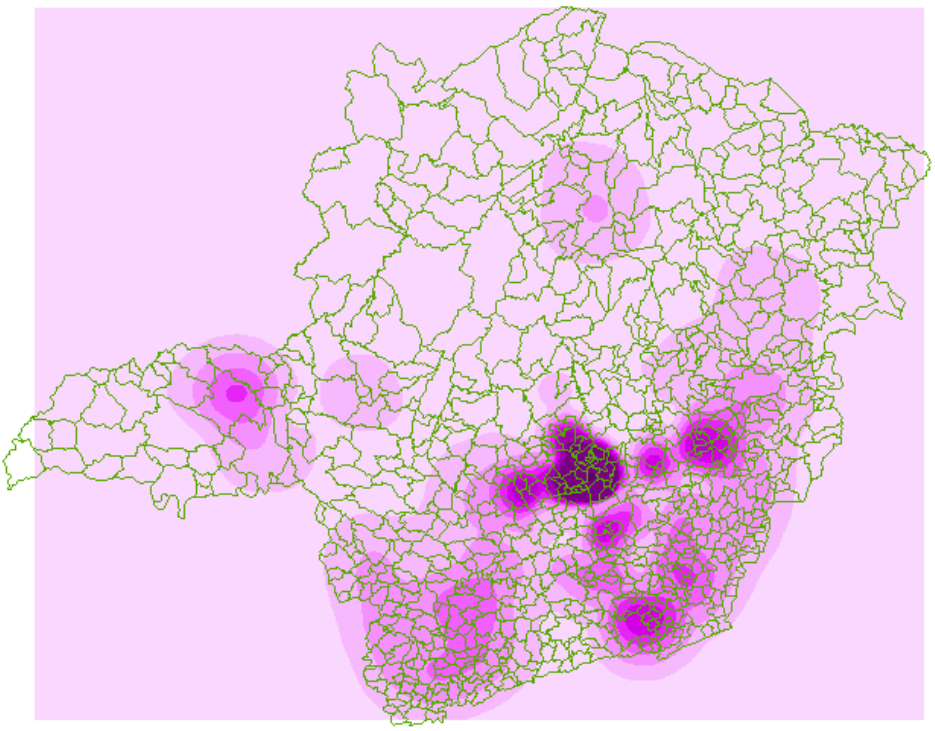
P_T = população total da região B_i,

5.2.4 Aplicações das modelagens propostas: Estudos de casos

Nesse estudo, foi aplicado o mapa de *Kernel* para o Estado de Minas Gerais, atribuído centróides aos seus municípios. Nosso objetivo aqui é encontrar uma estimativa que possa indicar como a população está distribuída pelo território. No *Kernel* Adaptativo não existe um raio fixo, ele varia para cada região de acordo com a distribuição dos pontos. O cálculo do *Kernel* disponível no *software* TerraView é adaptativo e nas situações em que a região apresentar alta densidade, o raio do

Kernel diminui proporcionalmente, dominando apenas uma parte da célula da grade de saída, onde os pesos são aplicados a cada célula envolvida pelo *Kernel* conforme a distância decresce. Estes pesos são usados para redistribuição dos valores do atributo no estudo a partir da localização do centróide para as células adjacentes. Contudo, tem-se o fato de que a cada centróide visitado teremos nas regiões com alta densidade células que receberão a população de vários centróides e nos locais de baixa densidade muitas células não apresentaram população.

Assumindo que a população em cada município se concentra em sua sede, podemos usar a informação da população total e da localização da sede para gerar uma superfície de distribuição de população para o Estado. Usando essa superfície, seria possível estimar a população em qualquer subdivisão do território mineiro. O estudo de caso aplica essa lógica a um exemplo, e verifica a qualidade do resultado. Sendo assim, o *Kernel* foi aplicado utilizando como centróides a localização das sedes municipais. Foi aplicada uma grade sobre as áreas dos municípios, tendo como evento as sedes ponderadas por suas respectivas populações. Neste caso a densidade dos pontos foi calculada através de uma a grade com 1000 colunas utilizando o algoritmo de função quártico e raio adaptativo. Sabemos que a população não está distribuída de forma homogênea sobre todo o território dos municípios e que a intensidade do estimador de *Kerne*l varia conforme as distâncias entre os centróides. Essa variação ocorre de forma inversamente proporcional, sendo maior a intensidade do *Kernel* quanto mais próxima a sede de um município estiver da sede do município vizinho, sendo assim, isso fez com que as regiões do estado que contenham municípios com pequenas áreas tenham seus centróides mais próximos, impactando numa maior intensidade nessas regiões, mapa 3. Como etapa de transformação temos o uso da estimativa de *Kernel* na tentativa de transformar a informação populacional com o objetivo de torná-la espacialmente compatível.



Mapa 3: Estimador de *Kernel* ponderado pela população das sedes municipais
Fonte: Geominas.
Censo Demográfico Brasileiro, 2000.

Para a etapa de pré-processamento o mapa de *Kernel* foi exportado para o *software* SPRING com o objetivo de explorar a grade numérica que foi gerada, pois acreditamos que a grade numérica obtida na construção do *Kernel* forneça um bom indicador sobre a distribuição da população sobre as sedes municipais. A através da ferramenta “Estatística de Imagem por Polígono”, foi possível extrair os valores da grade numérica da imagem por municípios, pois foi gerada uma tabela com algumas medidas estatísticas, dentre elas a estimativa da média de *pixels* por polígono e quantidade de *pixels* por polígono, ou seja, por município. A respectiva tabela foi exportada para o Excel, possibilitando a operacionalização da ponderação dos *pixels* totais do estado pelos pixels de cada município, da seguinte forma:

$$P_i = \frac{m_i}{m_E} * P_E \qquad (7)$$

$$m_i = \alpha_i * \beta_i$$

$$m_E = \sum_{i=1}^{n} m_i$$

Onde:

P_i = População total do polígono *i*;

m_i = Valor total dos *pixels* do polígono *i*;

α_i = Número de *pixels* do polígono *i*;

β = Valor médio do *pixel* do polígono *i*;

m_E = Valor total dos *pixels* do estado;

P_E = População do Estado.

O resultado dessa aplicação é verificado no gráfico 5. A correlação entre a população real e a população estimada segundo os *pixels* gerados pelo mapa de *Kernel* é de 0,73. No gráfico 6, em que é retirada a informação do município de Belo Horizonte, ocorreu pequeno aumento na correlação, que passou a ser de 0,75.

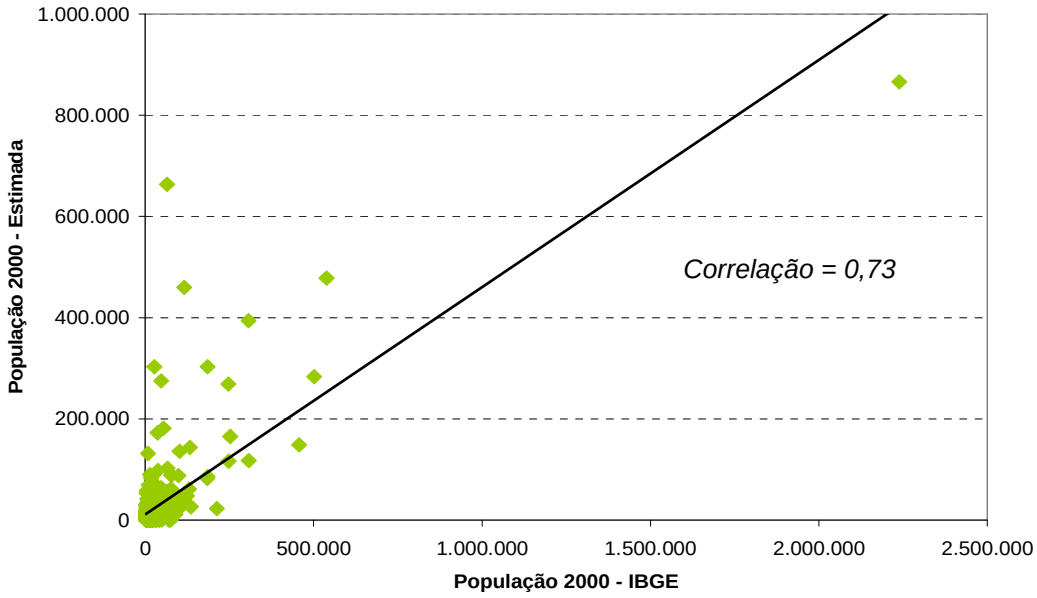


Gráfico 5: População Estimada com *Kernel* ponderado pelas populações das sedes municipais *versus* População dos municípios IBGE – Minas Gerais, 2000.
Fonte: Censo Demográfico Brasileiro, 2000.

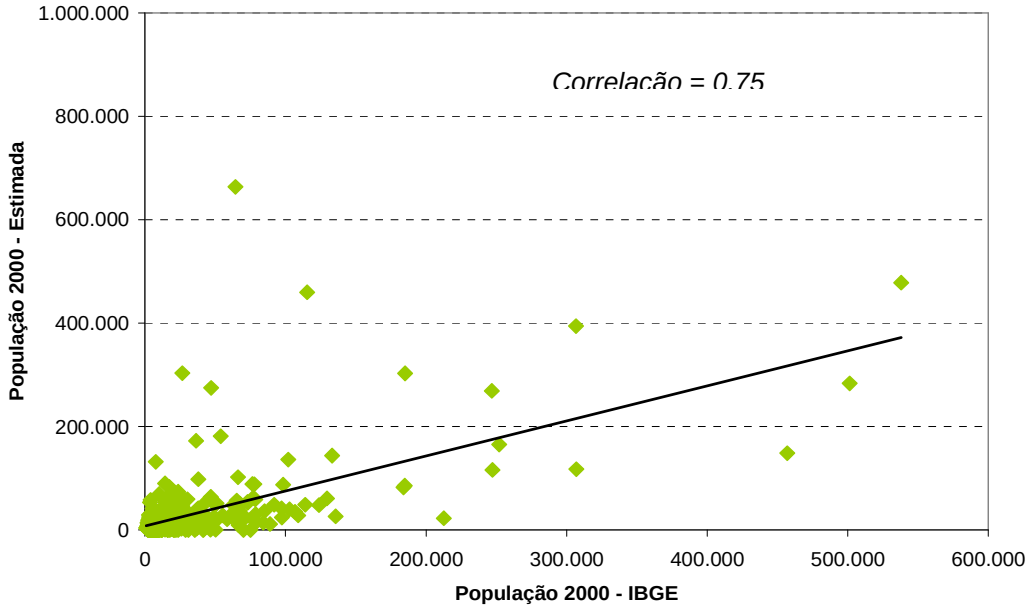
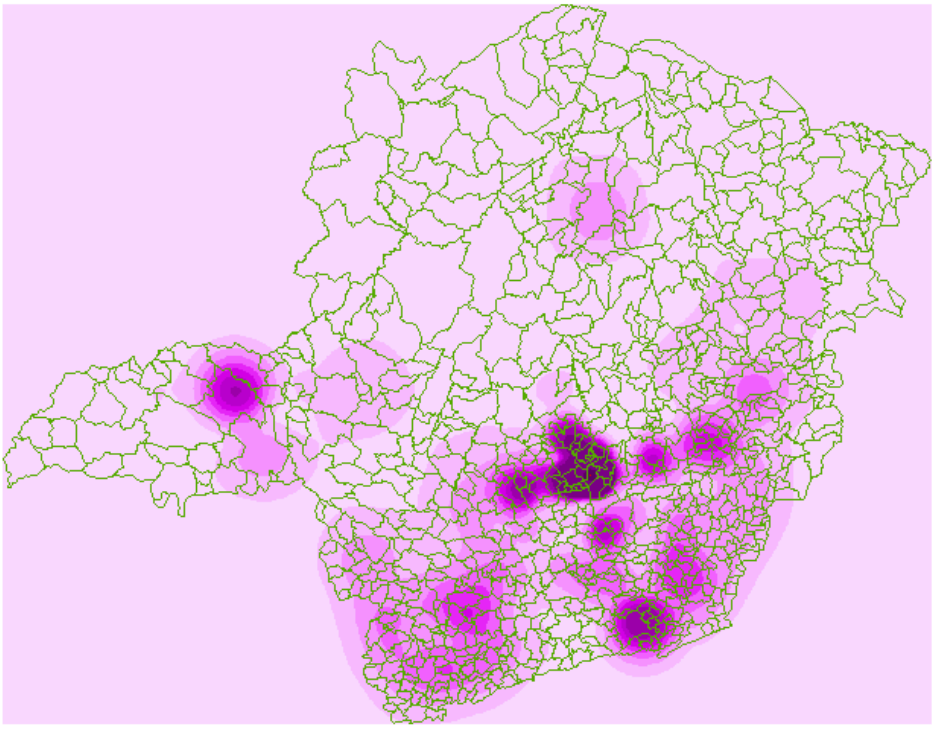


Gráfico 6: População Estimada com *Kernel* ponderado pelas populações sedes municipais *versus* População dos municípios IBGE (sem município mais populoso) – Minas Gerais, 2000.
Fonte: Censo Demográfico Brasileiro, 2000.

Naturalmente, como toda hipótese simplificadora, essa da concentração da população na sede também leva a erros, por exemplo, há outros núcleos populacionais fora da sede em vários municípios. Mas esse erro é menor do que o que ocorre quando se parte de outra hipótese simplificadora, ou seja, que a população se distribui uniformemente pelo território.

Com a intenção de melhorar as estimativas repetiu-se o processo usando como centróides para o mapa de *Kernel* a localização dos distritos municipais. Espera-se que com o maior espalhamento dos centróides pela região tenhamos uma melhor representação das manchas populacionais. Temos como resultado o mapa 4, em que verificamos pequeno aumento na intensidade das manchas em relação ao mapa anterior. Completando a análise, pelo gráfico 7 temos correlação de 0,76 e pelo gráfico 8 (quando tiramos a informação de Belo Horizonte) a correlação diminui para 0,75.



Mapa 4: Estimador de *Kernel* ponderado pela população dos distritos municipais
Fonte: Geominas.
Censo Demográfico Brasileiro, 2000.

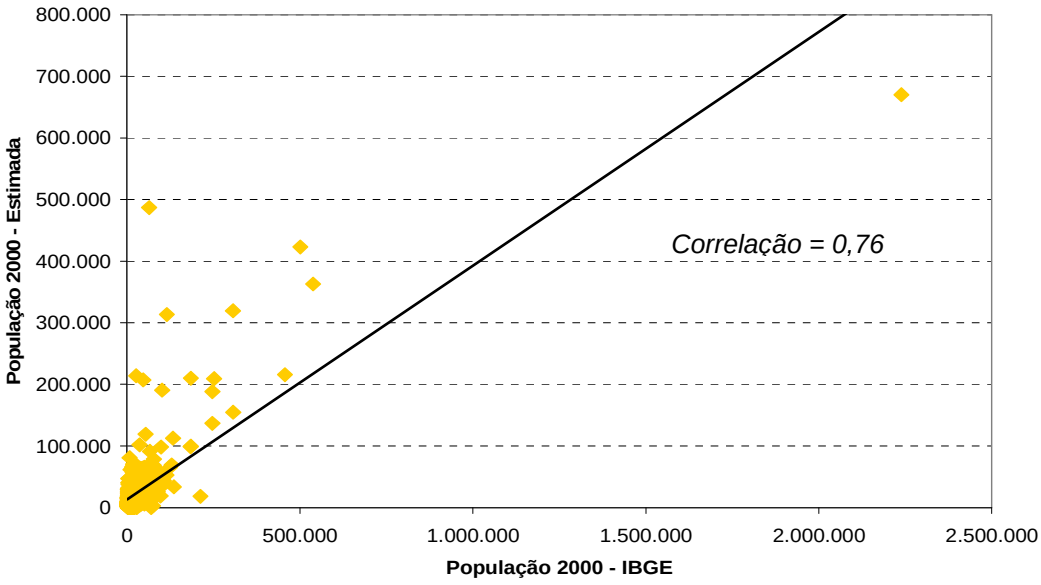


Gráfico 7: População Estimada com *Kernel* ponderado pelas populações dos distritos municipais *versus* População dos municípios.
Fonte: Censo Demográfico Brasileiro, 2000.

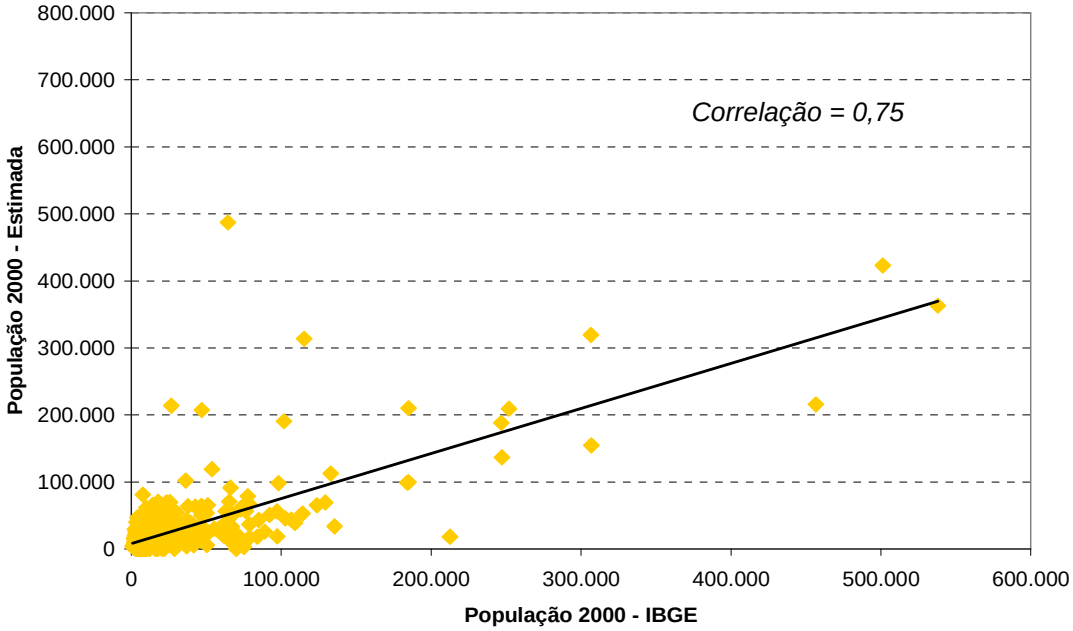


Gráfico 8: População Estimada com *Kernel* ponderado pelas populações dos distritos municipais *versus* População dos municípios IBGE (sem município mais populoso) – Minas Gerais, 2000.
Fonte: Censo Demográfico Brasileiro, 2000.

De modo geral, houve um aumento mais significativo, quando trabalhamos com os centróides dos distritos correlação entre os valores reais e os valores estimados passando de 0,73 para 0,76. Apesar de pequeno, esse aumento da correlação indica que, quanto mais informações tivermos para atribuir aos centróides, melhor será indicação de como a população está distribuída pelo território. Ao comparar os métodos apresentados neste capítulo, verificamos o ganho na estimativa da distribuição populacional através do uso da estimativa de *Kernel* como variável sintomática em relação ao uso do método de ponderação de áreas, figura 5.

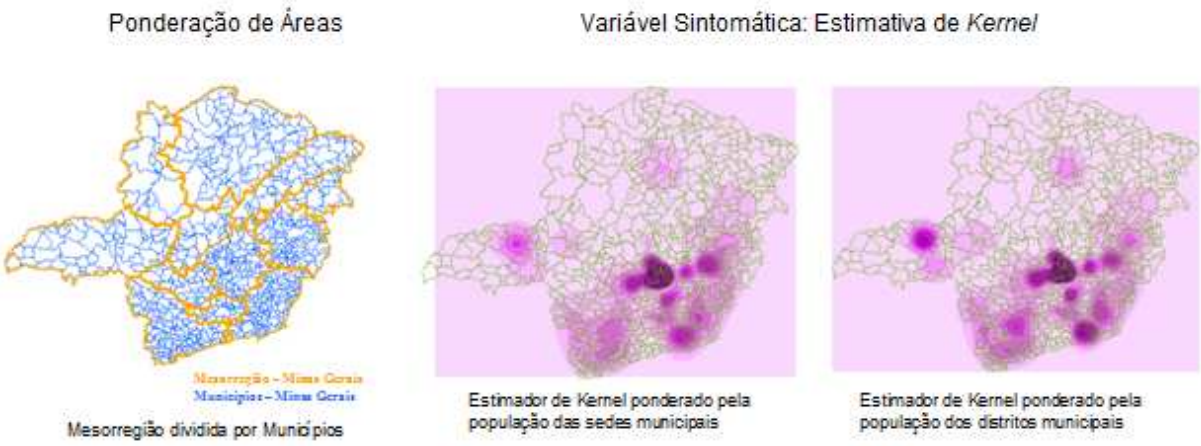


Figura 5: Comparação entre os métodos

6 CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS

Esta dissertação oferece uma contribuição para construção de indicadores complexos através de uma coletânea de metodologias, métodos e técnicas que permitem a compatibilização de dados cujas informações são provenientes de fontes distintas e ou apresentam diferentes granularidades espacial e temporal. No entanto, muitas dessas abordagens levam em consideração somente uma das dimensões apresentadas. Portanto, a importância desta coletânea se deve ao fato de reunir mecanismos das três complexidades em um único ambiente, formando um conjunto de recursos para compatibilização de dados complexos, que possa auxiliar à tomada de decisão. Foi verificado que existem propostas mais elaboradas, porém que demandam maior conhecimento tecnológico e tempo para serem desenvolvidas, assim como existem propostas mais simplificadas, porém eficazes. Ambos os casos apresentam restrições ou limitações metodológicas que interferem no alcance da estimativa, contudo ainda são boas estratégias para o processo de construção de indicadores complexos.

Nesse caminho foi proposto um modelo conceitual para trabalhar com informações com distinta granularidade espacial através da análise da distribuição populacional sobre o território em estudo. Primeiramente foi mostrado que o método de ponderação de áreas para desagregar a população da mesorregião (área maior) para estimar a população da microrregião (áreas menores) forneceu estimativas com baixa correlação (0,30 e 0,38) quando comparadas com os valores reais da região Utilizando o mesmo método para desagregar a população da mesorregião para estimar a população dos municípios, houve correlação de 0,17 a 0,30 entre as estimativas e os dados reais. Portanto, o método de ponderação de áreas não é uma boa alternativa para trabalhar as diferentes granularidades espaciais. Apresentou-se então uma metodologia utilizando a função de *Kernel* para avaliar a distribuição da população sobre o território. A função de *Kernel* foi aplicada considerando como centróides as sedes e distritos ponderados pelas respectivas populações resultando em mapas cujos *pixels* foram usados para estimar as distribuições da população sobre os municípios. Este processo procura modelar uma superfície de distribuição de população, a partir da idéia de que a população de um

município se concentra na sede e se distribui ao redor dela, com densidade que decresce com a distância. As estimativas encontradas apresentaram correlações maiores que 0,70, indicando alta proximidade com os valores reais e demonstrando ser uma boa alternativa para trabalhar dados com diferentes granularidades espaciais. Quando trabalhamos com os centróides dos distritos houve um pequeno aumento na correlação entre os valores reais e os valores estimados passando de 0,73 (ponderados pelas sedes) para 0,76. Apesar de pequeno, esse aumento da correlação indica que quanto mais informações tivermos para atribuir aos centróides, melhor será a indicação de como a população está distribuída pelo território. Para estudos futuros sugerimos utilizar imagens de satélites para obter um indicador da presença humana sobre o território e posteriormente utilizar esse valor para ponderar os centróides das áreas onde será aplicado o mapa de *Kernel* e comparar se existe uma melhor aproximação das estimativas da realidade local.

Foi apresentada também uma visão que considera as etapas envolvidas na construção de índices complexos similares às aplicadas no processo de descoberta de conhecimento em um *Data Warehouse (DW)*, tendo em vista a necessidade de compatibilizar as informações em um mesma realidade geográfica transformando-as em variáveis prontas para serem utilizadas nos respectivos cálculos. Dados de fontes distintas implicam muitas vezes em diferentes granularidades espaciais e temporais, e necessitam de tratamento adequado à informação visando compatibilizá-la, transformando-a em informação consistente para a etapa final de cálculo do indicador complexo. Também para estudos futuros sugerimos a construção de um *Data Warehouse (DW)* inserindo as etapas de geração de indicadores socioeconômicos complexos explorando os conceitos de agregação de dados com granularidades espaciais distintas.

REFERÊNCIAS

ALMEIDA, G. E. S.; QUINTANILHA, J. A; HO, L. L. Análise Envoltória de Dados (DEA) e Geoprocessamento para medir a eficiência na instalação de empresas no município de Osasco, 2007. Disponível em:http://www.lares.org.br/2007/7seminario trabalhos.html. Acesso em: 03 de fev. 2009.

BARBOSA, A. **Análise da Demanda do Álcool Utilizando os métodos de suavização exponencial**. 2005. 70f. Monografia (Bacharelado em Estatística Aplicada) - Universidade Estadual de Maringá, Paraná.

CÂMARA, G.; FEITOSA, F.; MONTEIRO, A. M. Compatibilização de Dados Censitários para Análises Temporais com o Auxílio de Imagens Landsat. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 12, 2005, São José dos Campos, São Paulo. **Anais...** São José dos Campos: INPE, 2005. v. 1. p. 2657-2664.

COLLAZOS L., K. ; BARRETO, J. M. . KDD para o Estudo Epidemiológico das Malformações. In: ENCUENTRO INTERNACIONAL DE TECNOLOGIA, GERENCIA Y MEDICINA PARA EL SECTOR SALUD, 1999, Lima, Peru. **Anais...** Lima: Universidad Catolica del Peru, 1999, v. 1. p. 113-115.

COLLAZOS L., K. ; BARRETO, J. M. ; PELLEGRINI, G. F. . Análise de Prontuário Médico para a Utilização com KDD, 2000. Disponível em: http://www .inf.ufsc.br /~l3c/artigos/Collazos00.pdf. Acesso em: 03 de jun. 2011

COLLAZOS L., K. ; BARRETO, J. M. ; ROISENBERG, M. . Dificuldades na Aplicação de KDD em Medicina. In: WORKSHOP DE INFORMÁTICA APLICADA À SAÚDE, 2, 2002, Itajaí, Santa Catarina. **Anais**... Itajaí: Univali, 2002. p158-169.

DAVIS JUNIOR, C. A. Geoprocessamento: dez anos de transformações. **Informática Pública**, v.4, n.1, p.17-23, 2002. Disponível em: http://www.ip.pbh.gov.br/ANO4_N1_PDF/ip0401davis_geo.pdf . Acesso em: 03 de jun. 2011

DAVIS JUNIOR, C. A. **Múltiplas Representações em Sistemas de Informação Geográficos**. Tese (Doutorado em Ciência da Computação). 2000. 115f. Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, Belo Horizonte.

DIAS, T. L.; OLIVEIRA, M. P. G.; CÂMARA, G.; CARVALHO, M. S. Problemas de Escala e a Relação Área-Indivíduo em Analise Espacial de Dados Censitários. **Informática Pública**, v. 1, n. 4, p. 89-104, 2002.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: An overview. **ADVANCES.** p. 1-27, 1996. Disponível em: https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1011/929. Acesso em 12 out. 2009.

FIDALGO, R. N.; TIMES, V. C,; SILVA, J.; SALGADO, A. C. Propondo uma Linguagem de Consulta Geografica Multidimensional, 2004. Disponível em: http://www.dpi.inpe.br/geoinfo/geoinfo2004/papers/6247.pdf. Acesso em: 23 ago. 2009.

FIDALGO, R. N.; TIMES, V. C.; SOUZA, F. F. GOLAPA: Uma Arquitetura Aberta e Extensível para Integração entre SIG e OLAP, 2001. Disponível em: http://www.geoinfo.info/geoinfo2001/papers/149robson.pdf. Acesso em: 10 jan. 2009.

FROZZA, A. A, MELLO, R. S. Um Método para Determinar a Equivalência Semântica entre Esquemas GML, 2006. Disponível em: http://www.dpi.inpe.br /geoinfo/geoinfo2006/papers/p25.pdf Acesso em: 3 jun. 2009.

GRANERO, J.C.; POLIDORI, M. C. Simulador da dinâmica espacial com representação em um ambiente SIG, 2002. Disponível em: http://www.dpi.inpe.br /geoinfo/geoinfo2002/papers/Granero.pdf Acesso em: 16 jan. 2009.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Disponível em: http://www.ibge.br. Acesso em: 12 jan. 2009.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Censo Demográfico Brasileiro, 2000. Disponível em CD ROM.

Inmon, W. H.; Richard D. H. Como usar o Data Warehouse, **Infobook**, Rio de Janeiro, 1997.

ÍNDICE MINEIRO DE RESPONSABILIDADE SOCIAL. Primeira Versão. Disponível em: http://www.datagerais.mg.gov.br/site/int_imrs.php. Acesso em: 10/11/2009.

ÍNDICE MINEIRO DE RESPONSABILIDADE SOCIAL. Segunda Versão. Disponível em: http://www.datagerais.mg.gov.br/site/int_imrs.php. Acesso em: 10/11/2009

JANNUZZI, P. M. Dossiê Indicadores: Indicadores sociais e as políticas públicas no Brasil. **ComCiência**, v. 96, p. 1-3, 2008.

LOBO, M. L. C.; AMATO, F. Sistema de Microplanejamento da Rede Estadual de Ensino. Universidade Federal do Paraná. **CENTRO INTEGRADO DE ESTUDOS EM GEOPROCESSAMENTO,** 2003. Disponível em: http://www.cieg.ufpr.br/ microplan .ciegufpr.pdf. Acesso em: 20/10/2008.

LÔBO, M. A. A. Método para compatibilizar setores censitários urbanos de 1991 e 2000 aplicado ao estudo da dinâmica populacional da Região Metropolitana de Belém (PA). **Revista Brasileira de Gestão Urbana**, v. 1, p. 71-84, 2009.

NAHAS, M. I. P. Mapeando a exclusão social em Belo Horizonte. **Planejar BH, Belo Horizonte**, v. Ano 2, n. 7, p. 29-34, 2000.

NAHAS, M. I. P. Metodologia de construção de índices e indicadores sociais como instrumentos balizadores da gestão municipal da qualidade de vida urbana: uma síntese da experiência de Belo Horizonte. **Migração e Ambiente nas aglomerações urbanas**, v.1, p. 465-487. 2001.

NAHAS, M. I. P.; PEREIRA, M. A. M.; ESTEVES, O. A.; GONÇALVES, É. Metodologia de construção do Índice de Qualidade de Vida Urbana dos municípios brasileiros (IQVU-BR).. In: ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS DA ASSOCIAÇÃO BRASILEIRA DE ESTUDOS POPULACIONAIS. 15, 2006, Caxambu, Minas Gerais. **Anais...** Caxambu: UNICAMP, 2006. v.1, p. 2598-2618.

OLIVEIRA, S. M. ; SOUSA, R. P. ; DAVIS JUNIOR, C. A. ; AMARAL, F. M. P. Adequação da delimitação dos setores censitários a outras unidades espaciais urbanas. Disponível em: http://homepages.dcc.ufmg.br/ ~clodoveu/files/ 100. 40/AC012.%201996%20Adequacao%20da%20delimitacao%20dos%20setores%20c ensitarios%20a%20outras%2 0unidades %20espaciais%20urbanas.pdf . Acesso em: 013 de jun. 2010

PALHETA DA SILVA, J.M.; NILANDER, R. A.; MATHIS, A.. O uso do geoprocessamento na definição das unidades espaciais para o índice de qualidade de vida urbana do município de Barcarena-PA. In: ENCONTRO LATINO-AMERICANO DE PÓS-GRADUAÇÃO, 14, 2004, São José dos Campos, São Paulo. **Anais...** São José dos Campos: UNIVAP, 2004. v. 1. p. 1-20.

Programa das Nações Unidas. Disponível em: Http://www.pnud.org.br. Acesso em: 12/01/2009.

REIS, I. A. **O estado da arte da integração entre Sistemas de Informação Geográfica e Modelos Inferenciais Bayesianos**. 2005a. 180f. Tese (Doutorado em Sensoriamento Remoto) - Instituto de Pesquisa Espaciais, São José dos Campos.

REIS, I. A.. Estimação da população dos setores censitários de Belo Horizonte usando imagens de satélite. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 12, 2005b, Goiânia, Goiás. **Anais...** Goiânia: INPE, 2005b. v. 1. p. 2741-2748.

ROCHA JÚNIOR, A. R.. Utilização do Consumo de Energia Elétrica como Indicador Sócio-Econômico na Área Urbana dos Municípios de Vitória, Vila Velha, Cariacica e Serra. 2007. 57f. Monografia (Bacharelado em Geografia) - Universidade Federal do Espírito Santo, Vitória.

SANTOS R. S. Aplicação de um modelo preditivo de mineração de dados para apoio à decisão de crédito. 2006. 92f. Dissertação (Mestrado em Ciência da Informação) - Universidade Federal de Minas Gerais, Belo Horizonte.

SOUSA, A. G. Concepção e Validação de um Modelo Multidimensional para Data Warehouse Espacial. 2007. 93f. Dissertação (mestrado em Informática) - Universidade Federal de Campina Grande.

SOUZA, I.M. Análise do Espaço Intra-Urbano para Estimativa Populacional Intercensitária Utilizando Dados Orbitais de Alta Resolução Espacial. 2004. 108f. Dissertação (mestrado em Planejamento Urbano e Regional) - Universidade do Vale do Paraíba, São José dos Campos.

SOUZA, G. O. C.; TORRES, H. G.. O estudo da metrópole e o uso de informações geográficas. **São Paulo em Perspectiva**, São Paulo, v. 17, n. 3, p. 35-44, 2003.

STEINER, M. T. A. ; SOMA, N. Y. ; SHIMIZU, T. ; NIEVOLA, J. C. ; STEINER NETO, P. J. . Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. **Gestão e Produção**, v. 13, p. 325-337, 2006.

TAFNER, P. S. B. ; FERREIRA, M. C . Renda e Consumo de Bens Duráveis no Brasil: uma aplicação da escala ECD nas regiões urbanas do Nordeste e Sudeste 1996/1997. **Revista de Economia**, v. v31, p. 63-82, 2006.

TASSINARI, W. S. ; PELLEGRINI, Debora ; SABROZA, Paulo ; CARVALHO, Marília Sá . Distribuição Espacial da Leptospirose no Município do Rio de Janeiroao Longo dos Anos de 1996 - 1999. **Cadernos de Saúde Pública**, v. 20, p. 1721-1729, 2004.

TORRES, H.G. Informação Demográfica e Políticas Públicas na Escala Regional e Local. **Reunión de expertos sobre población y desarrollo local**. Santiago, Chile: CELADE/CEPAL, 2005. Disponível em: http:// www.centrodametropole.org.br/pdf/ Texto_Celade___Haroldo_Torres2%5B1%5D.pdf. Acesso em: 03 de jun. 2010

UMBELINO, G.J.M ; BARBIERI, A. F. . Metodologia para a compatibilização de setores censitários e perímetros urbanos entre os censos de 1991, 2000 e 2010. In: **Encontro Nacional de Estudos Populacionais**, 16, 2008, Caxambu, Minas Gerais. **Anais**... Caxambu: UNICAMP, 2008. v.1 p. 2010-2028.