

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA URBANA

VERIFICAÇÃO DAS RELAÇÕES ESPACIAIS ENTRE A
DISTRIBUIÇÃO DE DADOS SÓCIO-ECONÔMICOS E A
LOCALIZAÇÃO DE ESCOLAS

CLARISSA TAGLIATTI

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Urbana da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Engenharia Urbana.

Orientação: Prof. Dr. Sergio Antonio Röhm

São Carlos
2010

**Ficha catalográfica elaborada pelo DePT da
Biblioteca Comunitária da UFSCar**

T127vr

Tagliatti, Clarissa.

Verificação das relações espaciais entre a distribuição de dados sócio-econômicos e a localização de escolas / Clarissa Tagliatti. -- São Carlos : UFSCar, 2010. 58 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2009.

1. Planejamento urbano. 2. Sistemas de informação geográfica. 3. Acessibilidade. 4. Escolas. 5. Data mining (Mineração de dados). I. Título.

CDD: 711 (20^a)

Este documento foi preparado utilizando-se o formatador de textos LATEX. Sua bibliografia é gerada automaticamente pelo BIBTEX, utilizando o estilo Chicago.

A meus pais,
pelo apoio e dedicação sempre.

Agradecimento

A Deus por me proporcionar condições para a realização deste trabalho.

Aos meus pais, Reginaldo e Cláudia, e meus irmãos Lilian e Marcos, pelo carinho, apoio e compreensão durante o período em que me dediquei a este trabalho.

Ao Diego pela paciência, compreensão, apoio, companheirismo, e principalmente, pelo incentivo, em todos os momentos.

Aos amigos Sérgio e Edson, muito mais que professores e orientadores. Obrigada pela amizade, apoio e confiança depositados em mim e no meu trabalho.

À professora Suely pelas contribuições e pela disponibilidade sempre.

Aos amigos verdadeiros e companheiros de todas as horas que nem a distância conseguiu separar: Fernanda, Roberta, Tellaroli, Dagoberto, Danilo, Leonardo e Carla. Obrigada pelo apoio e incentivo.

À Pâmela, por toda a ajuda e prontidão na extração de padrões, sempre muito dedicada em me ajudar.

À todos os amigos e funcionários do PPGEU e do DeCiv pela atenção e dedicação.

À Universidade Federal de São Carlos pela oportunidade.

Resumo

Do processo de crescimento urbano, ocorre o inchaço desordenado de regiões dotadas de pouca ou nenhuma estrutura. No caso de localização de escolas públicas esse processo pode gerar situações inusitadas, nas quais ocorre excesso de vagas escolares, em algumas regiões, enquanto em outras tem-se déficit, sujeitando assim os alunos, em alguns casos, a deslocamentos de várias horas. A fim de sanar tais problemas os governos têm investido em programas como o de financiamento ao transporte escolar, para a condução de alunos oriundos de regiões mais afastadas. Este estudo foi realizado com intuito de verificar a distribuição espacial de escolas públicas frente ao nível de renda, escolaridade e idade da população, utilizando Sistemas de Informação Geográfica associado a técnicas de Extração de Conhecimento a partir de Bases de Dados (Mineração de Dados). Esta prática possibilita a descoberta de informações ocultas dentro de grandes bases de dados, que podem ser consideradas interessantes, a medida que confirme, amplie ou conteste o conhecimento já existente sobre o assunto. Adicionalmente, viabiliza uma avaliação mais elaborada sobre o atendimento educacional, face às demandas existentes, contribuindo assim para uma análise crítica da situação destas cidades, do ponto de vista educacional, na busca da atribuição de melhor qualidade de vida à população. A cidade de Araraquara-SP, por ser de médio porte, e considerando a disponibilidade dos dados e a importância para aplicações futuras dos resultados extraídos, foi escolhida como objeto de estudo. Este estudo utilizou técnicas de Mineração de Dados Espaciais, possibilitando a descoberta de informações importantes, como por exemplo o fato de pessoas de baixa renda habitarem áreas periféricas distantes de escolas públicas. Além disso, a análise dos dados apresentou déficit e má distribuição de escolas de ensino médio no município de Araraquara. Estas evidências demonstram a viabilidade da metodologia empregada também para grandes bases de dados.

Abstract

The process of urban progress, cause the disordered regions growth with some or without structure. In case of school's location, this process can generate excess of school places in some regions while in others there is a deficit, thus subjecting the students, in some cases, the displacement of several hours. Trying to solve these problems, governments have been investing in programs such as financing school transportation, for the conduct of students from more remote regions. This study examined the spatial distribution of public schools comparing with the level of income, education and age of the population, using GIS techniques associated with Knowledge Extraction from Databases (Data Mining). This practice allows for the discovery of hidden information in large databases that may be considered interesting, as you confirm, extend, or challenges the existing knowledge on the subject. Additionally, it allows a more elaborated on the educational services provided, given the existing demands, contributing to a critical analysis of the situation of these towns, in educational point of view, searching the assignment of a better quality of life for people. The city of Araraquara-SP, is a medium size city, and considering the availability of data and the importance for future applications of the results extracted, was chosen as the object of study. This study used techniques of Spatial Data Mining, allowing the discovery of important information, such as the fact that poor people reside far away from public schools. Besides, the analysis of the data presented deficit and bad distribution of high schools in the city of Araraquara. These evidences indicated the feasibility of the methodology used also for large databases.

Sumário

Lista de Figuras	v
Lista de Tabelas	vi
1 Introdução	1
2 Acesso a educação no Brasil	4
3 Mineração de Dados Espaciais	8
3.1 Sistemas de Informação Geográfica	8
3.2 Mineração de Dados	10
3.2.1 Regras de Associação e Regras de Associação Espacial	12
3.2.2 Softwares de Mineração de Dados	17
4 Material e métodos	20
4.1 Objeto de estudo	20
4.2 Materiais	21
4.3 Metodologia	22
5 Resultados e discussão	29
5.1 Pré-processamento dos dados	29
5.2 Identificação de padrões	46
5.3 Comparação entre análises de Mapas Temáticos e Interpretação dos Pa- drões extraídos	48
5.4 Verificação da escalabilidade dos procedimentos em relação ao porte dos municípios	49
6 Considerações finais	51
6.1 Principais contribuições	51
6.2 Recomendações para trabalhos futuros	52
Referências Bibliográficas	58

Lista de Figuras

3.1	Sistemas de Informações Geográficas.	10
3.2	Processo de Mineração de Dados.	12
3.3	Tarefas de Mineração de Dados.	13
3.4	Exemplo de uma regra de associação.	14
4.1	Localização do município de Araraquara no Estado de São Paulo (sem escala).	21
4.2	Fluxograma do método utilizado.	22
4.3	Fluxograma metodológico do processo aplicado no estudo.	23
4.4	Associações realizadas no banco de dados durante o pré-processamento dos dados.	26
5.1	Distribuição de Escolas por Tipo de Ensino	39
5.2	Distribuição de Escolas por Porte	40
5.3	Densidade Demográfica por quadras	41
5.4	Distribuição de pessoas em idade de ensino Infantil	42
5.5	Distribuição de pessoas em idade de ensino Fundamental	43
5.6	Distribuição de pessoas em idade de ensino Médio	44
5.7	Distribuição de Renda Média per capita por quadras	45

Lista de Tabelas

3.1	Relação de itens comprados por transação.	15
3.2	Relação de itens frequêntes.	16
5.1	Tabela de análises estatísticas para o atributo <i>infantil</i>	31
5.2	Tabela de análises estatísticas para o atributo <i>fundamental</i>	32
5.3	Tabela de análises estatísticas para o atributo <i>medio</i>	33
5.4	Tabela de análises estatísticas para o atributo <i>0a6anos</i>	34
5.5	Tabela de análises estatísticas para o atributo <i>7a14anos</i>	35
5.6	Tabela de análises estatísticas para o atributo <i>15a18anos</i>	36
5.7	Trecho da tabela final gerada para entrada no software de extração de padrões.	37
5.8	Análise comparativa entre bases de dados dos municípios.	49

Introdução

A organização espacial das cidades brasileiras, oriunda da expansão urbana, tem provocado, entre outros problemas, a degradação da qualidade de vida da população, ocasionando espalhamentos desiguais das atividades, de maneira a não atingir toda a população (Amancio 2005).

O desenvolvimento de uma nação é condicionado por diversos fatores, entre os quais o nível de educação de sua população. Em todas as esferas da sociedade brasileira existe consciência da necessidade de investimentos em educação: criação de programas de fomento aos alunos; estabelecimento de percentuais mínimos de investimentos em função da arrecadação e subsídio ao transporte de alunos que moram em zonas rurais, são exemplos de ações governamentais realizadas neste sentido. Entretanto, apesar dos avanços divulgados oficialmente, a distribuição espacial da rede escolar no território municipal e sua relação com a distribuição da população é uma importante questão relacionada ao acesso desta à educação (Pizzolato, Barros, Barcelos, e Canen 2004).

A deficiência de infra-estrutura educacional ocorre predominantemente em áreas rurais, elevando o índice de evasão escolar. Para minorar esse problema, o governo vem criando alguns programas para transportar os estudantes até as escolas. O mais atual é o Caminho da Escola, financiado pelo BNDES (Banco Nacional de Desenvolvimento Econômico e Social), que fornece transporte para 8,4 milhões de alunos de áreas rurais, do estado do Mato Grosso do Sul (FNDE 2007).

Tradicionalmente, vêm sendo realizadas pesquisas de análises e interpretação manual de mapas e dados estatísticos, associados a um profundo conhecimento da localidade, por quem realiza o estudo. Pesquisas realizadas por Barcelos, Pizzolato, e Lorena (2002), que desenvolveram um estudo semelhante de avaliação de localização de escolas, utilizando modelos capacitado e não-capacitado, e Sistema de Informação Geográfica (SIG), com dados do município de Vitória (ES), aplicando uma metodologia que admitia que toda a população na faixa etária de 7 a 14 anos deveria estar matriculada no ensino fundamental.

Similarmente, o presente estudo utiliza dados sócio-econômicos, por possuírem importância ímpar no processo de tomada de decisão, pois contém conhecimentos e relações a serem explorados. O conhecimento encontrado pode facilitar a compreensão das influências do ambiente na vida do ser humano, e melhorar o planejamento do crescimento para determinada região. Nestes dados podem estar contidas informações de distribuição da população, demanda e oferta de instituições de ensino público, de saúde, e outras infra-estruturas urbanas, que demonstrem as condições e a qualidade de vida da população, principalmente das classes menos favorecidas. Uma análise crítica e profunda destes dados pode auxiliar em possíveis ações dos governos.

A associação dos dados sócio-econômicos a informações espaciais (geográficas) permite uma melhor visualização da distribuição espacial destes. Para isso utilizam-se os Sistemas de Informação Geográfica (SIGs), ferramentas que vêm se consolidando no Brasil em muitas áreas de aplicação com a finalidade de dar suporte a decisões administrativas e executivas. Uma das áreas que se destaca é a aplicação urbana, considerando a capacidade dos SIGs para integrar informações de fontes diferentes e em níveis distintos de responsabilidade em uma organização.

O objetivo principal deste estudo é obter um panorama do atendimento a demanda do sistema de ensino público para o município de Araraquara-SP, por meio de verificação do atendimento da rede pública de ensino infantil, fundamental e médio, em relação à população na faixa etária de 0 a 18 anos e suas características localizacionais e sócio-econômicas.

Os objetivos específicos deste trabalho são:

- Utilizar recursos de SIG combinados a uma técnica computacional denominada KDD (*Knowledge Discover in Databases*) em que se realiza a extração automatizada de conhecimento a partir de bases de dados, também conhecida como Mineração de Dados (MD), para obtenção de novos conhecimentos.

- Avançar no uso dos Sistemas de Informação Geográfica, não só como uma ferramenta para a associação de dados, mas para o uso aplicado para o Planejamento Urbano, com procedimentos de análise pautados em novos paradigmas de análise de dados.
- Verificar a escalabilidade dos procedimentos executados em Tagliatti (2007), em relação ao porte do município a ser analisado.
- Buscar novas formas de discretização de atributos, de acordo com cada tipo de dado, a fim de facilitar a interpretação das regras extraídas.

Este trabalho se justifica pela importância de um diagnóstico focado na situação da população quanto à abrangência dos sistemas de ensino público, municipal e estadual, de acordo com a classe social, idade e outros indicadores sociais que interferem na qualidade de vida dos cidadãos.

A dissertação desenvolvida está dividida em seis capítulos. O Capítulo 2 apresenta uma visão geral do acesso a educação no Brasil, bem como alguns estudos de acessibilidade a educação realizados com o auxílio de ferramentas computacionais. O capítulo 3 traz uma explanação de Sistemas de Informação Geográfica e bancos de dados geográficos, e uma abordagem da técnica de Mineração de Dados envolvendo Regras de Associação e Regras de Associação Espaciais, além de apresentar alguns softwares de Mineração de Dados. A apresentação da metodologia empregada, juntamente com o objeto de estudo e os materiais utilizados compõem o Capítulo 4. Os resultados e conclusões, e as considerações finais constam, respectivamente, dos Capítulos 5 e 6. Ao final são apresentadas as referências bibliográficas utilizadas no estudo.

Acesso a educação no Brasil

Quando na infância, a educação visa moldar o adulto em perspectiva. Para tanto, esta é um direito social de cidadania genuíno. Atualmente, o desenvolvimento tecnológico impõe necessidades de mudanças significativas no processo de integração entre pessoas e grupos. As soluções exigem maior planejamento, por tratarem de ambientes globalizados e não mais restritos a alguns segmentos da sociedade (Pinto 2001).

Martins (2004) analisa que a educação é um direito de todos e dever do Estado, portanto, é incumbência moral do Estado reconhecer que as crianças, marginalizadas social e economicamente, são, juridicamente, portadores dos mesmos direitos que provêm do Poder Público e que definem sua dignidade.

Se não se pode conceber qualquer sociedade contemporânea sem o funcionamento regular de um sistema escolar, menos ainda se pode imaginar um país democrático que não ofereça educação escolar aos seus habitantes. Desde que esta se tornou um direito fundamental, a existência de pessoas não assistidas pelos serviços escolares atenta contra o princípio da igualdade de direitos. Entretanto, ao contrário do que essas constatações geralmente levam a crer, ainda que a oferta dos serviços escolares seja indispensável à democracia, isso não distingue por si só o caráter democrático de uma sociedade no que diz respeito à educação (Ghanem 2004).

Pinto (2001) enfatiza que as metodologias de planejamento de redes escolares, usu-

almente utilizadas no Brasil, centram as discussões em questões relativas aos prédios escolares e avançam sobre alguns temas propriamente educacionais – políticas, metas e normas – sem o que não se viabilizariam. Esse tratamento segmentado é consequência das dificuldades de integração entre os órgãos educacionais que tratam apenas das questões que lhes são afetas. Devido a objetivos nem sempre convergentes que seguem agendas próprias e à adoção de métodos poucas vezes consistentes entre si, os planos superpõem-se mais do que se complementam.

O rápido crescimento das cidades brasileiras, não acompanhado de um planejamento prévio de expansão, vem acarretando transtornos à comunidade, havendo a necessidade das demarcações das áreas do município, tipos de serviços atualmente prestados a comunidade em cada área, sua abrangência e múltiplos aspectos ligados às mesmas (Dutra 1998).

No Brasil o déficit de escolas atinge regiões de extrema pobreza e marginalidade, tendo maior incidência nas regiões metropolitanas, etambém Norte e Nordeste do país. Além das grandes distâncias, as crianças ainda enfrentam dificuldades devido às más condições das vias de acesso, principalmente em períodos de chuva intensa, pois geralmente tratam-se de vias não asfaltadas e de superfície irregular.

Estudos desenvolvidos por Pegoretti e Sanches (2006) apontam que no transporte rural brasileiro, “educação” aparece como a maior motivação das viagens, sendo justificado pela desativação das escolas rurais isoladas, obrigando crianças a se deslocarem diariamente para as escolas na sede do município, nos povoados, aglomerados locais e distritos. Uma pesquisa desenvolvida pelo Fundo Nacional de Desenvolvimento da Educação (FNDE), em 2008, mostra que aproximadamente 4 dos 7 milhões de crianças e adolescentes em idade escolar, na zona rural, são atendidas pelo transporte escolar mantido pelos Estados e Municípios. Porém, muitas caminham cerca de 2 horas até o ônibus, e este possui condições muito precárias (Rocha 2007).

A tradição do planejamento escolar brasileiro recomenda que o tempo máximo que o aluno da zona rural deve levar de sua residência até a escola seja de 45 minutos. No entanto, muitas vezes, esses valores não são respeitados e os estudantes da zona rural precisam enfrentar um tempo de viagem mais longo, por estradas em veículos mal conservados (Pegoretti e Sanches 2006).

Visando a melhoria do acesso ao ensino público, alguns estudiosos desenvolvem análises da acessibilidade a escolas públicas no país, por meio de ferramentas computacionais, facilitando o cruzamento das informações.

Tradicionalmente, vêm sendo realizadas análises por interpretação manual de mapas e dados estatísticos, associados a um profundo conhecimento da localidade, por quem realiza esta análise. Pizzolato, Silva, e Mizubuti (1999), aplicaram uma metodologia de avaliação espacial da localização de escolas no município de Niterói/RJ. Neste estudo, foi admitido que o aluno procurasse a escola mais próxima de sua residência, a partir de análises simplificadas de mapas que continham a distribuição atual das escolas no município e a delimitação dos setores censitários, sendo a ferramenta computacional utilizada apenas para a elaboração de tabelas a partir das análises dos mapas.

Considerando que o geoprocessamento tem se mostrado eficaz na manipulação de dados municipais, gerando produtos que facilitam a interpretação, manipulação e geração de informações essenciais para uma boa administração municipal (Oliveira e Bastos 2007), este pode ser uma ferramenta de apoio a estudos relacionados à localização de escolas.

Barcelos, Pizzolato, e Lorena (2002), realizaram pesquisas na área e desenvolveram um estudo utilizando Sistema de Informação Geográfica para avaliação da localização de escolas, baseando-se na situação existente, apontando como as escolas deveriam se distribuir e supondo que cada escola fosse uma unidade pré-estabelecida. O estudo foi desenvolvido com dados do município de Vitória (ES), aplicando uma metodologia que admitia que toda a população na faixa etária de 7 a 14 anos deveria estar matriculada no ensino fundamental. Foram obtidas soluções em métodos heurísticos, a partir de análises de distribuição espacial por modelo de p-medianas. Trata-se de um algoritmo que emite um relatório de excesso ou escassez de vagas em cada escola, a partir de dados obtidos no SIG, do espaço de abrangência de cada escola, de acordo com a proximidade, divididos em micro regiões. O relatório emitido contém, para cada micro região, o código da escola escolhida e a quantidade de pessoas com idades entre 7 e 14 anos, a maior distância percorrida e o vértice mais distante.

Trad e Cunha (2005) estudaram o acesso à educação no município de Campinas/SP, utilizando software SIG para a confecção de mapas temáticos, a partir de dados do Censo Escolar de 2002, do Censo Demográfico de 2000, da Pesquisa Origem Destino (EMPLASA) de 2003, e do Saesp de 2000. Neste estudo concluiu-se que o relacionamento de dados espaciais com dados estatísticos gera uma análise minuciosa que pode caracterizar a mobilização por parte do poder público, quanto ao atendimento de necessidades básicas do cidadão, porém a vivência desta realidade faz-se necessária para que não ocorram generalizações.

Outra análise do acesso a educação utilizando software SIG foi realizada por Oliveira e Bastos (2007), com dados do município de Canoas/RS. No estudo foram coletados

dados de alguns bairros periféricos, por meio de questionários, e relacionados a dados censitários. Os dados trabalhados foram associados a um software SIG para a espacialização dos dados, que permitiu melhor análise da distribuição destes. Com a espacialização dos dados foi possível o cruzamento de informações de acessibilidade, grau de satisfação com o serviço educacional e atendimento das necessidades da população. O resultado do estudo permitiu a visualização das áreas de maior carência educacional das periferias do município.

Mineração de Dados Espaciais

Este capítulo apresenta uma abordagem sobre as ferramentas computacionais e técnicas utilizadas no estudo.

3.1 Sistemas de Informação Geográfica

Os Sistemas de Informação Geográfica (SIG), de maneira simplificada, integram características físicas do espaço geográfico e das atividades humanas (Chhetri, Corcoran, Stimson, Cooper, e Pullar 2007). Auxiliam na recuperação, classificação e análise de dados espaciais, tratando todas as funcionalidades em um único sistema e incorporando também a análise topológica (análise utilizando relações espaciais de proximidade e vizinhança, por exemplo).

Ferramentas computacionais como os SIGs permitem o acesso a bancos de dados geográficos, viabilizando a automatização do processo de gestão de informações geográficas, constituída de módulos como banco de dados, interface com o usuário, visualização e plotagem dos dados, consulta e análise espacial, integração e gerência dos dados (Câmara e Davis 1999). Melo (2001) enfatiza a importância dos SIGs no sentido de facilitar a produção e análise de informações referenciadas territorialmente, permi-

tindo que num mesmo ambiente sejam tratados e relacionados os dados alfanuméricos e cartográficos.

Os bancos de dados cartográficos ou geográficos contêm uma coleção de dados espaciais e de dados descritivos, relacionados e organizados para armazenagem e recuperação eficiente, além de permitir leitura, entrada, atualização e alteração dos dados por parte de seu administrador (Collaço 2003).

Desde a criação dos primeiros bancos de dados, os ambientes de informação geográfica têm-se diferenciado pelo suporte a decisões com uma dimensão a mais – a espacial, incorporando funcionalidades para a manipulação e análise de dados. Mas o acúmulo, a heterogeneidade e a diversidade dos dados nos dias atuais têm atingido enormes proporções, obrigando a inovações no processo de extração de informações (Barcellos e Ramalho 2002).

Para a construção de bancos de dados geográficos é necessária a escolha de representações computacionais adequadas para capturar a semântica de seu domínio de aplicação, ou seja, oferecer o conjunto mais amplo possível de estruturas de dados e algoritmos capazes de representar a grande diversidade de concepções do espaço (Casanova, Câmara, Davis, Vinhas, e Queiroz 2005).

A concepção de um banco de dados geográficos, de acordo com Casanova, Câmara, Davis, Vinhas, e Queiroz (2005), ocorre basicamente em quatro etapas. A primeira consiste em nomear os processos, de maneira a descrever suas estruturas e relações existentes no mundo real, definindo conceitos que compreendam características e limitações da representação. A segunda etapa é a geração de modelos formais estabelecendo um conjunto de processos lógicos que agrupem diferentes conceitos de aplicação, da forma mais abrangente possível, generalizando, com valores, os diferentes conceitos, de maneira a medir o mundo real. A terceira etapa implica na estruturação dos dados e algoritmos espacialmente (vetores) ou por matrizes. Finalmente, a quarta e última etapa, de implementação, implica na tomada de decisões concretas de programação, considerando aplicações para as quais o sistema é voltado, e disponibilidade de algoritmos para o tratamento de dados geográficos.

A compreensão da distribuição espacial de fenômenos, de acordo com Druck, Carvalho, Câmara, e Monteiro (2005) constitui hoje um desafio para a elucidação de questões importantes em diversas áreas do conhecimento, e tais estudos estão se tornando cada vez mais comuns devido à disponibilidade de SIGs, que permitem a apresentação espacial de variáveis de uma região por meio de mapas. O verdadeiro valor potencial de um SIG é sustentado por sua habilidade em permitir a análise de dados espaciais usando

técnicas de análise espacial, como apresentado na Figura 3.1 (Maguire, Goodchild, e Rhind 1991).

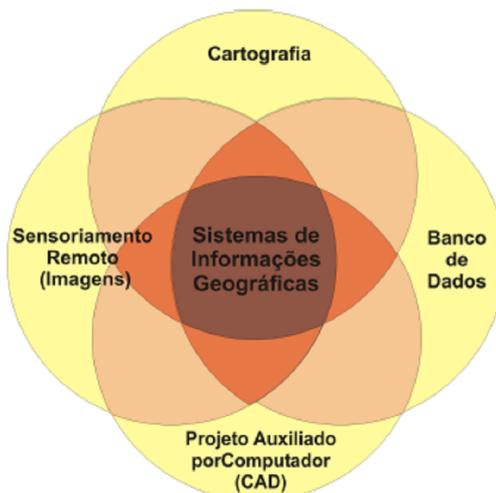


Figura 3.1: *Sistemas de Informações Geográficas.* Fonte: Maguire, Goodchild, e Rhind (1991).

Os SIGs geralmente possuem capacidade de busca e análise espaciais, podendo posicionar geograficamente as características de feições em suas posições relativas, descrevendo objetos do mundo real em termos de sua posição em relação a um sistema de coordenadas conhecido, seus atributos como cor, custo tipo e valor, e suas inter-relações espaciais, que descrevem como estão interligados e como se pode navegar entre os mesmos (Rodrigues 2008).

Rodrigues (2008) ainda afirma que com a propriedade de armazenar informações de natureza gráfica variada, como vetores e matrizes, os SIGs necessitam de módulos ou interfaces que permitam ao usuário incorporar dados e visualizá-los graficamente.

3.2 Mineração de Dados

Desde o início da década de 90 tem-se dado importância considerável a tecnologia de *Data Warehouse*, com características de um grande armazém de dados não volátil, e completa integração entre eles, possibilitando o resgate imediato das informações históricas por meio de algumas estatísticas em relatórios padronizados. Demsar (2007) evidencia que a análise de eventos complexos em pequenos intervalos de tempo, combinando importantes padrões em processos de tomada de decisão, é bem mais efetiva quando auxiliada por computador.

Várias outras ferramentas complementares como mineração de dados e OLAP (*Online Analytical Processing*) vêm sendo incorporadas ao ramo conhecido como KDD (*Knowledge Discovery in Databases* – Extração do conhecimento de bases de dados), para que pela agregação, relacionamento, estatística, visualização, etc., o analista possa obter informações de dimensões inatingíveis pelos métodos convencionais. Estas tecnologias já estão sendo incorporadas nos SIGs, derivando os conceitos do KDD, com o surgimento do *Spatial Data Warehouse*, além de ferramentas aplicadas para mineração de dados espaciais, e SIG/OLAP, com potencial nas instituições que necessitam de grande quantidade de informações precisas e acessadas rapidamente, já que todos os procedimentos para construção são automatizados por parâmetros estabelecidos antecipadamente (Barcellos e Ramalho 2002).

O processo KDD é definido por Fayyad, Piatetsky-Shapiro, e Smyth (1996) como o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis que estejam embutidos nos dados. Existem diferentes abordagens para a divisão em etapas do processo de descoberta de conhecimento, também denominado Mineração de Dados (MD). Este trabalho abordará a divisão do processo de MD proposta por Rezende, Pugliesi, Melanda, e Paula (2003), que define as etapas: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento (Figura 3.2). Trata-se de um processo interativo e iterativo, em que cada etapa, ou todo o processo, pode ser repetido para a obtenção do resultado desejado com ajuste de parâmetros. Neste contexto, a Regra de Associação é uma técnica utilizada na etapa de extração de padrões.

Na etapa inicial de identificação do problema, de acordo com Domingues (2004), é realizado um estudo do domínio da aplicação e identificados e selecionados os conjuntos de dados a serem utilizados para extração de conhecimento.

Os dados selecionados para o processo de MD normalmente não estão em formato compatível com o software de extração do conhecimento. Erros de digitação, geração de dados incorretos, limitações de memória, tempo de processamento, entre outros podem também inviabilizar a extração de padrões em todo o conjunto dos dados (Sionara 2006).

Faz-se necessário então um tratamento destes dados com seleção dos atributos relevantes à pesquisa, além da simplificação destes a fim de facilitar a interpretação das regras extraídas, como criação de intervalos, atribuição de valores, entre outros, para entrada no software de extração de padrões. Este tratamento é realizado na etapa de pré-processamento dos dados.



Figura 3.2: *Processo de Mineração de Dados.* Fonte: Rezende, Pugliesi, Melanda, e Paula (2003).

Após o pré-processamento dos dados, estes estão prontos para servirem de entrada no software de extração de padrões, para a obtenção do objetivo inicial do estudo. Os padrões extraídos são considerados conhecimento, em relação aos conceitos teóricos que os sustentam (Ramos e Santos 2003). A escolha do software de extração depende da linguagem de representação dos padrões a serem encontrados, ou seja, do tipo de conhecimento que se deseja extrair.

Na seqüência os dados são interpretados e simplificados a partir de métodos e procedimentos que na maioria das vezes utilizam medidas de avaliação para processar o conhecimento, tais como a filtragem do conhecimento, interpretação e explanação, avaliação, além de integração do conhecimento (Melanda 2004).

Depois de avaliado e validado, na etapa de utilização do conhecimento, este é utilizado como apoio a algum processo de tomada de decisão, podendo resolver conflitos potenciais entre o conhecimento existente e o obtido com o processo de MD (Domingues 2004).

3.2.1 Regras de Associação e Regras de Associação Espacial

A Mineração de Dados contempla duas categorias: a descritiva e a preditiva. A Associação é uma técnica descritiva e corresponde a extração de conhecimento a par-

tir de exemplos, construindo representações de fácil compreensão (Figura 3.3). Essas representações são regras que indicam o quanto a presença de um conjunto de itens X está relacionada com a presença de outro conjunto de itens Y distintos nos mesmos registros (exemplos) (Ladner, Petry, e Cobb 2007).

A possibilidade de descoberta de todas as associações existentes nas transações de uma base de dados é o aspecto de maior relevância da técnica. Entretanto, esta característica determina a geração de um número elevado de regras, dificultando sobremaneira a interpretação do conjunto de regras pelo usuário. Além disso, em alguns casos pode levar à descoberta de conhecimento muito óbvio ou irrelevante que pouco acrescenta ao conhecimento já existente sobre o domínio (Melanda e Rezende 2003).

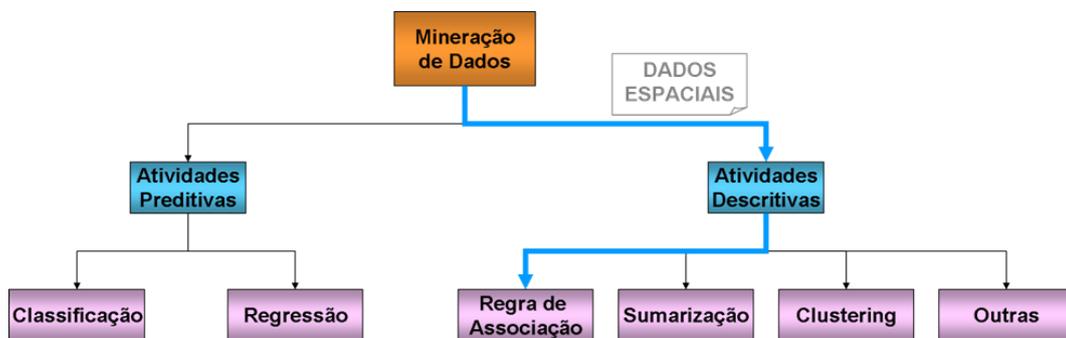


Figura 3.3: Tarefas de Mineração de Dados. Fonte: Rezende, Pugliesi, Melanda, e Paula (2003)

O formato de uma Regra de Associação pode ser representado como uma implicação na forma $LHS \rightarrow RHS$, em que LHS e RHS são, respectivamente, o lado esquerdo (*Left Hand Side*) e o lado direito (*Right Hand Side*) da regra, definidos por conjuntos disjuntos de itens (Domingues 2004).

Agrawal e Srikant (1995) e Koperski e Han (1995), definem Regra de Associação Espacial:

Seja $P = \{p_1, p_2, \dots, p_m\}$ um conjunto de predicados, $R = \{r_1, r_2, \dots, r_m\}$ um conjunto de predicados espaciais representando relações espaciais tal que $P \cap R \neq \phi$, D um banco de dados espaciais composto por um conjunto de objetos espaciais $O = \{o_1, o_2, \dots, o_n\}$, sendo cada objeto i composto por um conjunto de predicados de P , isto é, $i \subset P$. Os predicados de cada objeto i estão em ordem lexicográfica e n representa o número total de objetos. Considerando que $S = \{s_1, \dots, s_m\}$ e $Q = \{q_1, \dots, q_n\}$ são conjuntos de predicados, uma regra de associação espacial é uma implicação na forma

$S \rightarrow Q$, em que $S \subset P$, $Q \subset P$, $S \cap Q = \phi$ e $S \subset R$ ou $Q \subset R$. A regra $S \rightarrow Q$, ocorre no conjunto de objetos em D com $\text{Confiança} = \text{conf}$ se em $\text{conf} \%$ dos objetos em D que ocorre S , ocorre também Q . A regra $S \rightarrow Q$ tem $\text{Suporte } sup$ se em $sup \%$ dos objetos em D ocorre $S \cup Q$.

Da definição de Koperski e Han (1995) entende-se que uma Regra de Associação Espacial (RAE) é expressa $S = \{s_1, \dots, s_m\} \Rightarrow Q = \{q_1, \dots, q_n\}$ ($sup \%$, $conf \%$), onde pelo menos um dos predicados $\{s_1, \dots, s_m\}$ ou $\{q_1, \dots, q_n\}$ é espacial, considerando-se valores mínimos de Suporte e Confiança.

O fator Suporte indica a ocorrência relativa da regra de associação detectada dentro do conjunto de dados de transações. Por ser determinado pela razão entre o número de transações que sustentam a regra e o número total de transações, este fator é uma média relativa. O fator Confiança de uma regra de associação é o grau com o qual a regra é verdadeira entre os registros individuais, e é calculado pela razão do número de transações sustentando a regra pelo número de transações sustentando somente o corpo da regra. A Figura 3.4 apresenta um exemplo de Regra de Associação, na qual o fator Suporte é 40% e o fator Confiança 80% (Diniz e Neto 2000).

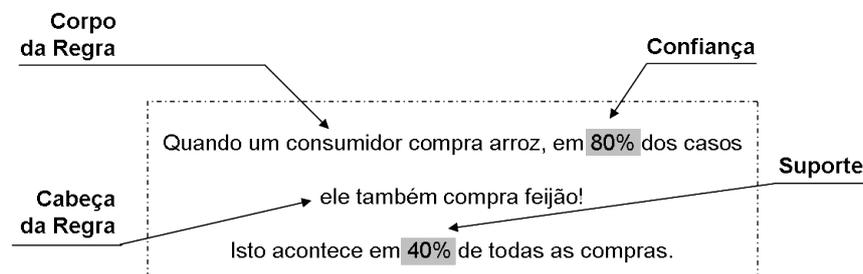


Figura 3.4: Exemplo de uma regra de associação. Fonte: Adaptado de Diniz e Neto (2000).

Grandes fatores de Suporte e Confiança representam um grau de relevância maior que os baixos fatores, entretanto, a implicação de baixos fatores de Suporte e Confiança é a existência de várias possíveis associações que não são, na realidade, o que se busca (Diniz e Neto 2000).

Um processo de Mineração de Dados que utilize Associação, gera um elevado número de Regras de Associação (RA). Essas regras podem representar alguma particularidade de uma transação ou não serem relevantes diante do número total de exemplos fornecidos. Assim, somente são consideradas regras que apresentem valores de Suporte e Confiança superiores a valores mínimos especificados durante o processo de MD.

O conceito de Mineração de Dados pode ser ampliado para considerar atributos

com características espaciais. Estas características são geralmente relações topológicas, como proximidade, adjacência, “está contido”. Desse modo, pode-se estender os métodos de Mineração de Dados para análise de dados espaciais, definindo-se a Mineração de Dados Espaciais (MDE). O que difere MDE das técnicas usuais de análise espacial é o uso de conceitos de Inteligência Artificial e de Aprendizado de Máquina.

Sendo assim, as RAE geradas descrevem a implicação de uma característica espacial por outra, em termos de atributos ou relações espaciais, em um determinado banco de dados. Por exemplo, $próximo(X, Y)$ e os objetos relacionados X e Y . O termo predicado refere-se aos atributos descritivos de objetos espaciais, considerando-se que o objeto espacial e o valor que o atributo assume estão relacionados. Essa relação poderia ser representada da seguinte maneira: $próximo(X, 900)$, sendo X escola cuja distância do setor censitário é 900 metros.

Domingues (2004) exemplifica Regra de Associação (RA) da seguinte maneira:

Supondo uma base de dados que contenha o conjunto de itens $A = \{bermuda, calça, camiseta, sandália, tênis\}$ e um conjunto de transações $T = \{1, 2, 3, 4\}$, no qual a relação de itens comprados por cada transação t_i é apresentada na Tabela 3.1.

Transações	Itens comprados
1	calça, camiseta, tênis
2	camiseta, tênis
3	bermuda, tênis
4	calça, sandália

Tabela 3.1: Relação de itens comprados por transação.

Considerando o valor de Suporte mínimo 50% (2 transações), e confiança mínima também 50%, é possível obter todas as Regras de Associação contidas na Tabela 3.1 utilizando dois passos:

1. Encontrar todos os itens contidos na Tabela 3.1 e que possuam Suporte (Sup) maior ou igual ao suporte mínimo (*itemsets frequentes*). Na Tabela 3.2 são apresentados todos os itens frequentes.
2. Com os *itemsets frequentes*, ou seja, itens que ocorreram com uma frequência maior ou igual a 2, gerar todas as RA contidas na Tabela 3.1 que neste exemplo são:

Itemsets Frequentes	Suporte
{tênis}	75%
{calça}	50%
{camiseta}	50%
{camiseta, tênis}	50%

Tabela 3.2: Relação de itens frequentes.

Regra 1: *tenis* → *camiseta*

Suporte = $\text{Sup}(\text{tênis}, \text{camiseta}) = 50\%$, que é igual ao Suporte mínimo.

$$\text{Confiança} = \frac{\text{Sup}(\text{tenis}, \text{camiseta})}{\text{Sup}(\text{tenis})} = \frac{50}{75} = 66,66\%,$$

que é maior que a Confiança mínima.

Regra 2: *camiseta* → *tenis*

Suporte = $\text{Sup}(\text{camiseta}, \text{tênis}) = 50\%$, que é igual ao Suporte mínimo.

$$\text{Confiança} = \frac{\text{Sup}(\text{camiseta}, \text{tenis})}{\text{Sup}(\text{camiseta})} = \frac{50}{50} = 100\%,$$

que é maior que a Confiança mínima.

Diversos trabalhos abordam a questão de extração de Regras de Associação Espaciais. Malerba, Espisito, Lisi, e Appice (2002), criou o ARES (*Association Rules Extractor from Spatial Data*), um sistema de descoberta de Regras de Associação Espacial, formado de vários algoritmos, entre eles o SPADA (*Spatial Pattern Discovery Algorithm*) que realiza a Mineração de Dados Espaciais. O SPADA trabalha com vários níveis de granularidade, ou seja, gera as regras com vários intervalos de suporte e confiança mínimos, ao mesmo tempo. Pivato (2006) desenvolveu o Módulo de Extração de Relações Topológicas (MERT), que busca relacionamentos necessários para extração de regras de associação espaciais a partir de relações implícitas nos dados. Seu diferencial é o fato de considerar os dados que, por sua própria definição, possuem algum relacionamento a ser utilizado como mais um critério na geração do conhecimento. Bogorny (2006) estudou a problemática da grande quantidade de conjuntos frequentes de regras de associação gerada pelos algoritmos de mineração de dados espaciais, e desenvolveu o aplicativo de extração de padrões Weka-GDPM (Waikato Environment for Knowledge Analysis – Geographic Data Preprocessing Module), que realiza automaticamente a

etapa inicial de pré-processamento dos dados, otimizando o tempo de trabalho, no que se refere às relações topológicas.

Mennis e Liu (2005) utilizaram uma das técnicas de Mineração de Dados, a Extração de Regras de Associação, utilizando dados espaço-temporais. Esta foi aplicada na exploração relações espaciais e temporais em dados geográficos com características sócio-econômicas, com dados da região do Colorado (EUA).

Já Santos (2004) desenvolveu o sistema Padrão, que funciona a partir do ambiente de trabalho da ferramenta Clementine, um aplicativo comercial de Mineração de Dados. O Clementine foi “alimentado” com um conjunto de bases de dados que permitem a identificação de padrões, considerando a componente espacial dos dados. Os princípios adotados no Padrão podem ser utilizados em qualquer outra ferramenta de Mineração de Dados, já que o importante é a informação disponibilizada nas bases de dados, que permitem aos algoritmos de Mineração de Dados considerarem a componente espacial na procura dos padrões. No estudo realizado por Santos (2004), foram implementados dados de municípios portugueses de uma das ilhas dos Açores, e as relações espaciais do tipo direção e distância, existentes entre regiões adjacentes. Utilizou-se o SIG GeoMedia, Intergraph Co., para identificar as relações espaciais existentes, a partir de áreas e respectivos centróides, com o auxílio de programação Visual Basic. Para o desenvolvimento do estudo também foi utilizado o processo de MDE proposto por Santos (2004).

Tagliatti e Melanda (2008) concluíram um trabalho pioneiro, onde observou-se que a utilização de dados sócio-econômicos juntamente com a técnica de Mineração de Dados Espaciais (MDE) fornece informações que retratam a relação entre as características e a abrangência do ensino, de acordo com a renda e a idade escolar da população. Porém, dependendo da base de dados utilizada, o conhecimento extraído pode ser óbvio, sem informações inovadoras (Tagliatti, Melanda, e Pivato 2006).

3.2.2 Softwares de Mineração de Dados

Existem diversos softwares de Mineração de Dados para extração de Regras de Associação. Tagliatti e Melanda (2007) desenvolveram uma análise de viabilidade de utilização de alguns destes softwares. São eles:

Apriori

O software Apriori desenvolvido por (Agrawal e Srikant 1994) utiliza a definição

de Suporte diferente da original, permitindo a utilização de métricas complementares na geração de regras, mas exige a especificação do valor mínimo de métrica. O Apriori busca encontrar os itens individuais e seus respectivos suportes para determinar quais são frequentes, formando o primeiro conjunto de *itemsets*. No passo subsequente, esse conjunto é utilizado para gerar *itemsets*, denominados *itemsets* candidatos, acrescentando um item a cada elemento do conjunto. A seleção de *itemsets* frequentes a partir dos candidatos é realizada por meio do cálculo do Suporte para os candidatos, desprezando aqueles com Suporte abaixo do mínimo. Esse processo é iterativo uma vez que os *itemsets* frequentes encontrados darão origem aos novos *itemsets* frequentes, pois o Suporte de todos os futuros *itemsets* candidatos seriam inferior ao Suporte mínimo (Pivato 2006).

Magnum Opus

O aplicativo Magnum Opus é comercializado pela Rule Quest e permite a definição do número máximo de regras a serem geradas e do número máximo de atributos no lado direito, gerando regras com apenas um item do lado esquerdo. Ele possui um filtro para as regras consideradas improdutivas e desinteressantes, e recursos limitados para o tratamento de atributos contínuos (Webb 1995).

WEKA

Desenvolvido pela Universidade de Waikato na Nova Zelândia, o aplicativo WEKA (Waikato Environment for Knowledge Analysis) possui ferramentas para pré-processamento dos dados, classificação, regressão, clustering, Regras de Associação e visualização, além da implementação do algoritmo Apriori. Este aplicativo inclui novos algoritmos de maneira simplificada, e permite considerar, adicionalmente ao Suporte e Confiança, métricas para avaliação de regras com Lift, Leverage e Conviction. Permite também o estabelecimento do número de regras a serem geradas, e exige, para Regras de Associação, que os atributos sejam categóricos (Witten e Frank 1999).

WEKA-GDPM

O WEKA-GDPM (Geographic Data Processing Prototype), desenvolvido por Bogorny (2006), é um módulo de processamento de dados geográficos que integra ferramentas de mineração de dados clássicas do WEKA, desenvolvido pelas especificações do OGC (Open GIS Consortium). Fornece usos de interfaces gráficas para a execução do processo todo de KDD (*Knowledge Discovery in Databases*), estando implementada uma variedade de algoritmos de Mineração de Dados. Seu enfoque principal é automatizar o processamento de dados geográficos para a redução do tempo na mineração de dados. A vantagem deste aplicativo é o fato

da etapa de pré-processamento dos dados ser realizada automaticamente demandando menos trabalho e tempo.

MERT

O Módulo de Extração de Relações Topológicas, desenvolvido por Pivato (2006), extrai relações topológicas para posterior aplicação de um algoritmo de mineração de Regras de Associação, quantificando as ocorrências das relações entre os objetos espaciais. Para tanto os dados georreferenciados devem estar armazenados em um banco de dados espaciais. Este módulo foi desenvolvido em C++, e as relações topológicas com as quais ele trabalha são disjunção, adjacência, cruzamento, sobreposição, inclusão, intersecção, igualdade, cobertura e proximidade. Também possui métodos de importação de dados para diferentes formatos, sendo o principal para arquivos *Shapefile*, além de um filtro para eliminação das regras que não possuem relações topológicas, abordagem possível somente para o algoritmo Apriori de extração de regras de associação, disponível na própria interface do MERT.

Material e métodos

Neste capítulo são apresentados o objeto de estudo, os materiais e a metodologia utilizados no desenvolvimento deste trabalho.

4.1 Objeto de estudo

O Município de Araraquara está localizado a uma altitude de 664 metros entre as coordenadas 21° 47' 40" Latitude sul e 48° 10' 32" Longitude oeste, na região central do Estado de São Paulo, Brasil, juntamente com mais 14 municípios.

Os municípios vizinhos de Araraquara são Américo Brasiliense, São Carlos, Ibaté, Ribeirão Bonito, Boa Esperança do Sul, Gavião Peixoto, Matão, Motuca, Rincão e Santa Lúcia, e a área ocupada pelo município é de 1006 km² (Figura 4.1).

Conta com cerca de 200 mil habitantes, sendo, portanto classificado como município de porte médio, densidade demográfica de 191 hab/km² e grau de urbanização de 96 por cento, de acordo com o IBGE (2008).

Segundo o Sistema Nacional de Indicadores Urbanos (SNIU 2008), o município de Araraquara possui uma taxa de analfabetismo decrescente nas últimas décadas, caindo

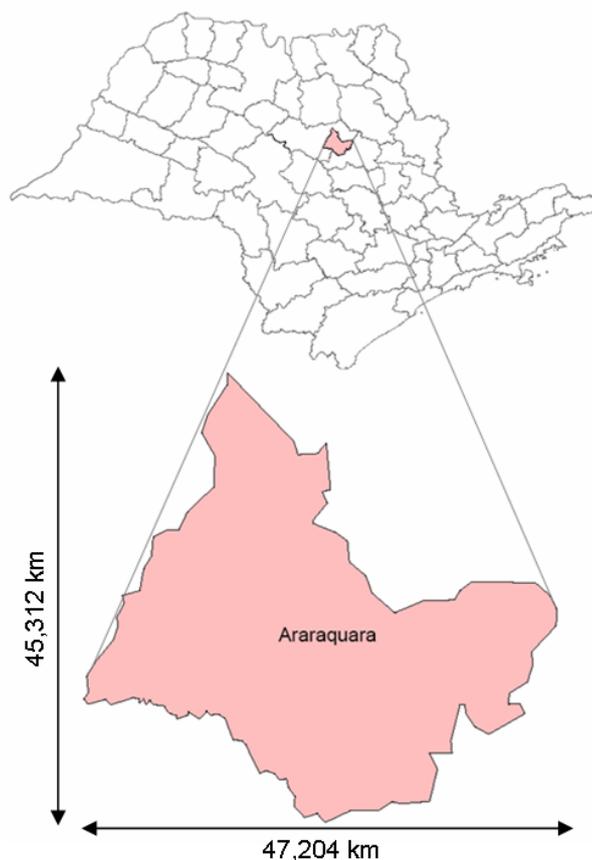


Figura 4.1: *Localização do município de Araraquara no Estado de São Paulo (sem escala).*

de 17,4% em 1970 para 8,8% em 1991. Em 1998 o município contava com 3 escolas municipais e 33 estaduais, ambas de ensino fundamental, e 5 escolas estaduais de ensino médio.

4.2 Materiais

Os materiais utilizados no decorrer do estudo estão listados a seguir:

- Base de dados do Censo demográfico IBGE 2000 – Resultados do universo: Agregados de setores censitários, referente ao município de Araraquara (IBGE 2003). Esta base contém 3200 variáveis relativas aos 130 setores e também as delimitações dos setores censitários;
- Dados espaciais fornecidos pela prefeitura: malha urbana, localização, nome das escolas e limites de bairros;

- Base de dados das escolas públicas do ano 2000, fornecida pela Secretaria de Estado da Educação de São Paulo (SEESP 2007), com 42 variáveis relacionadas ao número de matrículas, número de turmas e tipo de ensino;
- Software para análises estatísticas *R*;
- Software SIG *GeoMedia, Intergraph Co*;
- Software de Extração de Padrões *Apriori*.

4.3 Metodologia

Os métodos a serem empregados neste estudo estão sintetizados no fluxograma apresentado na Figura 4.2: entrada → processo → saída, ou seja, Dados → MD → Conhecimento.

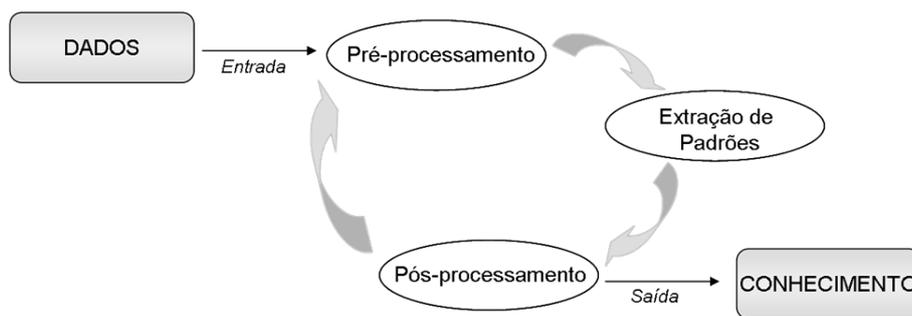


Figura 4.2: Fluxograma do método utilizado.

O Processo, ou seja, a Mineração de Dados, utilizado neste estudo, é o desenvolvido por Rezende, Pugliesi, Melanda, e Paula (2003) do Instituto de Ciências Matemáticas e de Computação – USP São Carlos, e conta com quatro fases: conhecimento do domínio, pré-processamento dos dados, extração de padrões e interpretação dos resultados Figura 4.3, apresentadas a seguir.

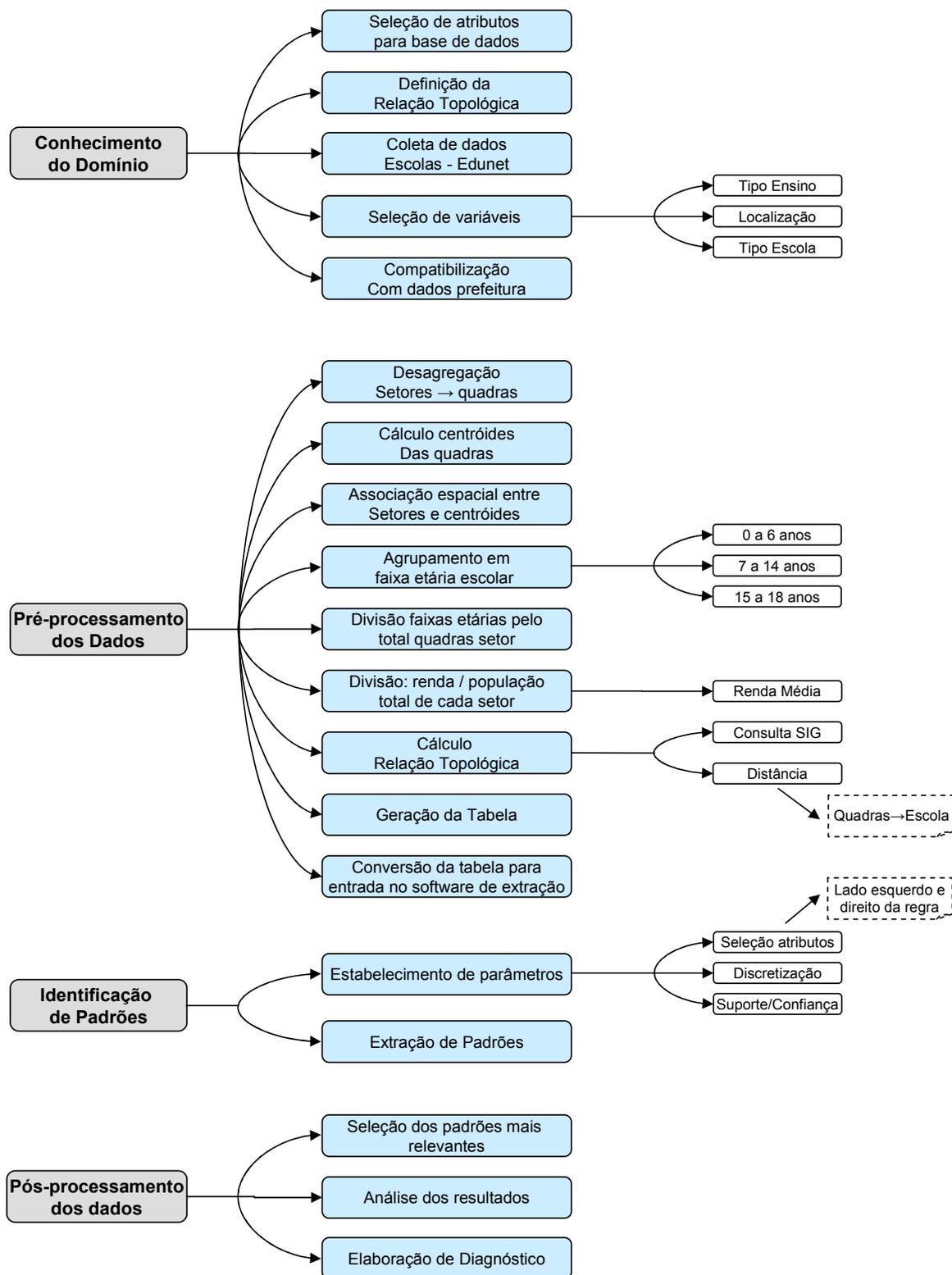


Figura 4.3: Fluxograma metodológico do processo aplicado no estudo.

Conhecimento do domínio

Na fase de conhecimento do domínio, foram verificados os atributos base de dados do IBGE referentes ao município, relevantes para o estudo, e escolhido o conjunto de dados mais adequado à extração de padrões – renda média, população total, e população em idade escolar – para a elaboração de um modelo conceitual da base de dados, para posterior manipulação. Também foram analisados os dados espaciais disponibilizados pela Prefeitura Municipal, para conhecimento do tipo de informação fornecido.

Foi definida como relação topológica a distância euclidiana (em linha reta) entre o centróide de cada quadra e a localização de cada escola. Considerando que podem ocorrer algumas distorções, já que a distância radial é generalizada, ou seja, não corresponde corretamente com a realidade das vias urbanas, caminho percorrido pelas pessoas.

Pré-processamento dos dados

O início da fase de pré-processamento dos dados ocorreu com a desagregação dos dados disponibilizados pelo censo de 2000 do IBGE. Estes são agregados por setores censitários e sua utilização com o referido nível de agregação tende a mascarar os resultados, devido à ocorrência de setores com grandes dimensões e poucas quadras (Tagliatti 2007) e (Tagliatti, Melanda, e Röhm 2007). Considerando que os cálculos de distância são efetuados em relação ao centróide dos objetos espaciais, no caso de setores censitários a distância entre o centróide e a localização das escolas poderia ser muito maior do que entre o centróide das quadras e a escola. Portanto, foi realizada a desagregação dos dados do censo do IBGE de setores censitários para quadras como já realizado em (Tagliatti, Melanda, e Röhm 2007), a fim de tornar a associação entre os dados mais representativa e homogênea, especialmente no que se refere às distâncias.

Na sequência foi realizada uma pesquisa de dados referentes as escolas públicas do município de Araraquara, na Secretaria de Estado da Educação de São Paulo (SEESP), e foram selecionados dados referentes ao ano 2000, pelo fato do último censo do IBGE se referir ao mesmo período. Selecionaram-se as principais variáveis relativas às características das escolas – tipo de ensino, quantidade de alunos por tipo de ensino, número de turmas etc. Estes dados então foram compatibilizados com os dados fornecidos pela Prefeitura, para posterior associação às representações gráficas das escolas.

Com o auxílio do software SIG os dados espaciais foram georreferenciados em um

mesmo Sistema de Projeção Geográfica: South American 1969, zona 23 sul (SAD69-23S).

A delimitação dos setores censitários disponibilizada pelo (IBGE 2003) foi confrontada com os dados selecionados anteriormente, para seleção somente dos objetos contidos dentro dos limites dos setores censitários, para maior eficiência na desagregação dos dados.

Dados espaciais de delimitação das quadras e localização de escolas, disponíveis em formato CAD, foram importados no software SIG. Como o intuito é distribuir a população dos setores nas quadras, não teria sentido manter os polígonos das quadras referentes às áreas verdes e industriais, portanto, estas foram excluídas. Porém, como não foi possível a obtenção de dados de usos do solo do município, surgiu outro fator limitante do estudo: as quadras comerciais, mas este fator não impactou nos resultados.

Na associação dos dados referentes ao número de habitantes por idade, admitiu-se a condição ideal de faixas etárias para cada tipo de ensino, que podem diferir da realidade. Os dados foram agrupados para cada tipo de ensino: infantil (de 0 a 6 anos), fundamental (de 7 a 14 anos) e médio (de 15 a 18 anos). Os dados fornecidos pelo IBGE são referentes ao censo demográfico de 2000, por isso optou-se pela busca de dados escolares do mesmo ano, a fim de se obter maior veracidade nos resultados.

Para iniciar a associação dos dados no SIG, os referentes as faixas etárias escolares, renda média do setor e população total do setor foram associados aos setores censitários (*join de tabela*). Na sequência foram gerados os centróides das quadras, e associados espacialmente (*join espacial*) aos dados dos setores censitários.

Assumiu-se que a distribuição dos dados acontece de maneira igual para cada quadra, e por meio de consultas, os dados de faixa etária e da população total foram distribuídos pelas quadras igualmente. Foi estudada a viabilidade de distribuir por meio de ponderação, mas devido a homogeneidade das quadras dentro de um mesmo setor, esta ação apresentou-se desnecessária. Os dados referentes à renda foram divididos pela população total de cada setor censitário, obtendo-se a renda média per capita. Em cada setor a discretização dos dados ocorreu de maneira semelhante ao realizado por Tagliatti, Melanda, e Pivato (2006).

As representações gráficas das escolas foram associados a tabela com os atributos referentes as escolas selecionados (*join de tabela*). Os dados disponibilizados englobavam universidades, colégios técnicos e secretarias de educação, totalizando 134 pontos, não relevantes para o objetivo da pesquisa, portanto foram excluídos. Algumas esco-

las, devido a ausência de informações, ou nos dados da Prefeitura, ou nos dados da SEESP, foram excluídas, restando 65 escolas estaduais e municipais de ensinos infantil, fundamental e médio.

A determinação da relação topológica *distância* também foi realizada no software SIG, obtendo-se uma tabela com a distância euclidiana de cada quadra para cada escola. Esta foi associada à tabela dos centróides das quadras (*join de tabela*).

O relacionamento de dados das escolas foi associado espacialmente aos centróides das quadras (*join espacial*), originando uma tabela com todos os dados necessários para a Mineração de Dados, onde cada linha desta representa uma distância. A Figura 4.4 apresenta a sequência de associações no banco de dados para composição da tabela final, conforme descrito anteriormente.

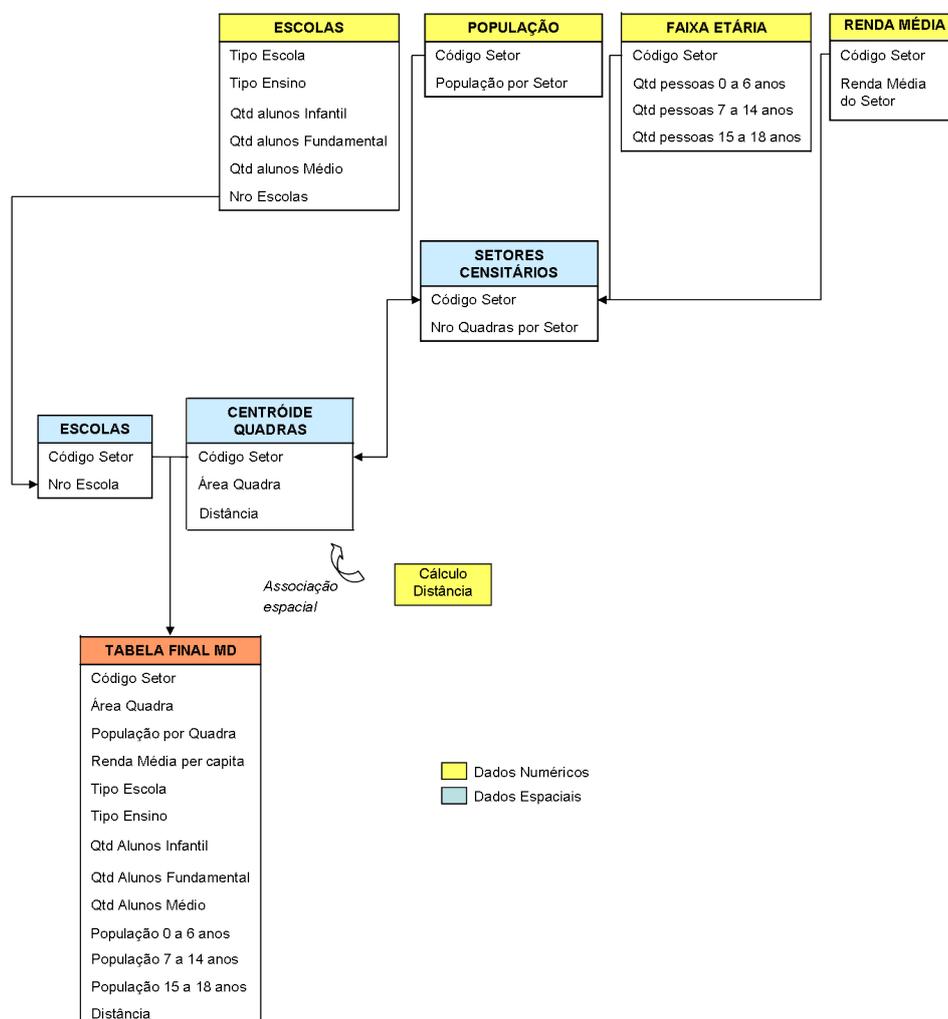


Figura 4.4: Associações realizadas no banco de dados durante o pré-processamento dos dados.

A tabela final gerada foi analisada e teve alguns de seus ítems discretizados, a fim de facilitar a interpretação dos padrões quando gerados. A discretização foi realizada com o auxílio do software *R* de análises estatísticas. Cada atributo a ser discretizado foi analisado pela frequência de seus dados antes da definição dos intervalos de dados. O atributo *distância* foi discretizado relacionando-se com a faixa de idade escolar das pessoas, já o atributo *renda per capita* foi classificado de acordo com os critérios da ABEP (2003).

A tabela final, com os dados discretizados foi exportada para o formato texto, compatível com o formato de entrada no software de extração de padrões *Apriori*.

Identificação de padrões

Para a extração de padrões foi necessária a seleção dos atributos a serem analisados, definição da posição do atributo em relação ao lado da regra ($LHS \rightarrow RHS$), como tipo de ensino, idade escolar e renda do lado esquerdo da regra e a distância do lado direito, por exemplo. Para melhor ilustrar a seguir é apresentado um exemplo de RAE de Tagliatti (2007):

```
Se RendaMédia=alta e PessoasEmIdadeFundamental=médio  
então Proximidade=longe [Confiança=84,2%; Suporte=31,0%]
```

Interpretação: Considerando as quadras com uma quantidade média de pessoas em idade de ensino fundamental, e com renda média per capita superior ou igual a 300 reais, em 31,0% dos casos, estas estão localizadas a uma distância superior ou igual a 1000 metros de uma escola. A partir desta regra, observa-se também que a relação **RendaMédia=alta e PessoasEmIdadeFundamental=médio** ocorre em 84,2% dos casos.

Parâmetros como Suporte e Confiança, número mínimo e máximo de atributos por regra também foram definidos no início desta fase, para a realização da extração de padrões.

Na sequência, os padrões extraídos passaram por análise e refinamento com o apoio do ARINE (*Association Rules Interactive Exploration*), ambiente para identificação de regras de associação interessantes, desenvolvido por Melanda (2004), os mais relevantes selecionados para interpretação.

As medidas utilizadas na filtragem dos padrões, no ambiente ARINE foram *Suporte*, *Confiança* e *Novidade*. A medida *Suporte*, já apresentada anteriormente, indica a ocorrência relativa da regra de associação detectada dentro do conjunto de dados de transações. A medida *Confiança*, também já apresentada, indica a frequência com que os dois lados da regra ocorrem juntos, em relação ao número total de ocorrências de determinada transação do lado direito da regra. Já a medida *Novidade* objetiva identificar quão inovadora, interessante ou não-usual é uma dada regra (Melanda 2004), ou seja, quanto maior o valor da medida *Novidade*, mais interessante deve ser a regra e, de acordo com (Lavrac, Flach, e Zupan 1999), pode ser demonstrado que esta medida varia de -0,25 a +0,25.

Pós-processamento dos dados

Na fase de pós-processamento dos dados, foi realizada uma análise comparativa do grau de inovação das regras, e a comparação das interpretações das regras com os dados apresentados nos mapas temáticos gerados, tais como distribuição de escolas por porte e por tipo de ensino, densidade demográfica, distribuição de pessoas em idade escolar, e renda média per capita.

Os mapas temáticos resultaram de sucessivas associações de dados alfa-numéricos do (IBGE 2003) a objetos espaciais no software SIG e desagregação de setores censitários IBGE (2003) para quadras. Algumas ferramentas de cálculos e análises estatísticas foram utilizadas na elaboração dos mesmos.

Resultados e discussão

Este capítulo apresenta os resultados obtidos e discussões em cada etapa do estudo e algumas conclusões.

5.1 *Pré-processamento dos dados*

A tabela de dados de entrada no software de extração de padrões Apriori, gerada ao final do pré-processamento, contém as distâncias entre o centróide de cada quadra e cada escola calculadas, dados referentes as quadras, como código do setor censitário, quantidade de pessoas nas faixas de idade escolar (0 a 6 anos, 7 a 14 anos e 15 a 18 anos), e renda média per capita, e dados referentes as escolas, como tipo de escola (municipal ou estadual), tipo de ensino (infantil, fundamental e médio), e quantidade de alunos em cada tipo de ensino.

Foram criados 3 intervalos para o atributo *distância* de acordo com as faixas de idade e tipo de ensino, ou seja, uma criança até 6 anos de idade, do ensino infantil, caminha por até 500 metros. Já uma criança com idade entre 7 e 14 anos caminha de 500 a 1000 metros, e pessoas na faixa etária de 15 a 18 anos conseguem caminhar mais do que 1000 metros. Então os intervalos definidos para o atributo *distância* foram respectivamente: *proxima*, *media* e *longe*.

Definiram-se classes para o atributo *renda média per capita*, com base nos critérios adotados pela Associação Brasileira de Empresas de Pesquisa (ABEP 2003), para dados de 2000, a partir da quantidade de itens como televisão, rádio, automóvel, geladeira e banheiros, existentes nas residências, além do grau de instrução do chefe de família. Foi adotada a classificação de renda familiar para a Região Metropolitana de Araraquara, e considerado uma média de 3 habitantes por família, para obtenção de intervalos de rendas médias mensais per capita no município. As faixas de renda definidas foram:

- A1 = acima de 1.549,01 reais
- A2 = de 934,01 a 1.549 reais
- B1 = de 556,01 a 934 reais
- B2 = de 309,01 a 556 reais
- C = de 141,01 a 309 reais
- D = de 69,01 a 141 reais
- E = de 0,1 a 69 reais
- 0 = zero reais

Os atributos analisados no software de análises estatísticas *R* foram os referentes à quantidade de alunos matriculados nos ensinos infantil, fundamental e médio, à quantidade de pessoas com idades entre 0 e 6 anos, 7 e 14 anos e 15 e 18 anos.

Os intervalos definidos para o atributo *infantil* – alunos matriculados no ensino infantil – cujas análises estatísticas são apresentadas na Tabela 5.1 foram:

- zero alunos
- de 34 a 298 alunos
- de 299 a 397 alunos

Infantil		
Frequencia	Percentual	Percentual Acumulado
0,00	59,42	59,42
34,00	1,06	60,48
119,00	0,80	61,27
161,00	1,19	62,47
206,00	1,33	63,79
233,00	2,25	66,05
248,00	1,72	67,77
250,00	3,71	71,49
252,00	0,66	72,15
269,00	2,65	74,80
280,00	2,25	77,06
291,00	1,59	78,65
297,00	0,13	78,78
298,00	1,06	79,84
299,00	1,06	80,90
304,00	1,99	82,89
309,00	2,39	85,28
320,00	4,24	89,52
323,00	2,12	91,64
329,00	4,11	95,76
341,00	0,13	95,89
347,00	1,59	97,48
375,00	0,27	97,75
392,00	0,80	98,54
397,00	1,46	100,00
Total	100,00	

Tabela 5.1: Tabela de análises estatísticas para o atributo infantil.

Os intervalos definidos para o atributo *fundamental* – alunos matriculados no ensino fundamental – cujas análises estatísticas são apresentadas na Tabela 5.2 foram:

- zero alunos
- de 50 a 553 alunos
- de 581 a 1117 alunos

Fundamental		
Frequencia	Percentual	Percentual Acumulado
0,00	42,44	42,44
50,00	1,06	43,50
73,00	0,66	44,16
76,00	2,25	46,42
208,00	1,06	47,48
272,00	0,27	47,75
323,00	1,33	49,07
331,00	1,72	50,80
334,00	1,33	52,12
345,00	2,79	54,91
354,00	3,71	58,62
406,00	2,25	60,88
409,00	2,12	63,00
424,00	1,46	64,46
493,00	2,25	66,71
529,00	1,46	68,17
553,00	2,25	70,42
581,00	1,33	71,75
649,00	1,06	72,81
674,00	0,80	73,61
706,00	0,93	74,54
721,00	3,32	77,85
722,00	0,80	78,65
726,00	1,19	79,84
736,00	1,59	81,43
767,00	2,39	83,82
775,00	1,06	84,88
790,00	3,85	88,73
792,00	1,19	89,92
865,00	1,19	91,11
903,00	2,52	93,63
931,00	2,12	95,76
1.026,00	2,25	98,01
1.117,00	1,99	100,00
Total	100,00	

Tabela 5.2: Tabela de análises estatísticas para o atributo fundamental.

Os intervalos definidos para o atributo *médio* – alunos matriculados no ensino médio – cujas análises estatísticas são apresentadas na Tabela 5.3 foram:

- zero alunos
- de 186 a 1339 alunos
- de 1638 a 2806 alunos

Médio		
Frequencia	Percentual	Percentual Acumulado
0,00	91,38	91,38
186,00	1,46	92,84
568,00	2,12	94,96
1.339,00	1,06	96,02
1.638,00	1,72	97,75
2.806,00	2,25	100,00
Total	100,00	

Tabela 5.3: Tabela de análises estatísticas para o atributo medio.

Os intervalos definidos para o atributo *0a6 anos* – pessoas com idades entre 0 e 6 anos por quadra – cujas análises estatísticas são apresentadas na Tabela 5.4 foram:

- zero pessoas
- de 0,03 a 0,10 pessoas
- de 0,11 a 0,24 pessoas
- de 0,25 a 4,00 pessoas

Os intervalos definidos para o atributo *7a14anos* – pessoas com idades entre 7 e 14 anos por quadra – cujas análises estatísticas são apresentadas na Tabela 5.5 foram:

- zero pessoas
- de 0,50 a 4,47 pessoas
- de 4,50 a 6,56 pessoas
- de 6,67 a 73,00 pessoas

0 a 6 anos					
Frequencia	Percentual	Percentual Acumulado	Frequencia	Percentual	Percentual Acumulado
0,00	7,03	7,03	0,24	4,51	66,05
0,03	3,71	10,74	0,24	2,79	68,83
0,07	1,99	12,73	0,25	4,24	73,08
0,08	1,72	14,46	0,28	2,79	75,86
0,09	1,46	15,92	0,29	2,25	78,12
0,10	4,11	20,03	0,31	1,72	79,84
0,10	2,65	22,68	0,33	3,18	83,02
0,11	4,64	27,32	0,40	3,98	87,00
0,12	2,12	29,44	0,42	1,59	88,59
0,13	1,99	31,43	0,44	1,19	89,79
0,14	3,85	35,28	0,50	2,12	91,91
0,14	0,93	36,21	0,55	1,46	93,37
0,15	2,65	38,86	0,56	1,19	94,56
0,15	1,72	40,58	0,74	2,52	97,08
0,17	1,59	42,18	0,91	1,46	98,54
0,18	6,76	48,94	1,00	0,27	98,81
0,18	3,71	52,65	1,37	1,06	99,87
0,19	2,12	54,77	4,00	0,13	100,00
0,20	1,99	56,76	Total	100	
0,22	4,77	61,54			

Tabela 5.4: Tabela de análises estatísticas para o atributo 0a6anos.

Os intervalos definidos para o atributo *15a18anos* – pessoas com idades entre 15 e 18 anos por quadra – cujas análises estatísticas são apresentadas na Tabela 5.6 foram:

- zero pessoas
- de 1,00 a 2,44 pessoas
- de 2,60 a 3,81 pessoas
- de 3,82 a 42,5 pessoas

7 a 14 anos					
Frequencia	Percentual	Percentual Acumulado	Frequencia	Percentual	Percentual Acumulado
0,00	0,13	0,13	6,53	2,25	63,40
0,50	0,80	0,93	6,54	1,72	65,12
1,40	1,33	2,25	6,56	1,19	66,31
2,13	1,06	3,32	6,67	1,59	67,90
2,18	3,71	7,03	6,70	1,33	69,23
2,95	2,79	9,81	6,75	2,65	71,88
3,00	3,71	13,53	6,81	2,12	74,01
3,11	1,19	14,72	7,06	2,25	76,26
3,36	3,32	18,04	7,18	1,46	77,72
3,47	2,25	20,29	7,50	2,12	79,84
3,67	2,79	23,08	7,62	1,72	81,56
4,11	1,19	24,27	7,89	2,25	83,82
4,37	1,06	25,33	8,00	1,46	85,28
4,38	3,85	29,18	8,54	1,72	87,00
4,47	4,51	33,69	10,10	1,33	88,33
4,50	0,80	34,48	11,25	1,06	89,39
4,64	1,46	35,94	11,33	0,80	90,19
4,67	1,99	37,93	12,00	0,80	90,98
4,81	2,12	40,05	12,78	3,58	94,56
5,00	1,06	41,11	13,33	1,19	95,76
5,13	4,11	45,23	17,00	1,59	97,35
5,42	2,52	47,75	17,20	1,33	98,67
5,56	1,19	48,94	21,20	0,66	99,34
5,75	1,59	50,53	32,00	0,27	99,60
6,00	4,91	55,44	51,00	0,13	99,73
6,18	1,46	56,90	73,00	0,27	100,00
6,33	1,99	58,89	Total	100,00	
6,35	2,25	61,14			

Tabela 5.5: Tabela de análises estatísticas para o atributo 7a14anos.

15 a 18 anos					
Frequencia	Percentual	Percentual Acumulado	Frequencia	Percentual	Percentual Acumulado
0,00	0,13	0,13	3,82	4,51	65,52
1,00	1,06	1,19	4,08	1,72	67,24
1,09	2,79	3,98	4,22	1,19	68,44
1,67	0,80	4,77	4,31	2,12	70,56
1,68	3,32	8,09	4,39	2,25	72,81
1,78	3,71	11,80	4,42	1,59	74,40
1,89	3,85	15,65	4,64	1,46	75,86
1,90	1,33	16,98	4,67	1,59	77,45
2,25	3,71	20,69	4,71	0,93	78,38
2,33	0,80	21,49	4,82	1,46	79,84
2,35	2,25	23,74	4,82	2,25	82,10
2,37	1,06	24,80	5,00	1,33	83,42
2,41	2,25	27,06	5,44	1,19	84,62
2,44	1,19	28,25	5,92	1,59	86,21
2,58	4,11	32,36	6,08	1,72	87,93
2,80	1,99	34,35	6,28	2,39	90,32
2,80	2,79	37,14	6,40	0,66	90,98
2,91	1,46	38,59	7,20	1,33	92,31
2,94	2,12	40,72	7,50	1,33	93,63
3,15	1,72	42,44	7,67	1,19	94,83
3,25	1,06	43,50	8,00	1,19	96,02
3,33	1,99	45,49	8,12	1,06	97,08
3,37	1,06	46,55	8,33	0,80	97,88
3,53	2,25	48,81	10,33	0,80	98,67
3,59	2,25	51,06	16,80	0,66	99,34
3,62	2,12	53,18	19,50	0,27	99,60
3,63	2,52	55,70	20,00	0,13	99,73
3,75	2,65	58,36	42,50	0,27	100,00
3,78	1,19	59,55	Total	100,00	
3,81	1,46	61,01			

Tabela 5.6: Tabela de análises estatísticas para o atributo 15a18anos.

A Tabela 5.7 apresenta uma amostra desta tabela, pelo fato de conter 49.764 linhas, referentes às distâncias entre o centróide das quadras e as escolas.

Foram gerados alguns mapas temáticos com o intuito de facilitar a visualização da relação dos dados trabalhados com o espaço urbano. A seguir serão descritas as manipulações dos dados para a concepção de cada mapa:

- Distribuição de escolas por tipo de ensino – os dados de quantidade de alunos para cada tipo de ensino foram destacados por cores, ou seja, os referentes ao tipo de ensino *infantil*, *fundamental* e *médio* (Figura 5.1). Duas escolas foram descartadas por não se localizarem dentro dos setores censitários, e portanto, não conterem dados do IBGE para associação.

setor	distância	QUADRAS				renda per capita	ESCOLA				
		até 6 anos	de 7 a 14 anos	de 15 a 18 anos			escola	ensino	infantil	fundamental	médio
5000171	proxima	0,25A4	6,7A73	3,82A42,5	C	municipal	infantil	34A298	0	0	
5000171	proxima	0,25A4	6,7A73	3,82A42,5	C	municipal	infantil	34A298	0	0	
5000171	proxima	0,25A4	6,7A73	3,82A42,5	C	municipal	infantil	34A298	0	0	
5000171	proxima	0,25A4	6,7A73	3,82A42,5	C	municipal	infantil	34A298	0	0	
5000171	proxima	0,25A4	6,7A73	3,82A42,5	C	municipal	infantil	34A298	0	0	
5000171	proxima	0,25A4	6,7A73	3,82A42,5	C	municipal	infantil	34A298	0	0	
5000171	proxima	0,25A4	6,7A73	3,82A42,5	C	municipal	infantil	34A298	0	0	
5000171	proxima	0,25A4	6,7A73	3,82A42,5	C	municipal	infantil	34A298	0	0	
25000036	longe	0,25A4	6,7A73	3,82A42,5	B2	estadual	fundamental	0	581A1117	0	
25000036	longe	0,25A4	6,7A73	3,82A42,5	B2	estadual	fundamental	0	581A1117	0	
25000036	longe	0,25A4	6,7A73	3,82A42,5	B2	estadual	fundamental	0	581A1117	0	
25000036	longe	0,25A4	6,7A73	3,82A42,5	B2	estadual	fundamental	0	581A1117	0	
25000036	longe	0,25A4	6,7A73	3,82A42,5	B2	estadual	fundamental	0	581A1117	0	
25000036	longe	0,25A4	6,7A73	3,82A42,5	B2	estadual	fundamental	0	581A1117	0	
25000036	longe	0,25A4	6,7A73	3,82A42,5	B2	estadual	fundamental	0	581A1117	0	
25000036	longe	0,25A4	6,7A73	3,82A42,5	B2	estadual	fundamental	0	581A1117	0	
25000036	longe	0,25A4	6,7A73	3,82A42,5	B2	estadual	fundamental	0	581A1117	0	
25000036	longe	0,25A4	6,7A73	3,82A42,5	B2	estadual	fundamental	0	581A1117	0	

Tabela 5.7: Trecho da tabela final gerada para entrada no software de extração de padrões.

- Distribuição de escolas por porte – os dados referentes a quantidade de alunos nos ensinos infantil, fundamental e médio foram somados, compondo o atributo referente a quantidade total de alunos. Foram definidos três intervalos referentes a quantidade pequena, média e grande de alunos, respectivamente até 500 alunos, de 500 a 1000 alunos, e acima de 1000 alunos (Figura 5.2).
- Densidade Demográfica por quadra – os dados de população total do IBGE foram discretizados dividindo-se pelo total de quadras de cada setor, e distribuídos pelas áreas das quadras (em hectares). Criaram-se três intervalos de Densidade - baixa, média e alta - correspondendo respectivamente aos intervalos - até 50 hab/ha, de 50 a 150 hab/ha, e acima de 150 hab/ha (Figura 5.3).
- Distribuição de pessoas em idade escolar – apresentação gráfica da distribuição de pessoas em idade escolar para os ensinos *infantil*, *fundamental* e *médio*, ou seja, de 0 a 6 anos, de 7 a 14 anos e de 15 a 18 anos, em intervalos definidos de acordo com a frequência (Figuras 5.4, 5.5 e 5.6).
- Distribuição de renda per capita – a renda média de cada setor censitário foi distribuída pela população referente ao mesmo. Criaram-se então, intervalos de renda per capita, de acordo com os critérios da ABEP (2003), descritos a seguir, cuja distribuição é apresentada na Figura 5.7.

Comparando-se a Distribuição de Escolas por Porte (Figura 5.2) com os Tipos de Ensino das mesmas (Figura 5.1), com a Distribuição de Pessoas em Idade de Ensino Infantil (Figura 5.4), observa-se que as escolas de pequeno porte (até 500 alunos

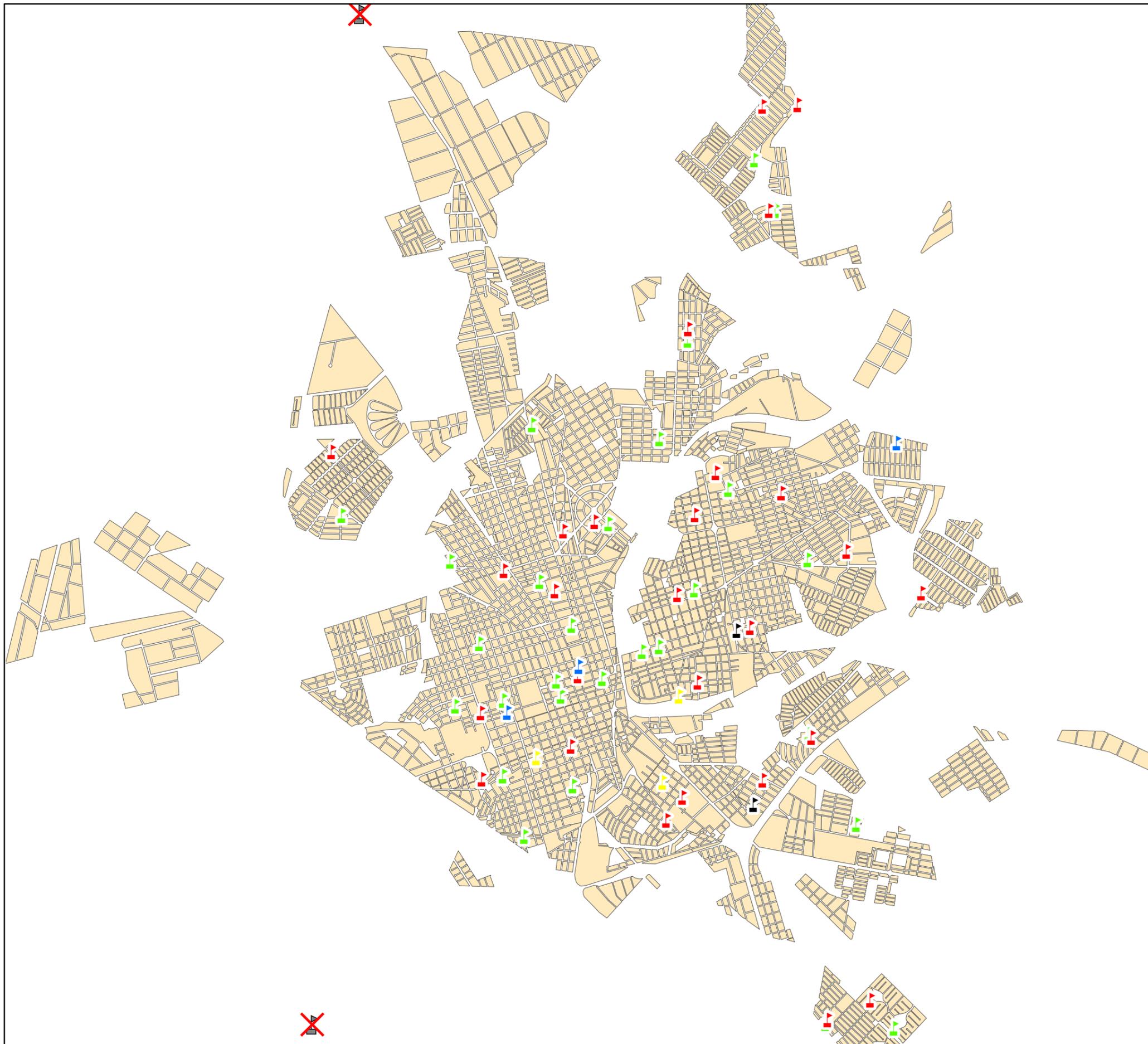
matriculados) são, em sua maioria, de ensino Infantil, portanto Municipais. Estas localizam-se em regiões que apresentam até 3 pessoas com idade de ensino infantil por quadra, ou seja, poucas crianças, podendo-se inferir que as escolas infantis, de pequeno porte, atendem a demanda de crianças com esta idade de ensino (até 6 anos).

Já comparando-se as Figuras 5.2 e 5.1, com a Distribuição de Pessoas em Idade de Ensino Fundamental (Figura 5.5), notas-se que quase todas as escolas de médio porte (de 500 a 100 alunos matriculados) são de ensino Fundamental, e localizam-se, em sua maioria, em áreas cujas quadras apresentam até 10 habitantes compatíveis com esta idade de ensino, apresentando um bom atendimento a demanda de alunos com idades entre 7 e 14 anos.

Relacionando-se as mesmas figuras com a Distribuição de Pessoas em Idade de Ensino Médio (Figura 5.6), pode-se dizer que existem poucas escolas de ensino Médio, porém estas são consideradas de grande porte (acima de 1000 alunos matriculados). Na sua maioria estas estão localizadas em regiões com até 5 pessoas com idade de ensino médio por quadra, não apresentando um atendimento a toda a população, devido a pouca quantidade de escolas e concentração destas na parte sul do município.

A Densidade Demográfica do município (Figura 5.3), analisada juntamente com a Distribuição de Renda Média Per Capita (Figura 5.7), indicou que nas áreas mais periféricas, onde existem até 50 hab/ha, e pouca concentração de escolas públicas, as classes de renda variam entre B2, C, D-E, ou seja, de 0,1 a 556 reais por habitante. Já na região central do município as classes de renda variam entre B1, B2 e C (de 141,01 a 934 reais), região com densidade demográfica de 50 a 150 hab/ha, com algumas quadras sortidas apresentando densidade acima de 150 hab/ha. Poucas quadras, e na região central, apresentaram rendas per capita elevadas (A1 e A2 – acima de 934,01 reais), estas contam com densidade demográfica variando de 50 a 150 hab/ha.

As áreas não cobertas pela rede escolar (mais periféricas em sua maioria) possuem densidade demográfica de 50 hab/ha, com renda per capita variando de 0,1 a 556 reais. Nessas áreas existem habitantes em idade escolar de ensino Fundamental e Médio, e as escolas mais próximas são de ensinos Infantil e Fundamental.



Universidade Federal de São Carlos
 Departamento de Engenharia Civil
 Programa de Pós-graduação em Engenharia Urbana

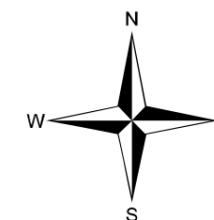
- Dissertação de Mestrado -
 Verificação das relações espaciais
 entre a distribuição de dados sócio-econômicos
 e a localização de escolas

DISTRIBUIÇÃO DE ESCOLAS POR TIPO DE ENSINO - ARARAQUARA/SP -

Figura 5.1

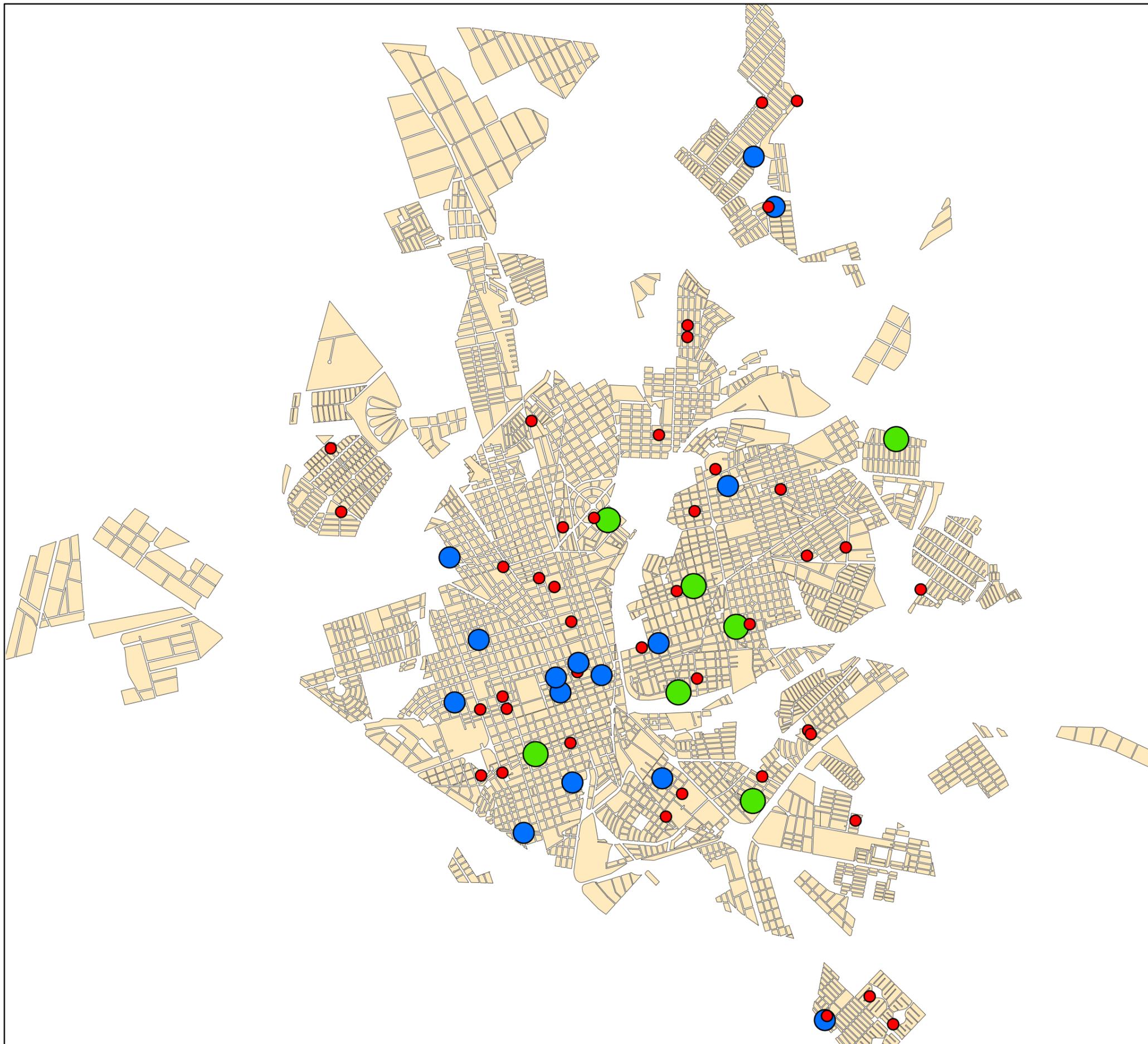
Legenda

-  infantil
-  fundamental
-  infantil-fundamental
-  fundamental-medio
-  medio
-  escolas descartadas



Datum: SAD 1969 - UTM - ZONA 23S
 Escala - 1:50.000





Universidade Federal de São Carlos
 Departamento de Engenharia Civil
 Programa de Pós-graduação em Engenharia Urbana

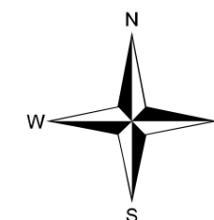
- Dissertação de Mestrado -
 Verificação das relações espaciais
 entre a distribuição de dados sócio-econômicos
 e a localização de escolas

**DISTRIBUIÇÃO DE ESCOLAS
 POR PORTE (QUANTIDADE
 DE ALUNOS)
 - ARARAQUARA/SP -**

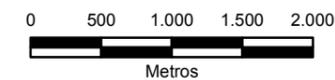
Figura 5.2

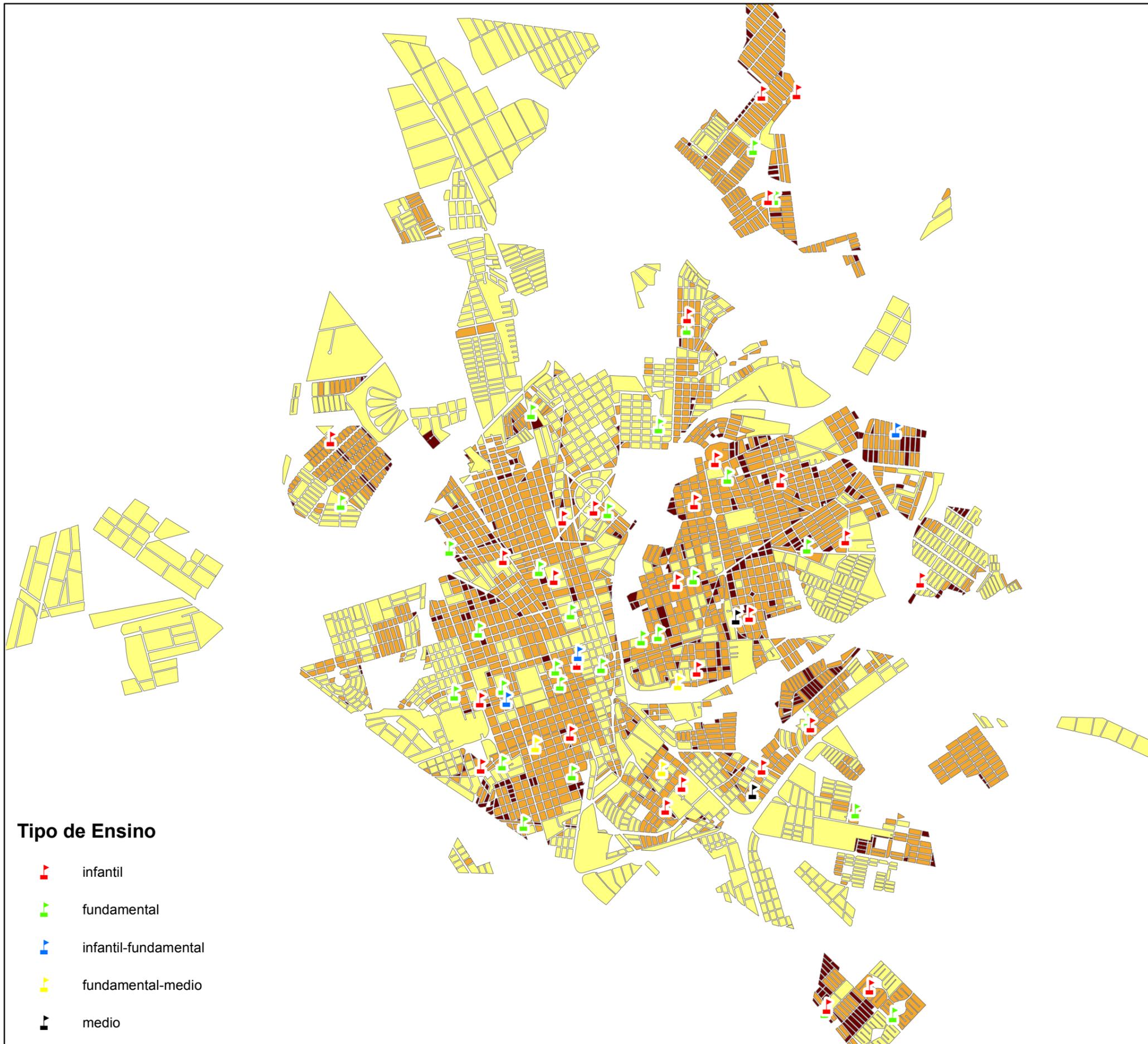
Legenda

- pequena: até 500 alunos
- média: de 500 a 1000 alunos
- grande: acima de 1000 alunos



Datum: SAD 1969 - UTM - ZONA 23S
 Escala - 1:50.000





Tipo de Ensino

-  infantil
-  fundamental
-  infantil-fundamental
-  fundamental-medio
-  medio



Universidade Federal de São Carlos
 Departamento de Engenharia Civil
 Programa de Pós-graduação em Engenharia Urbana

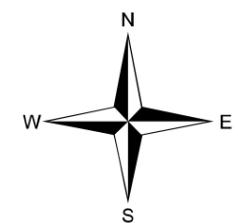
- Dissertação de Mestrado -
 Verificação das relações espaciais
 entre a distribuição de dados sócio-econômicos
 e a localização de escolas

**DENSIDADE DEMOGRÁFICA
 POR QUADRA
 - ARARAQUARA/SP -**

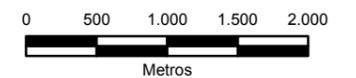
Figura 5.3

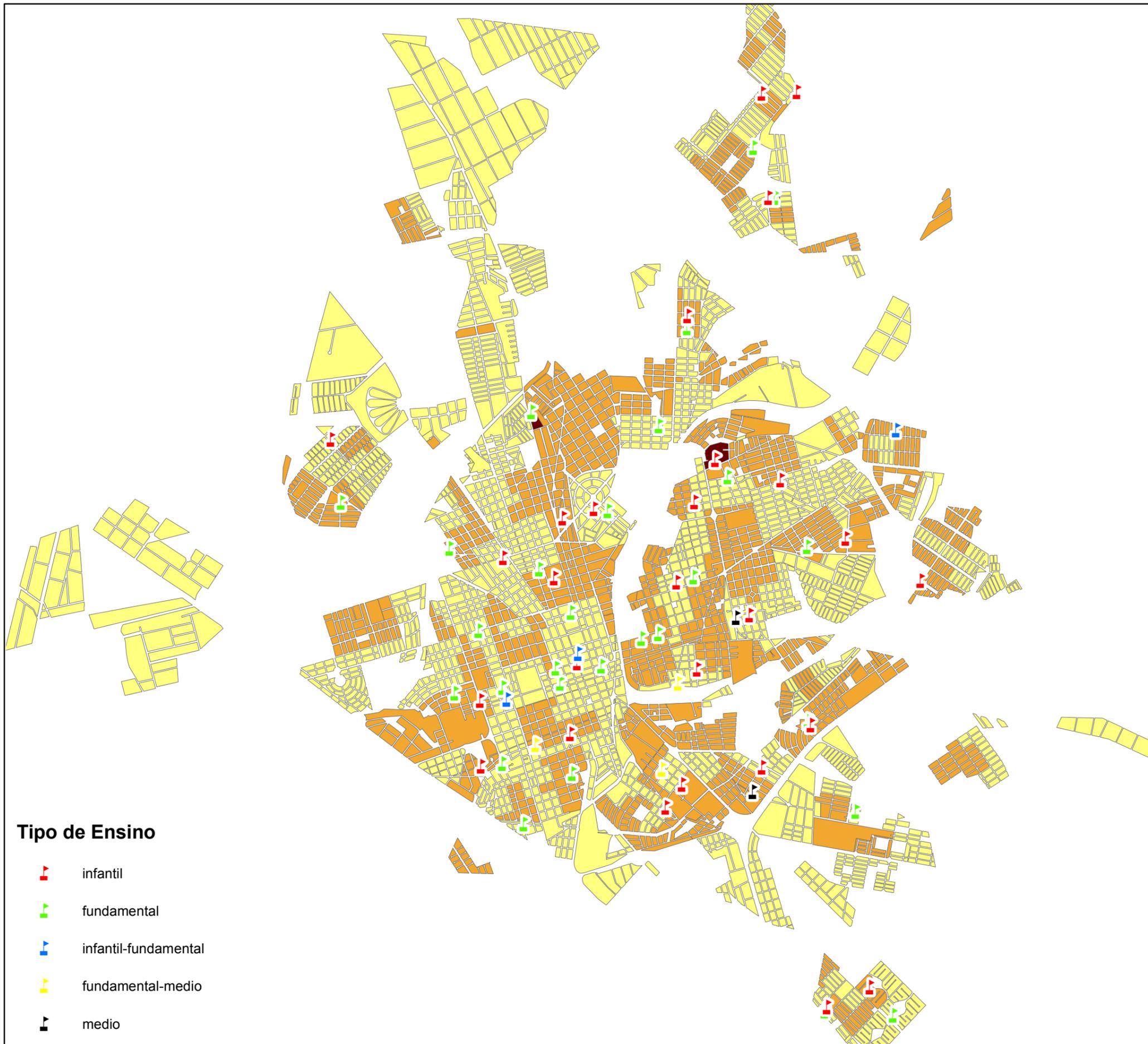
Legenda

-  até 50 hab/ha
-  de 50 a 150 hab/ha
-  acima de 150 hab/ha



Datum: SAD 1969 - UTM - ZONA 23S
 Escala - 1:50.000





Tipo de Ensino

-  infantil
-  fundamental
-  infantil-fundamental
-  fundamental-medio
-  medio



Universidade Federal de São Carlos
 Departamento de Engenharia Civil
 Programa de Pós-graduação em Engenharia Urbana

- Dissertação de Mestrado -
 Verificação das relações espaciais
 entre a distribuição de dados sócio-econômicos
 e a localização de escolas

**DISTRIBUIÇÃO DE PESSOAS
 EM IDADE ESCOLAR
 DE ENSINO INFANTIL
 - ARARAQUARA/SP -**

Figura 5.4

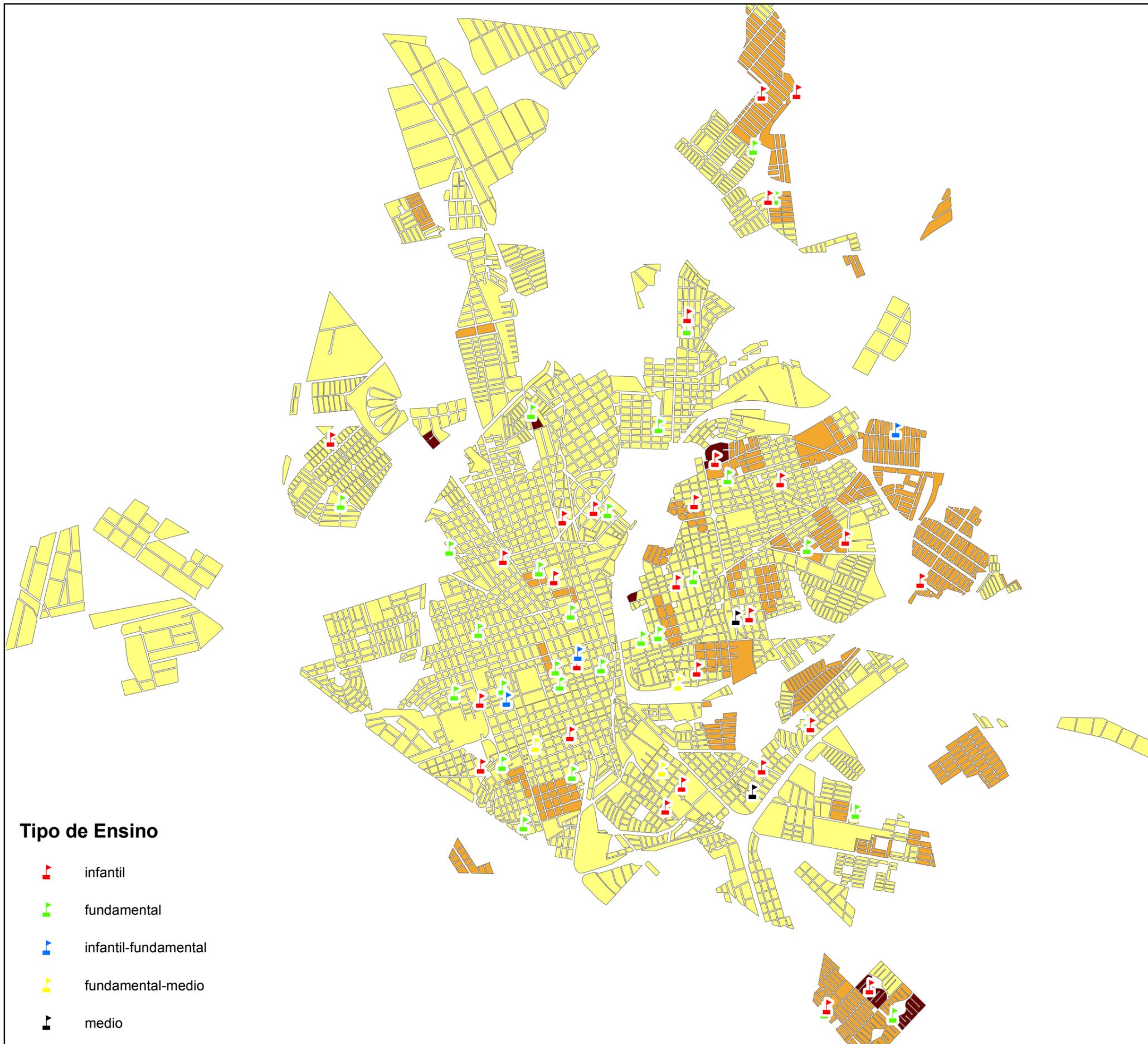
Legenda

-  nenhuma pessoa por quadra
-  até 3 pessoas por quadra
-  acima de 3 pessoas por quadra



Datum: SAD 1969 - UTM - ZONA 23S
 Escala - 1:50.000





Tipo de Ensino

-  infantil
-  fundamental
-  infantil-fundamental
-  fundamental-medio
-  medio



Universidade Federal de São Carlos
 Departamento de Engenharia Civil
 Programa de Pós-graduação em Engenharia Urbana

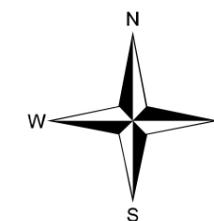
- Dissertação de Mestrado -
 Verificação das relações espaciais
 entre a distribuição de dados sócio-econômicos
 e a localização de escolas

**DISTRIBUIÇÃO DE PESSOAS
 EM IDADE ESCOLAR
 DE ENSINO FUNDAMENTAL
 - ARARAQUARA/SP -**

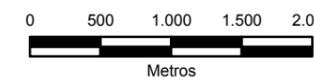
Figura 5.5

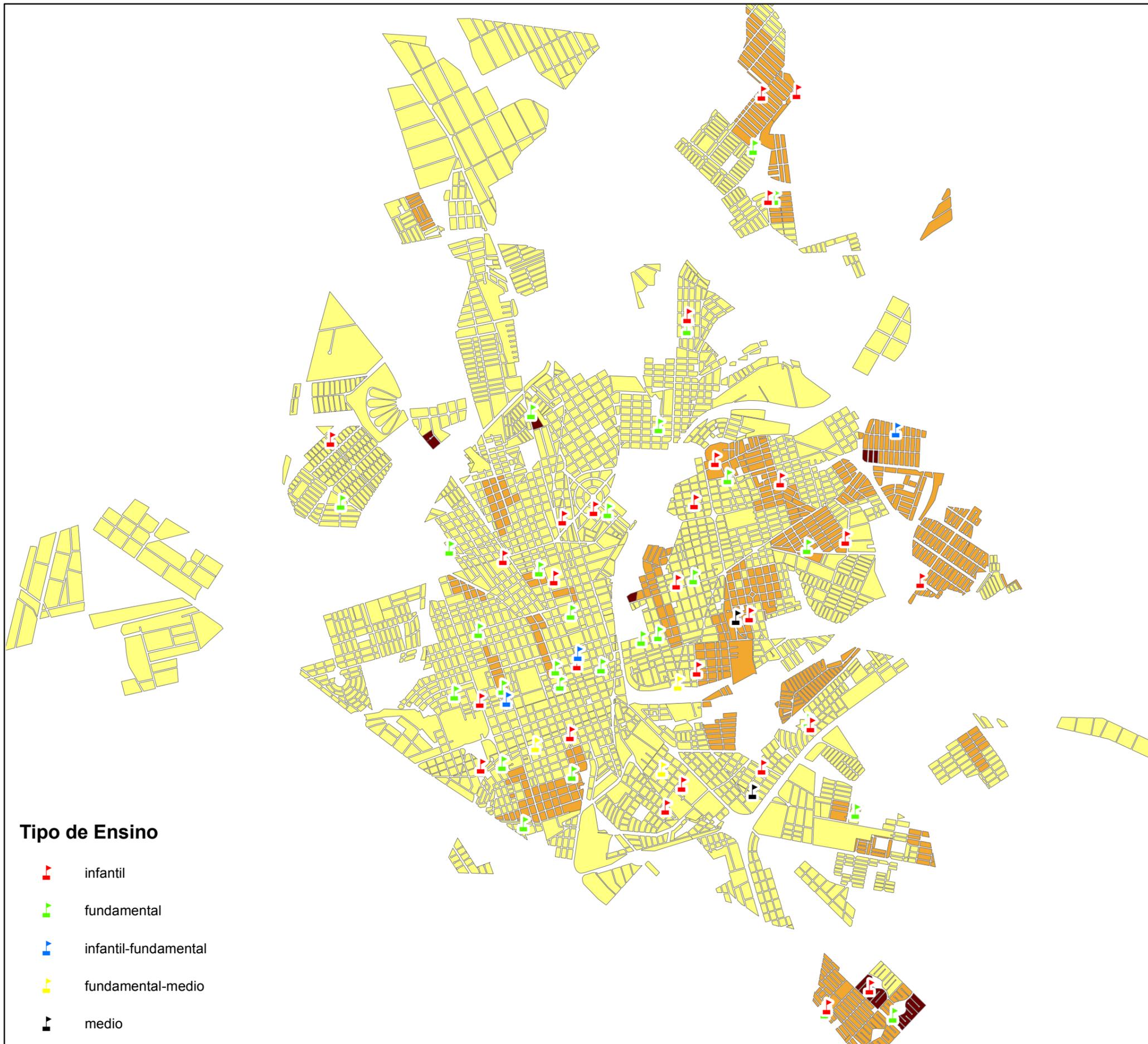
Legenda

-  até 10 pessoas por quadra
-  de 10 a 50 pessoas por quadra
-  acima de 50 pessoas por quadra



Datum: SAD 1969 - UTM - ZONA 23S
 Escala - 1:50.000





Tipo de Ensino

-  infantil
-  fundamental
-  infantil-fundamental
-  fundamental-medio
-  medio



Universidade Federal de São Carlos
 Departamento de Engenharia Civil
 Programa de Pós-graduação em Engenharia Urbana

- Dissertação de Mestrado -
 Verificação das relações espaciais
 entre a distribuição de dados sócio-econômicos
 e a localização de escolas

**DISTRIBUIÇÃO DE PESSOAS
 EM IDADE ESCOLAR
 DE ENSINO MÉDIO
 - ARARAQUARA/SP -**

Figura 5.6

Legenda

-  até 5 pessoas por quadra
-  de 5 a 20 pessoas por quadra
-  acima de 20 pessoas por quadra



Datum: SAD 1969 - UTM - ZONA 23S
 Escala - 1:50.000

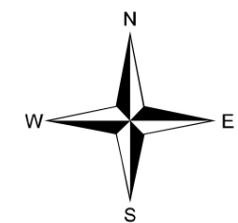


**DISTRIBUIÇÃO DE RENDA
 MÉDIA PER CAPITA
 POR QUADRAS
 - ARARAQUARA/SP -**

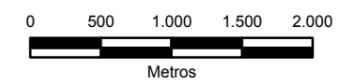
Figura 5.7

Legenda

-  A1 - acima de 1549,01
-  A2 - de 934,01 a 1549,00
-  B1 - de 556,01 a 934,00
-  B2 - de 309,01 a 556,00
-  C - de 141,01 a 309,00
-  D-E - de 0,01 a 141,00

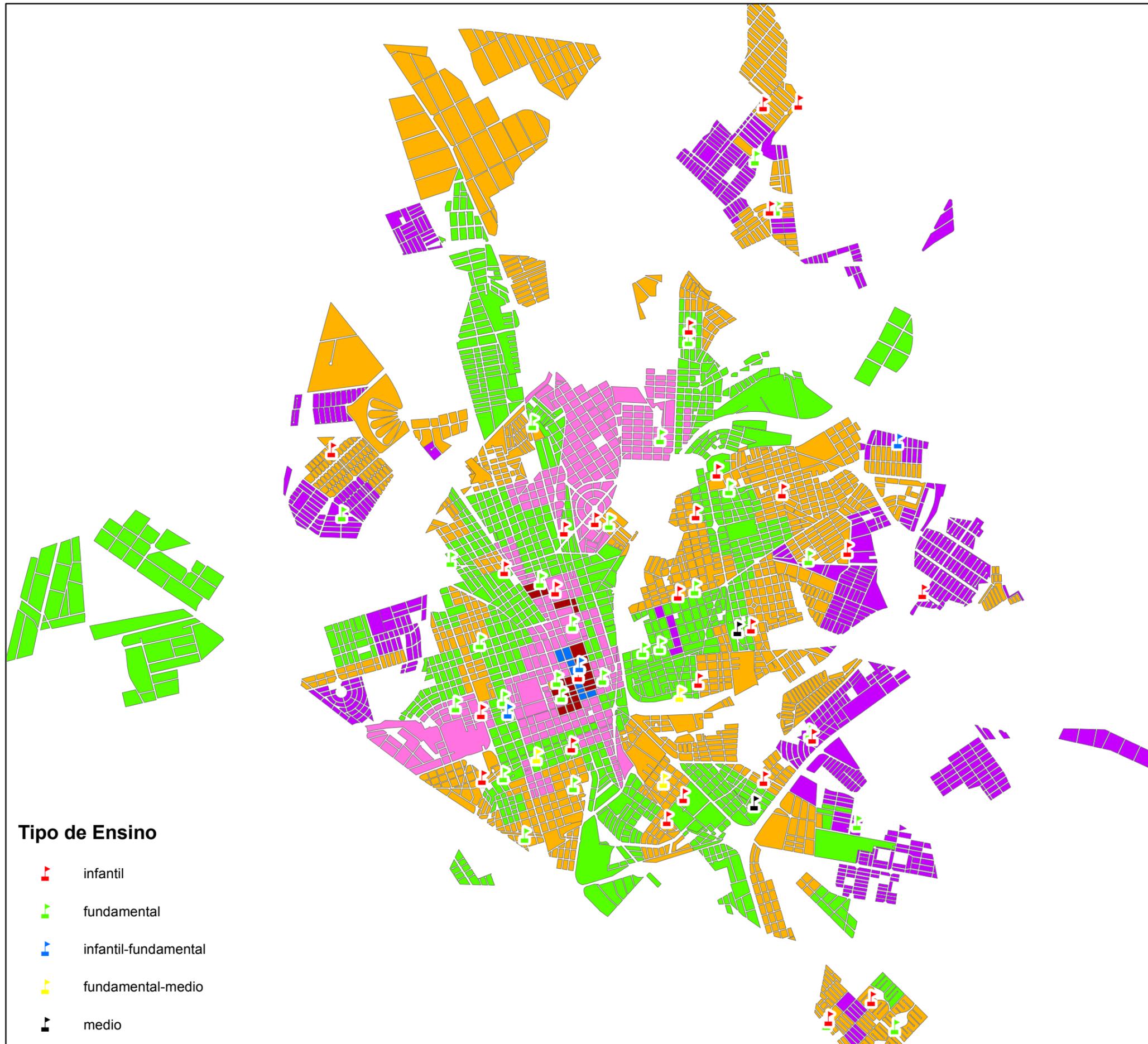


Datum: SAD 1969 - UTM - ZONA 23S
 Escala - 1:50.000



Tipo de Ensino

-  infantil
-  fundamental
-  infantil-fundamental
-  fundamental-medio
-  medio



5.2 Identificação de padrões

No software de extração de padrões *Apriori* extraíram-se 1030 padrões – 85 para distância próxima, 272 para distância média, e 673 para distância longe – dos quais foram filtrados no ambiente ARINE 50 para cada um dos 3 tipos do atributo distância (próxima, média e longe), de acordo com o maior suporte, totalizando 150 padrões. Destes, foram selecionadas algumas amostras das mais interessantes para análise.

Pelo fato do BD apresentar muitos dados de *distância=longe*, o valor do *Suporte* teve de ser minimizado para abranger os atributos de menor ocorrência (*distância=média* e *distância=próxima*). Com o intuito de se obter melhores valores de *Suporte* e maior quantidade de regras, o BD foi particionado para cada tipo de distância, trabalhando-se então, com valores distintos de *Suporte*: 0,70% para *distância=próxima* e *distância=média* e 10% para *distância=longe*. O valor da confiança foi padronizado para todos os tipos em 10%.

Na sequência são apresentadas as amostras de regras selecionadas com suas respectivas interpretações.

REGRA 1

Se DISTANCIA=longe então RENDA_PER_CAPITA=C
[Suporte=36,53%; Confiança=39,05%; Novidade=0,0042]

Interpretação: Considerando as escolas localizadas a uma distância superior ou igual a 1000 metros das quadras, em 36,53% dos casos, os habitantes dessas quadras possuem Renda per capita classe C, ou seja, entre 141,01 e 309 reais. Observa-se com esta regra que o atributo DISTANCIA=longe ocorre em 39,05% dos casos em que DISTANCIA=longe está associada ao atributo RENDA_PER_CAPITA.

REGRA 2

Se DISTANCIA=longe então ESCOLA=municipal
e ALUNOS_INFANTIL=299_A_397 [Suporte=19,13%; Confiança=20,45%;
Novidade=0,0027]

Interpretação: Considerando as escolas localizadas a uma distância superior ou igual a 1000 metros das quadras, em 19,13% dos casos, estas são

municipais, e possuem de 299 a 397 alunos matriculados no ensino infantil. Observa-se com esta regra que o atributo DISTANCIA=longe ocorre em 20,45% dos casos em que DISTANCIA=longe está associada ao atributo ALUNOS_INFANTIL.

REGRA 3

Se DISTANCIA=media então RENDA_PER_CAPITA=B2
[Suporte=2,23%; Confiança=86,98%; Novidade=0,0224]

Interpretação: Considerando as escolas localizadas a uma distância de 500 a 1000 metros das quadras, em 2,23% dos casos, os habitantes destas quadras possuem Renda per capita entre 309,01 e 556 reais. Observa-se com esta regra que o atributo DISTANCIA=media ocorre em 86,98% dos casos em que DISTANCIA=media está associada ao atributo RENDA_PER_CAPITA.

REGRA 4

Se DISTANCIA=proxima então DE7A14=6,7_A_73
[Suporte=0,88%; Confiança=32,66%; Novidade=0,0083]

Interpretação: Considerando as escolas localizadas a uma distância de até 500 metros das quadras, em 0,88% dos casos, de 6,7 a 73,0 habitantes das quadras possuem idades entre 7 e 14 anos, ou seja, de ensino fundamental. Observa-se com esta regra que o atributo DISTANCIA=proxima ocorre em 32,66% dos casos em que DISTANCIA=proxima está associada ao atributo DE7A14.

REGRA 5

Se DISTANCIA=proxima então RENDA_PER_CAPITA=B2
e ESCOLA=estadual [Suporte=0,98%; Confiança=36,24%;
Novidade=0,0091]

Interpretação: Considerando as escolas localizadas a uma distância de até 500 metros das quadras, em 0,98% dos casos, estas são estaduais, e os habitantes das quadras possuem Renda per capita entre 309,01 e 556 reais. Observa-se com esta regra que o atributo DISTANCIA=proxima ocorre em 36,24% dos casos dos casos em que DISTANCIA=proxima está associada ao atributo RENDA_PER_CAPITA.

Das regras selecionadas, a que apresenta maior valor de novidade é a **REGRA 3** – Novidade=0,0224, e a que apresenta menor valor de novidade é a **REGRA 2** – Novidade=0,0027. Ou seja, a **REGRA 3** apresenta um conteúdo mais inovador do que as outras apresentadas, e a **REGRA 2** não traz muita novidade, sendo que os valores da medida *Novidade* foram padronizados variando de -0,25 a 0,25.

Comparando-se os valores de novidade de todas as regras obtidas que variou de 0,0023 a 0,032 com o intervalo teórico de -0,25 a +0,25, observa-se que as regras possuem valores muito próximos de zero, representando um conhecimento que tende a confirmar o conhecimento existente, próximo do senso comum.

5.3 Comparação entre análises de Mapas Temáticos e Interpretação dos Padrões extraídos

Analisando as interpretações dos padrões extraídos juntamente com a ilustração da distribuição dos dados (Figuras 5.2 a 5.7 – mapas temáticos), pode-se inferir que a informação apresentada pela Regra 1, de que em cerca de 36% dos casos os habitantes de quadras distantes de escolas públicas possuem renda média per capita tipo C (141,01 a 309 reais), e pela Regra 2 que em 19% dos casos estas escolas distantes são municipais, possivelmente de ensino Infantil, com quantidade de alunos matriculados entre 299 e 397, fato também evidenciado na análise das Figuras 5.2 e 5.7, acrescentando-se a evidência de que a densidade demográfica destas áreas é de 50 hab/ha.

Porém, a Regra 1 apresenta um valor de novidade (0,0042) relativamente maior que a Regra 2 (0,0027), pode-se dizer que deve-se ao fato de possuir um suporte maior. Outro fator que pode justificar o grau de novidade superior da Regra 1 é o tipo de informação trazido por ela, pois evidencia que a população que reside distante de escolas possui Renda per capita média relativamente baixa, já a Regra 2 apresenta informações do tipo de escola e quantidade de alunos. Para uma prefeitura esse tipo de informação seria mais útil, na análise de atendimento a demanda de ensino, para um estudo de locais passíveis de implantação de escolas públicas, ou de ampliação do transporte público de maneira atingir toda a população de baixa renda em idade escolar.

5.4 Verificação da escalabilidade dos procedimentos em relação ao porte dos municípios

Em outros estudos a mesma metodologia foi aplicada a municípios de pequeno porte, por possuírem bases de dados menores. Tratam-se de São Pedro e Rio das Pedras, ambas localizadas no interior de São Paulo, com atualmente cerca de 30 mil habitantes.

No estudo realizado foram extraídos padrões no Software Magnum Opus e comparados a mapas temáticos de Distribuição de escolas e Densidade demográfica. A metodologia apresentou-se eficiente para pequenas bases de dados, indicando que os municípios são bem servidos de escolas públicas e mostrou que os resultados podem vir a subsidiar as prefeituras municipais em ações que propiciem a melhoria da qualidade de vida da população.

A base de dados do município de Araraquara, utilizada neste estudo, apresentou-se cerca de 5 vezes maior que as dos municípios de São Pedro e Rio das Pedras, utilizadas nos experimentos de Tagliatti (2007), nas quais foi aplicada a mesma metodologia, porém, para cada um destes estudos aplicou-se um software de extração de padrões diferente, de acordo com a necessidade requerida pelo tamanho da base de dados.

A Tabela 5.8 apresenta uma comparação entre as bases de dados citadas:

Município	São Pedro	Rio das Pedras	Araraquara
População	27.897	23.494	182.471
Escolas	19	18	62
Setores	43	38	262
Quadras	815	470	3.380
Exemplos	7.073	10.764	49.764
Regras Extraídas	180	246	1.030
Software utilizado	Magnum Opus	Magnum Opus	Apriori

Tabela 5.8: Análise comparativa entre bases de dados dos municípios.

Os municípios de São Pedro e Rio das Pedras possuem, cada um, cerca de 15% dos habitantes de Araraquara, portanto suas bases de dados são bem menores, ocasionando a extração de uma quantidade menor de padrões, possibilitando assim a seleção manual dos mais interessantes. Devido a elevada quantidade de exemplos, a extração dos padrões com a base de dados de Araraquara gerou uma grande quantidade de regras, necessitando do ambiente ARINE para a seleção das mais interessantes para interpretação.

Quanto ao tempo demandado nos experimentos, pode-se dizer que está diretamente ligado ao tamanho da base de dados, principalmente na etapa de pré-processamento dos mesmos, por se tratar de um trabalho minucioso e requerer maior atenção durante as associações. Por este fato, a base de dados do município de Araraquara apresentou maior grau de complexidade em relação aos demais.

Considerações finais

Conforme evidenciado, a deficiência de infra-estrutura educacional ocorre predominantemente nas áreas mais periféricas e rurais, requerendo estudos e análises focadas na distribuição de escolas pelos municípios, para implementação de programas por parte dos governos, que melhorem esse quadro.

Estudos que associam dados sócio-econômicos a espaciais, com o auxílio da ferramenta SIG, e aplicação da técnica de Mineração de Dados Espaciais podem contribuir na extração automatizada de conhecimentos inovadores, escondidos em grandes quantidades de dados, apresentando um panorama do atendimento a demanda por ensino público, como é o caso do município de Araraquara.

6.1 Principais contribuições

A verificação do atendimento da rede pública de ensino infantil, fundamental e médio em relação as faixas de idade condizentes com cada tipo de ensino, evidenciou conhecimentos importantes. Um exemplo disso são pessoas de baixa renda, habitantes das áreas periféricas do município de Araraquara, que não possuem escolas públicas nas proximidades de suas residências. Resta-lhes a opção de caminhar grandes distâncias até a instituição de ensino mais próxima, que ofereça o tipo de ensino adequado para

sua idade, pois vale ressaltar, que a instituição mais próxima, nem sempre oferece o tipo de ensino condizente com a idade da pessoa. Estes fatos podem ser confirmados nos mapas temáticos elaborados neste estudo.

Quanto a metodologia aplicada, a utilização da técnica de Mineração de Dados Espaciais para dados sócio-econômicos apresentou-se viável na descoberta de conhecimentos em grandes bases de dados, no caso, a de um município de médio porte. Contando com quase 50 mil associações (exemplos) a composição da base de dados sócio-econômicos e de ensino do município de Araraquara apresentou elevado grau de complexidade, requerendo bastante atenção para que, durante as associações, alguns dados não fossem perdidos, o que poderia descaracterizar a base e comprometer a veracidade das informações resultantes.

A discretização dos dados foi, dentro da etapa de pré-processamento dos dados, de suma importância para viabilizar a interpretação dos padrões extraídos, pois transformou dados numéricos em intervalos mais próximos de nossa realidade, como o caso da distância (*próxima, média e longe*) e das classes de renda (*A1, A2, B1, B2, C, D e E*), tornando a regra mais fácil de ser analisada, agilizando a busca por conhecimentos interessantes após a filtragem no ambiente ARINE. Porém dados numéricos como quantidade de pessoas em cada tipo de ensino apresentaram uma distribuição muito variada, de modo que, se fossem discretizados da mesma maneira que os demais dados, poderiam mascarar a realidade. Para tanto um software estatístico foi utilizado, e por análise de frequência, criados intervalos numéricos.

A metodologia utilizada também pode ser aplicada a quaisquer bases de dados, contanto que seja refinada para a devida finalidade. Para o caso de dados urbanos, sugere-se a aplicação da mesma às demais infra-estruturas municipais, como demandas por habitação, saúde, transporte público, etc, com o intuito de subsidiar as autoridades competentes em tomadas de decisão, contribuindo na melhoria da qualidade de vida da população.

6.2 *Recomendações para trabalhos futuros*

São apresentadas na sequência algumas propostas para trabalhos futuros, visando a continuidade e complementação deste estudo:

- desenvolver o mesmo estudo com os dados do próximo censo demográfico do IBGE, a fim de comparar os resultados, já que se trata de um dado que varia

bastante de censo para censo (as crianças que tinham 6 anos em 2000, atualmente têm 15 anos, ou seja, mudaram de faixa etária escolar);

- utilizar outra técnica de Mineração de Dados que não a de Regras de Associação, para extração de dados;
- aprofundar os estudos sobre a discretização em classes por Regras de Associação numérica;
- realizar o mesmo procedimento considerando intervalos de distâncias variáveis de acordo com o tipo de ensino, e incluindo dados de escolas particulares.

Referências

- ABEP (2003). Critério de classificação econômica brasil. Disponível em: http://www.abep.org/codigosguias/ABEP_CCEB.pdf. 27, 30, 37
- Agrawal, R. e R. Srikant (1994). Fast algorithms for mining association rules. In *Proceedings of 20th International Conference on Very Large Data Bases*, pp. 487–499. 17
- Agrawal, R. e R. Srikant (1995). Mining sequential patterns. In *Eleventh International Conference on Data Engineering, VLDB*, Taipei, Taiwan, pp. 3–14. IEEE. 13
- Amancio, M. A. (2005). Relacionamento entre a forma urbana e as viagens a pé. Dissertação de Mestrado, Programa de Pós Graduação em Engenharia Urbana – UFSCar. 1
- Barcellos, C. e W. Ramalho (2002). Situação atual do geoprocessamento e da análise de dados espaciais em saúde no Brasil. *Revista Informática Pública* vol. 4, Belo Horizonte, MG. Disponível em: http://www.ip.pbh.gov.br/AN04_N2_PDF/ip0402barcellos.pdf. Acesso em: 24 set. 2008. 9, 11
- Barcelos, F. B., N. D. Pizzolato, e L. A. N. Lorena (2002). Avaliação da localização de escolas com modelos capacitado e não-capacitado e uso de uma ferramenta GIS: Estudo de caso de Vitória/ES. Disponível em: <http://www.lac.inpe.br/~lorena/sbpo2002/SBP02002-lorena.pdf>. Acesso em: 03 mar. 2008. 2, 6
- Bogorny, V. (2006). Enhancing Spatial Association Rule Mining in Geographic Databases. Tese de Doutorado, Instituto de Informática, Universidade Federal do Rio Grande do Sul. 16, 18
- Casanova, M., G. Câmara, C. Davis, L. Vinhas, e R. G. Queiroz (2005). *Banco de dados geográfico*. Curitiba, PR: MundoGEO. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/bdados/>. Acesso em: 25 set. 2008. 9

- Chhetri, P., J. Corcoran, R. Stimson, J. Cooper, e D. Pullar (2007). *Subjectively Weighted Development Scenarios for Urban Allocation: A Case Study of South East Queensland*. Ph. D. thesis, Oxford – UK. 8
- Câmara, G. e C. Davis (1999). *Introdução à ciência da geoinformação*, pp. 307–317. INPE. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/introd/cap1-introducao.pdf>. Acesso em: 04 abr. 2007. 8
- Collaço, D. L. (2003). *Caracterização geológico-geotécnica do município de São Leopoldo para fins de planejamento como subsídio ao Plano Diretor*. Ph. D. thesis, Instituto de Ciências Matemáticas e de Computação - USP São Carlos, São Carlos. Dissertação de Mestrado – EESC-USP. 9
- Demsar, U. (2007). *Knowledge Discovery in the Environmental Sciences: Visual and Automatic Data Mining for Radon Problems in Groundwater*. Ph. D. thesis, Oxford – UK. 10
- Diniz, R. A. A. e F. L. Neto (2000). *Data Mining: Uma Introdução*. Associação Brasileira de Estatística. Caxambu. 14
- Domingues, A. M. (2004). *Generalização de regras de associação*. Ph. D. thesis, Instituto de Ciências Matemáticas e de Computação - USP São Carlos, São Carlos. Dissertação de Mestrado – EESC-USP. 11, 12, 13, 15
- Druck, S., M. S. Carvalho, G. Câmara, e A. M. V. Monteiro (2005). *Análise espacial de dados geográficos*. Embrapa Cerrados. 9
- Dutra, N. G. S. (1998). *Planejando uma rede escolar municipal para reduzir custos de deslocamentos*. Dissertação de Mestrado, Departamento de Transportes. 5
- Fayyad, U. M., G. Piatetsky-Shapiro, e P. Smyth (1996). *From Data Mining to Knowledge Discovery: An Overview*, pp. 6–22. Menlo Park, CA: AAAI. 11
- FNDE (2007). Programa caminho da escola. Disponível em: http://www.fnde.gov.br/home/index.jsp?arquivo=caminho_escola.html. Acesso em: 28 fev. 2007. 1
- Ghanem, E. (2004). A democratização do acesso à escola. in: *Educação escolar e democracia no Brasil*. In *Autêntica – Ação Educativa*, Belo Horizonte, pp. 39–63. Disponível em: <http://pedagogia.incubadora.fapesp.br/portal/SOCEDII>. Acesso em: 22 mai. 2008. 4
- IBGE (2003). *Censo demográfico 2000. Resultados do universo: Agregados de setores censitários – DVD 2 Rio de Janeiro, Rio Grande do Sul, Rio Grande do Norte, Rondônia, Roraima, Santa Catarina, São Paulo, Sergipe, Tocantins*. 21, 25, 28
- Koperski, K. e J. Han (1995). *Discovery of spatial association rules in geographic*

- information databases. In *Proceedings of the 4th International Symp. Advances in Spatial Databases*, pp. 47–66. Springer-Verlag. 13, 14
- Ladner, R., F. E. Petry, e M. A. Cobb (2007). Fuzzy set approaches to spatial data mining of association rules. *Transactions in GIS. Blackwell Publishing Ltd v.7, n.1*, 123–138. 13
- Lavrac, N., P. Flach, e B. Zupan (1999). *Rule Evaluation Measures: A Unifying View*. Springer Berlin / Heidelberg. 28
- Maguire, J. D., M. F. Goodchild, e D. W. Rhind (1991). *Geographical Information Systems, Principles and Applications*. Longman Scientific Technical. 10
- Malerba, D., F. Espisito, F. Lisi, e A. Appice (2002). *Mining Spatial Association Rules in Census Data*. Ph. D. thesis, Dipartimento di Informatica – Università degli Studi di Bari, Via Orabona, Bari, Itália. 16
- Martins, V. (2004). Como garantir igualdade de acesso à escola. Disponível em: <http://kplus.cosmo.com.br/materia.asp?co=134&rv=Direito>. Acesso em: 22 mai. 2008. 4
- Melanda, E. A. (2004). Pós-processamento de Regras de Associação. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação - USP São Carlos. 12, 27, 28
- Melanda, E. A. e S. O. Rezende (2003). Sintaxe padrão para representar Regras de Associação. Relatórios técnicos do ICMC - USP São Carlos. 13
- Melo, M. F. (2001). Planta de valores genéricos: um produto cartográfico para a equidade tributária e o planejamento municipal. Dissertação submetida a avaliação do mestrado em Estudos Populacionais e Pesquisas Sociais, área de concentração em Produção e Análise da Informação Geográfica, como requisito parcial para obtenção do grau de mestre. ENCE – Rio de Janeiro. 8
- Mennis, J. e J. W. Liu (2005). Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. *Transactions in GIS. v.9, n.1*, 5–17. 17
- Oliveira, C. L. e L. C. Bastos (2007). *O uso do geoprocessamento no auxílio à tomada de decisão na alocação de serviços públicos. Estudo de caso: rede educacional de ensino do bairro Harmonia – Canoas, RS*. Ph. D. thesis, Florianópolis. Disponível em: <http://martem.dpi.inpe.br/coldpi.inpe.br/sbsr@80200611.15.04.28doc/5443-5450.pdf>. Acesso em: 22 Maio 2008. 6
- Pegoretti, M. S. e S. P. Sanches (2006). Indicador de acessibilidade dos alunos da zona rural às escolas da zona urbana. *Revista dos Transportes Públicos* 28, 83–98.

- Pinto, M. V. (2001). Cadastramento escolar: Democratização do acesso à escola pública. *Informática Pública* 1(2). Disponível em: http://www.ip.pbh.gov.br/AN01_N2_PDF/ip0102viniciuspinto.pdf. Acesso em: 22 Maio 2008. 4
- Pivato, M. A. (2006). Mineração de Regras de Associação em dados geo-referenciados. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – USP São Carlos. 16, 18, 19
- Pizzolato, N. D., A. G. Barros, F. B. Barcelos, e A. G. Canen (2004). *Localização de Escolas Públicas: Síntese de algumas linhas de experiências no Brasil*. Ph. D. thesis. 1
- Pizzolato, N. D., G. G. Silva, e S. Mizubuti (1999). *Avaliação da oferta de ensino fundamental pela rede pública e sua distribuição espacial: aplicação ao município de Niterói (RJ)*. Ph. D. thesis. 6
- Ramos, I. e M. Y. Santos (2003). Mining no suporte à construção de conhecimento organizacional. Edição em CD-ROM Ú- ISBN 972-9354-42-1. 12
- Rezende, S. O., J. B. Pugliesi, E. A. Melanda, e M. F. Paula (2003). *Sistemas Inteligentes: Fundamentos e Aplicações* (1 ed.), Chapter 12 – Mineração de Dados, pp. 307–335. Barueri, SP: Manole. 11, 12, 13, 22
- Rocha, D. (2007). Transporte escolar rural é precário em vários municípios, revela pesquisa. Disponível em: <http://brasilatual.com.br/sistema/?p=809>. Acesso em: 15 jan. 2009. 5
- Rodrigues, V. L. (2008). Mapeamento geotécnico como base para o planejamento urbano e ambiental: Município de Jaú/SP. Dissertação de Mestrado, Geotecnia – USP São Carlos. 10
- Santos, M. Y. C. (2004). Padrão: um Sistema de Descoberta de Conhecimento em Bases de Dados Geo-referenciadas. Tese de Doutorado, Departamento de Sistemas de Informação. 17
- SEESP (2007). Arquivo de escolas municipal. Disponível em: <http://escola.edunet.sp.gov.br/Download/downloads.htm>. Acesso em: 28 fev. 2007. 22
- Sionara, R. A. (2006). Identificação de regras de associação interessantes por meio de análises com medidas objetivas e subjetivas. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – USP São Carlos. 11
- SNIU (2008). Perfil municipal – araraquara-sp. Disponível em: <http://www.cidades.gov.br/secretarias-nacionais/saneamento-ambiental/indicadores/Sniu.zip/view>. Acesso em: 10 mar. 2008. 20
- Tagliatti, C. (2007). *Estudo de relações espaciais entre a distribuição de dados sócio-*

- econômicos e a localização de escolas*. Ph. D. thesis, Universidade Federal de São Carlos, São Carlos. Trabalho de Graduação – UFSCar. 3, 24, 27, 49
- Tagliatti, C. e E. A. Melanda (2007). *Análise de viabilidade na aplicação de algoritmos de extração de padrões em associações entre objetos espaciais e dados censitários*. Ph. D. thesis, São Carlos. Anais... CD-ROM. 17
- Tagliatti, C. e E. A. Melanda (2008). *Estudo da relação entre escolas: Avaliação da distribuição espacial de escolas usando Regras de Associação Espaciais*. Ph. D. thesis, Santos. Anais... CD-ROM. 17
- Tagliatti, C., E. A. Melanda, e M. A. Pivato (2006). *Aplicação de Regras de Associação na análise de dados geográficos da cidade de Jaboticabal*. Ph. D. thesis, São Carlos. Anais... CD-ROM. 17, 25
- Tagliatti, C., E. A. Melanda, e S. A. Röhm (2007). *Mineração de Dados Espaciais na identificação da relação de proximidade de escolas de acordo com a faixa etária da população das cidades de São Pedro e Rio das Pedras*. Ph. D. thesis, São Carlos. Anais... CD-ROM. 24
- Trad, L. I. e J. M. P. Cunha (2005). *Acessibilidade aos equipamentos da educação pública em um grande centro urbano: o caso de Campinas*. Ph. D. thesis, Campinas. 6
- Webb, G. I. (1995). OPUS: An efficient admissible algorithm for unordered search. *Journal of Artificial Intelligence Research* 3, 431–465. 18
- Witten, I. H. e E. Frank (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. São Francisco, Califórnia: Morgan Kaufmann. Disponível em: <http://www.cs.waikato.ac.nz/~ml/weka/book.html>. 18