UNIVERSIDADE ESTADUAL PAULISTA
Faculdade de Ciências e Tecnologia
Câmpus de Presidente Prudente
Programa de Pós-Graduação em Ciências Cartográficas

Sergio Ricardo Ribas Sass

Abordagens de Descoberta de Conhecimento em Bases de Dados Aplicadas ao Cadastro Territorial Multifinalitário

PRESIDENTE PRUDENTE
2013

UNIVERSIDADE ESTADUAL PAULISTA
Faculdade de Ciências e Tecnologia
Câmpus de Presidente Prudente
Programa de Pós-Graduação em Ciências Cartográficas

Sergio Ricardo Ribas Sass

Abordagens de Descoberta de Conhecimento em Bases de Dados Aplicadas ao Cadastro Territorial Multifinalitário

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciências Cartográficas da Universidade Estadual Paulista – Campus de Presidente Prudente, como requisito final para a obtenção do título de Mestre em Ciências Cartográficas. Área de Concentração: Aquisição, Análise e Representação da Informação Espacial.

Orientador: Prof. Dr. Amilton Amorim
Co-Orientador: Prof. Dr. Milton Hirokazu Shimabukuro

PRESIDENTE PRUDENTE

2013

	Sass, Sergio Ricardo Ribas.
S264a	Abordagens de descoberta de conhecimento em Bases de Dados Aplicadas ao Cadastro Territorial Multifinalitário / Sergio Ricardo Ribas Sass. - Presidente Prudente : [s.n], 2013
	73 f.
	Orientador: Amilton Amorim Coorientados: Milton Hirokazo Shimabukuro
	Dissertação (mestrado) - Universidade Estadual Paulista, Faculdade de Ciências e Tecnologia
	Inclui bibliografia
	1. Cadastro Territorial Multifinalitário. 2. Representação da Informação Espacial. 3. Mineração de dados (Computação). I. Amorim, Amilton. II. Shimabukuro, Milton Hirokazo. III. Universidade Estadual Paulista. Faculdade de Ciências e Tecnologia. IV. Abordagens de descoberta de conhecimento em Bases de Dados Aplicadas ao Cadastro Territorial Multifinalitário.

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Presidente Prudente

BANCA EXAMINADORA

PROF. DR. **AMILTON AMORIM**
ORIENTADOR

PROF. DR. **RONALDO CELSO MESSIAS CORREIA**
UNESP/FCT

PROFA. DRA. **LIA CAETANO BASTOS**
UFSC

SERGIO RICARDO RIBAS SASS

Presidente Prudente (SP), 26 de fevereiro de 2013.

RESULTADO: APROVADO

Faculdade de Ciências e Tecnologia
Seção Técnica de Pós-Graduação
Rua Roberto Simonsen, 305 CEP 19060-900 Presidente Prudente SP
Tel 18 3229-5318 fax 18 3223-4519 posgrad@fct.unesp.br

Dedico a um grande homem, Sr. Augusto, desbravador, amigo, companheiro e que ficou entre nós por um século, e no dia 06/02/2013 Deus resolveu levá-lo.

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus, que me fortalece, me guia e me ampara nos momentos de aflição.

Aos meus pais, pelos conselhos, pela insistência e pela educação transmitida e investida.

À minha esposa, maravilhosa, companheira, dedicada, amorosa. Não tenho adjetivos suficientes para elogiar. Foram vários tropeços e nunca, em nenhum momento ela deixou de acreditar. E claro, várias alegrias também que espero compartilhar pelo resto de minha vida.

Aos meus filhos que tiveram que entender a ausência do pai em vários momentos de estudos.

Ao Professor Amilton Amorim que, sempre presente, acompanhou minha trajetória e foi direcionando e aconselhando quando os caminhos se mostravam tortuosos.

Ao Professor Milton Hirokazu Shimabukuro que, por meio de seus conhecimentos em computação me auxiliou em diversas etapas do trabalho.

Aos meus amigos que fizeram parte dessa família UNESP, um agradecimento especial ao amigo Marcelo Solfa pela dedicação e companheirismo. E também aos amigos de fora, amigos mais particulares, que nunca entenderam e nem fizeram questão de entender o motivo de tanto estudo. Sempre precisei deles em momentos de descontração.

À UNESP de Presidente Prudente pela estrutura oferecida para que o trabalho fosse feito da melhor maneira possível. Estrutura física e pessoal, dedico também a todos os funcionários que me atenderam sempre com a maior dedicação.

Aos órgãos CNPQ e CAPES pelas bolsas oferecidas, incentivo extremamente importante para a dedicação do aluno.

RESUMO

O Banco de Dados na gestão pública é um recurso computacional que precisa ser administrado com a mesma importância de um ativo financeiro de uma organização, pois dá suporte à qualidade de suas operações.

Com o grande crescimento da quantidade de dados armazenados nesses Bancos de Dados, os gestores passaram a depender não só de informações, mas também de conhecimentos extraídos desses dados como suporte no processo de tomada de decisão.

O Cadastro Territorial Multifinalitário (CTM) é a ferramenta que gerencia os dados da organização pública, e juntamente com ele, para extrair conhecimento desses dados, novas técnicas computacionais se tornam grandes aliadas, como *Data Warehouse* (*DW*) e *Data Mining(DM).*

Essa pesquisa discute a tecnologia de *DM, e,* aliado ao CTM, mostra os resultados de alguns experimentos para a cidade de Ribeirão dos Índios-SP.

Palavras chaves: Banco de Dados; tomada de decisão; Cadastro Territorial Multifinalitário; *Data Warehous;, Data Mining.*

ABSTRACT

The Database in the public management is a computational resource which needs to be administered with the same importance as a financial asset of an organization, because it supports the quality of its operations.

With the large growth in the amount of data stored in these Databases, the managers have come to depend not only of information but also knowledge extracted from these data to help in the decision-making process.

The Multipurpose Cadastre (CTM) is the tool that manages the public organization's data, and along with it, to extract knowledge of these data, new computational techniques become great allies, such as the Data Warehouse (DW) and the Data Mining (DM).

This research discusses the technology of the DM, and, allied to the CTM, shows the results of some experiment for the town of Ribeirão dos Índios, SP.

Keywords: Database, decision-making; Multipurpose Cadastre; Data Warehouse; Data Mining.

LISTA DE FIGURAS

LISTA DE SIGLAS

a. C.	antes de Cristo
ABNT	Associação Brasileira de Normas Técnicas
BDM	Banco de Dados Multidimensional
BDR	Banco de Dados Relacional
BIC	Boletim de Informações Cadastrais
CTM	Cadastro Territorial Multifinalitário
DW	*Data Warehouse*
ETL	*Extraction, Transformation and Load*
FIG	*Fedération Internationale de Gèométres*
IBGE	Instituto Brasileiro de Geografia e Estatística
IPTU	Imposto Predial e Territorial Urbano
ISO	*International Organization for Standardization*
OLTP	*On-Line Transaction Processing*
PE	*Processing Elements*
RNA	Redes Neurais Artificiais
ROLAP	*Relational On-Line Analytical Processing*
SDA	Análise de Dados Simbólicos
SGBD	Sistema de Gerenciamento de Banco de Dados
SGBDR	Sistema de Gerenciamento de Banco de Dados Relacional
SI	Sistema de Informação
SIT	Sistema de Informação Territorial
SQL	*Structure Query Language*
SOM	*Self Organizing Maps*
TI	Tecnologia da Informação

SUMÁRIO

1 INTRODUÇÃO

Com o avanço da tecnologia, diversas ferramentas computacionais surgiram com objetivo de facilitar, agilizar e dar mais qualidade nas tarefas executadas pelas organizações empresariais, públicas ou privadas. Um exemplo disso são as Bases de Dados informatizadas que, juntamente com os Sistemas de Informação, são responsáveis por armazenar e gerenciar os dados da organização, buscando alcançar maior qualidade em suas operações administrativas e planejamentos estratégicos.

Com o aumento da demanda pela informação, a quantidade de dados coletados e acumulados vem crescendo muito rapidamente nos últimos anos em virtude do processo de informatização da sociedade e do rápido desenvolvimento de ferramentas de coleta e armazenamento de dados (HAN; KAMBER; PEI, 2005). Associado a isso, a crescente demanda por conhecimento novo, voltado para decisões estratégicas tem despertado o interesse em descobrir novos conhecimentos intrínsecos nas bases de dados (ROMÃO, 2002).

O crescimento dessas bases de dados as tornaram importantes fontes de informações e conhecimentos, recursos que auxiliam analistas de negócios no processo de tomada de decisão (O'BRIEN 2003). A utilização de ferramentas, técnicas e tecnologias apropriadas ao melhoramento da obtenção, tratamento, apresentação e disponibilização desses recursos é um fator que pode influenciar no aumento da competitividade da organização (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Porém, para uma organização pública, não é vital o ganho de competitividade. Então qual seria a importância da extração de informações e conhecimentos a partir de suas bases de dados?

De acordo com Pacheco (*apud* Kurahassi, 1999), desde a Constituição de 1988, o poder público municipal perdeu a característica de unidade apenas administrativa e assumiu o papel de unidade gestora e corresponsável pelo atendimento das necessidades sociais. Os municípios brasileiros tiveram suas responsabilidades e recursos expandidos. Por outro lado, as demandas sociais aumentaram e os desafios trazidos pela globalização impuseram novos campos de ação aos municípios.

O Ministério do Planejamento propõe um modelo de excelência na gestão pública onde informações e conhecimentos são as bases de todo o processo, como mostra a Figura 1.

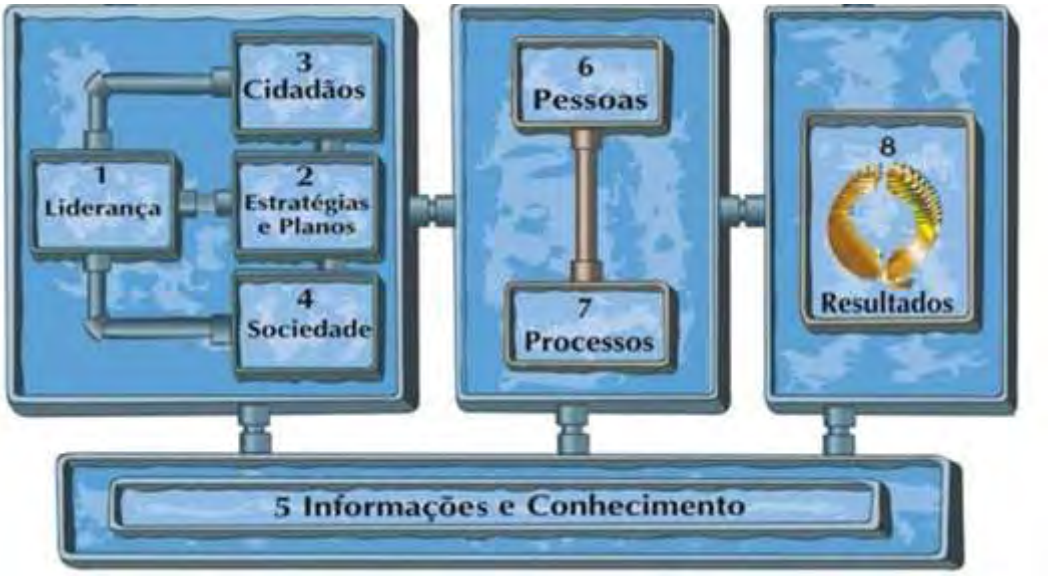


Figura 1: Modelo de Excelência na Gestão Pública
Fonte: (BRASIL, 2008)

Percebe-se então, que também nas prefeituras, informações e conhecimentos se tornaram vitais, tanto na tomada de decisão em níveis estratégicos e gerenciais, como em simples tarefas rotineiras e processos de trabalho.

Nas prefeituras brasileiras, o Cadastro Territorial Multifinalitário (CTM) é considerado a Base de Dados que armazena a identificação e caracterização de parcelas cadastrais[1], indivíduos ou elementos, e o Sistema de Informação Territorial (SIT) responsável por informatizar o armazenamento e gerenciamento desses dados (GARCIA 2007).

O CTM deve possuir a capacidade de integrar dados de áreas teoricamente distintas (fiscais, legais, socioeconômicos, etc), mas com o objetivo comum de nortear políticas públicas. Sua evolução no decorrer dos anos mostra que ele passou de um simples gerenciador de cobranças de impostos a uma complexa Base de Dados de gestão territorial. Com isso, dois problemas surgiram:

1. Aumento da quantidade de dados armazenados - no momento em que as prefeituras resolveram aplicar todos os objetivos do CTM, ocorreu um aumento considerável na estrutura da Base de Dados e na quantidade de registros armazenados;
2. Heterogeneidade dos dados – apesar da intenção de colocar o CTM em aplicação, ainda não existe uma padronização na coleta e armazenamento desses dados. Isso ocasiona heterogeneidade no formato dos mesmos.

[1] De acordo com a Portaria 511/2009 do Ministério das Cidades " parcela cadastral é a menor unidade do cadastro, definida como uma parte contígua da superfície terrestre com regime jurídico único."

Para resolver esses problemas, e proporcionar informações e novos conhecimentos ao gestor público, existem técnicas computacionais que auxiliam os Administradores de Banco de Dados (*Data Base Administrator DBA*) na análise automatizada de dados e extração de conhecimentos úteis a partir de grandes Bases de Dados. Essas técnicas são conhecidas como Mineração de Dados (*Data Minig - DM*) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Para a aplicação dessas técnicas, todo um processo de tarefas deve ser respeitado e executado. Esse processo é conhecido como Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Database – KDD*).

1.1 *Objetivo*

Apresentar uma proposta de aplicação de Técnicas de *DM* em um conjunto de bases de dados cadastrais urbanas buscando resultados que auxiliem os gestores no processo de tomada de decisão. A proposta visa aplicar as técnicas de *DM* sem a utilização de um *DW*.

Objetivos Específicos

- Mostrar a aplicação do processo de KDD e de técnicas de DM em bases de dados heterogêneas;
- Buscar novas técnicas de visualizações de resultados obtidas em bases de dados além das comumente executadas por Linguagem de Consulta Estruturada (Structure Query Language – SQL);

1.2 *Justificativa*

A evolução do CTM ocasionou um aumento na quantidade e na heterogeneidade dos dados armazenados nas suas Bases, causando uma grande complexidade na obtenção de conhecimento adequado e útil como auxílio ao gestor. Essa complexidade se deu em virtude das adaptações feitas nos bases de dados cadastrais no decorrer dos anos para atender as características multifinalitárias do Cadastro Territorial Urbano.

A falta de Análise de Requisitos no processo de concepção e construção de uma base de dados bem estruturada e capaz de armazenar dados históricos resultou em várias anomalias responsáveis por encobrir informações relevantes no planejamento estratégico.

Atualmente, as bases de dados cadastrais urbanas “escondem” conhecimentos importantes que poderiam ser utilizados como parte do processo de tomada de decisão à nível municipal. Porém, a quantidade de variáveis envolvidas, muitas vezes não deixam visíveis determinados relacionamentos na busca por esse conhecimento.

Para esses casos, a tecnologia de *DM* com todas as suas etapas, surge com o objetivo de buscar e extrair o conhecimento dessas bases de dados auxiliando o gestor público no processo de tomada de decisão.

1.3 *Estrutura do Trabalho*

Capítulo 1: Introdução apresenta uma visão geral do trabalho, objetivos, justificativa e a estrutura.

Capítulo 2: Levantamento Bibliográfico apresenta um estudo sucinto sobre CTM, Sistema Gerenciador de Banco de Dados (SGBD), Sistema de Apoio a Decisão envolvendo: Dado x Informação x Conhecimento, *KDD*, *Data Mining* e Técnicas de Visualização.

Capítulo 3: Estudo de caso apresenta um estudo preliminar de *DM* sobre bases de dados cadastrais da cidade de Ribeirão dos Índios.

Capítulo 4: Considerações finais dos experimentos preliminares.

2 LEVANTAMENTO BIBLIOGRÁFICO

2.1 *Cadastro Territorial Multifinalitário*

O sistema cadastral auxilia no planejamento urbano, arrecadação do Imposto Predial e Territorial Urbano (IPTU), fiscalização do uso do solo, otimização dos recursos e gestão de equipamentos urbanos. No decorrer dos anos, esse sistema vem evoluindo e ganhando importância principalmente no processo de tomada de decisão dos gestores públicos.

O CTM é uma ferramenta que armazena dados e auxilia na análise econômica (valor do imóvel e do imposto), geométrica (localização, forma e dimensões da parcela), jurídica (principalmente no registro de imóveis), sociais (perfil do proprietário e outros) e ambientais de um determinado lugar geográfico. Esses dados são obtidos, geralmente, por meio de censos e levantamentos cadastrais específicos(MALAMAN; AMORIM, 2010).

2.1.1 Histórico

Muito antes do termo CTM ser apresentado, a palavra Cadastro era usada para representar demarcações territoriais que existiam antigamente. Evidências mostram que conceitos de Cadastro apareceram por volta de 4000 a.C. na Babilônia com intuito fiscal e como forma de organização da sociedade com a demarcação da terra (PHILIPS, 2004).

Porém, é na França, com Napoleão no ano de 1807, que acontece o marco da revolução cadastral. Após a revolução francesa foi decretado um completo levantamento de todo território nacional bem como das terras ocupadas, com fins estratégicos, de estímulo à cidadania e de tributação justa dos imóveis (LARSSON, 1996).

Não existe consenso entre os autores sobre a definição do termo e das funções do Cadastro. Diferentes concepções são apresentadas e até mesmo na etimologia é difícil precisar seu significado. O dicionário Aurélio da língua portuguesa mostra que, Cadastro deriva do termo francês Cadastre, que significa registro público dos bens imóveis de um determinado território, os registros dos bens privados de um determinado indivíduo (ERBA, 2005).

No Brasil, o marco do Cadastro foi o ano de 1854 quando, pelo Decreto nº 1318[2], foi regulamentada a Lei nº 601[3] diferenciando os bens de domínio público do particular, criando o registro paroquial das terras e obrigando proprietários rurais a registrarem suas terras.

[2] Disponível em: http://www.planalto.gov.br/ccivil_03/decreto/Historicos/DIM/DIM1318.htm
[3] Disponível em: http://www.jusbrasil.com.br/legislacao/104056/lei-601-50

Porém, a relação desse tipo de Cadastro com o Cadastro Territorial só aconteceu em 1964 com a criação do Estatuto da Terra por meio da Lei 4.504[4] que regulamenta os direitos e obrigações relacionados aos bens e imóveis rurais com o objetivo de desenvolver a reforma agrária e promoção de políticas agrícolas (ANTUNES, 2007; LOCH, 2007).

Em 2001 foi instituída a Lei 10.267[5] que padroniza os procedimentos desde a caracterização do imóvel até a sua localização por meio das coordenadas dos vértices definidores dos limites. Mas, até então, nada referenciava os imóveis urbanos. Somente em 1998 a Associação Brasileira de Normas Técnicas (ABNT) publicou a norma NBR – 14166/1998[6] estabelecendo regras para a implantação e manutenção da Rede de Referência Cadastral Municipal (AMORIM et al, 2007).

E em 2009, o Ministério das Cidades estabeleceu as diretrizes para o CTM na portaria 511[7].

2.1.2 Funções do CTM

Os primeiros cadastros não tinham a visão Multifinalitária, portanto visavam apenas à arrecadação. Foram projetados para apoiar a tributação territorial, registravam o valor da parcela, a partir do qual era calculado o valor do imposto territorial. Tempo depois a preocupação com o ordenamento territorial adicionou a visão jurídica ao cadastro, melhorando a eficiência e segurança das transações em relação à posse da terra em alguns países. (FIG, 2010).

Atualmente, muitos dos cadastros implementados nas prefeituras, ainda perseguem os objetivo tributários, porém, novos atributos foram inseridos para obter métodos de avaliações mais precisos. Detalhes construtivos, localização, forma e dimensão dos terrenos, são exemplos desses novos atributos utilizados na avaliação (ERBA, 2005).

Na década de 1990, dois eventos marcaram uma mudança de paradigma referente ao Cadastro Territorial: a Conferência das Nações Unidas sobre Meio Ambiente e Desenvolvimento realizada na cidade do Rio de Janeiro no ano de 1992 e a Segunda Conferência das Nações Unidas sobre Assentamentos Humanos. A partir desses eventos, fica

[4] Disponível em: http://www.planalto.gov.br/ccivil_03/leis/L4504.htm.

[5] Disdponível em: http://www.planalto.gov.br/ccivil_03/leis/leis_2001/l10267.htm

[6] Disponível em: http://www.abntcatalogo.com.br/norma.aspx?ID=3961z

[7] Disponível em: http://www.cidades.gov.br/images/stories/ArquivosCapacitacao/Capacita%C3%A7%C3%A3o/Editais/Portaria511_CTM.pdf

clara a importância da informação territorial confiável para apoiar os processos de tomada de decisões, para preservação do meio ambiente e promoção do desenvolvimento sustentável. O Cadastro Territorial então soma a seus dados econômico-físico-jurídicos, dados ambientais e sociais de seus ocupantes, consolidando assim a nova visão de Cadastro Territorial Multifinalitário (CTM) (ERBA, 2005).

Com essa nova visão, as prefeituras tentaram adaptar o modelo cadastral que já estava sendo empregado, para atender a característica multifinalitária proposta nesses eventos. No entanto, uma adaptação nem sempre consegue atingir os objetivos necessários. Não foi diferente nesse caso, muitas limitações apareceram no intuito de inserir o caráter social no cadastro.

Esse fato tornou necessário o estudo de um novo sistema cadastral, que, começando em 1994 pela Comissão 7 da Federação Internacional de Geômetras (FIG), desenvolveu uma visão futura de um cadastro moderno a ser instrumentado nos 20 anos seguintes. O resultado desse trabalho de pesquisa foi denominado Cadastro 2014, que torna mais amplo o registro de dados no cadastro e o transforma em um inventário público metodicamente ordenado de todos os objetos territoriais legais de determinado país ou distrito (ERBA, 2005).

A Figura 2 mostra como foi a evolução das funções do Cadastro.

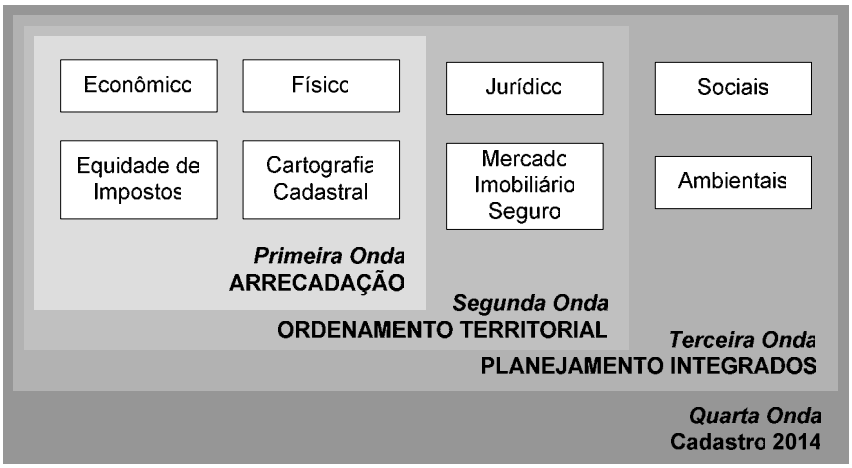


Figura 2: Evolução das visões do Cadastro
Fonte: (ERBA, 2005)

2.1.3 Estrutura do Cadastro

Os Cadastros consistem de textos e mapas cuja organização é baseada em uma parcela territorial e são ligados por um identificador único, como mostra a Figura 3. Esses dados são coletados, armazenados e referenciados (DALE; MCLAUGHLIN, 1990).

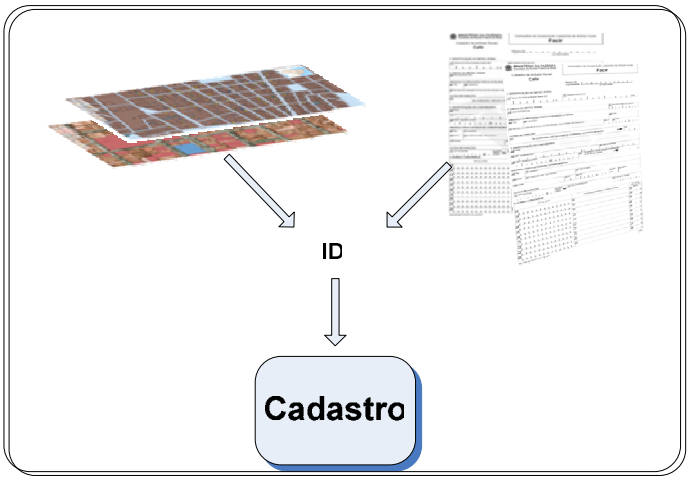


Figura 3: Estrutura Cadastral.
Fonte (SASS, 2012)

O levantamento cadastral de campo produz documentos que descrevem a origem das informações sobre as parcelas cadastrais, e das pessoas a elas relacionadas, exemplo, identificador da parcela, nome do proprietário, área e uso. A carta cadastral contém a representação cartográfica do levantamento sistemático das parcelas territoriais, em grande escala, com elementos físicos e naturais (OLIVEIRA, 2010). De forma geral, as etapas de execução do Cadastro, para a geração dos documentos definidos pela Portaria 511, seriam a de planejamento, trabalho de campo e trabalho de escritório.

Na etapa de planejamento, um diagnóstico da situação atual é realizado e um estudo sobre a viabilidade técnica e financeira da implantação do Cadastro. Isso porque, normalmente o Cadastro era voltado somente para a tributação, porém, com as mudanças na filosofia do sistema cadastral e com as necessidades atuais dos gestores públicos, novos dados são inseridos no Cadastro, como dados sobre educação, saúde, meio ambiente, entre outros.

Na etapa de campo é feita a cartografia cadastral e a coleta dos dados sobre as parcelas e seus proprietários. A cartografia cadastral normalmente é feita por empresas privadas, que utilizam o aerolevantamento ou levantamento topográfico. A coleta dos dados é feita por meio do preenchimento do Boletim de Informações Cadastrais (BIC) que consta dos dados sócioeconômicos e de caracterização do imóvel.

A etapa de escritório é responsável pelo processamento, integração e armazenamento dos dados cartográficos e textuais coletados. (LOCH; ERBA, 2007).

Para dar suporte a todas as etapas do Cadastro, os sistemas cadastrais utilizam recursos de *hardwares*, *softwares*, pessoas e redes. O Sistema de Informação Territorial (SIT) é o sistema que auxilia no armazenamento e gerenciamento dos dados coletados. No Brasil, o

termo SIT não é muito utilizado, embora em outros países ele seja bastante conhecido quando integra o Cadastro ao Registro de Imóveis (AMORIM; SOUZA; YAMASHITA, 2008). Porém, os Art. 4º e 5º, da Portaria 511, dizem que, quando os dados do CTM estão relacionados aos dados do Registro de Imóveis, formam um Sistema de Cadastro e Registro Territorial (SICART), e acrescentando dados de cadastros temáticos cria-se o SIT.

2.1.4 BIC (Boletim de Informações Cadastrais)

O Boletim de Informações Cadastrais (BIC), considerado o principal documento do Cadastro Territorial Urbano tem como função registrar os dados técnicos e informações cadastrais de cada um dos elementos levantados em campo. A partir das informações do BIC, são gerados produtos do cadastro como a Planta de Referência Cadastral, Planta Cadastral, Planta de Valores Genéricos, entre outros. O processo de elaboração do BIC, geralmente é dividido em 2 partes: dados geométricos e dados descritivos, como mostra a Figura 4.

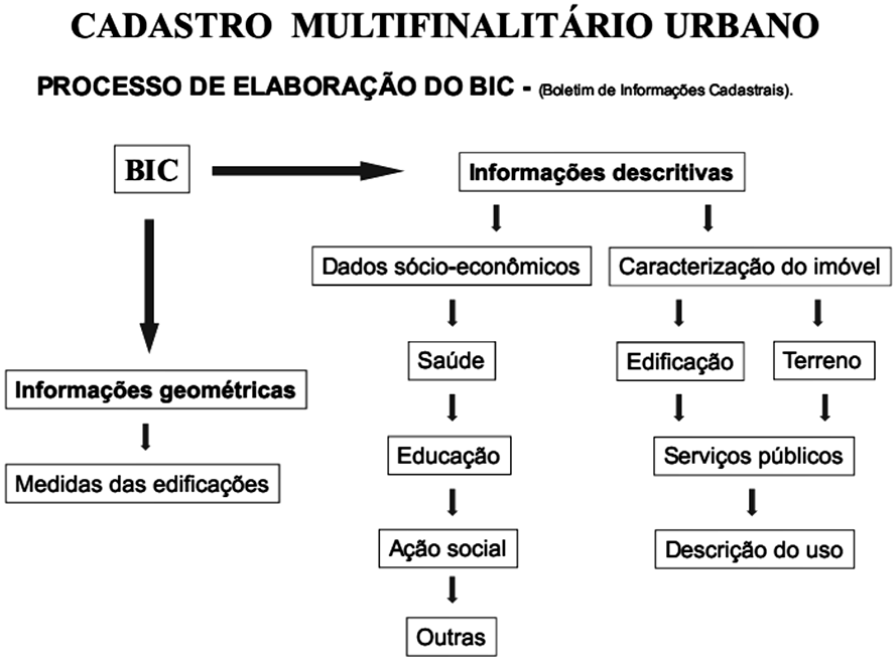


Figura 4: Processo de Elaboração do BIC

Dentre os dados que constam no BIC, alguns são listados abaixo:

- Inscrição cadastral: (Setor, Quadra, Lote e Fração Ideal, SSQQLLFF) campo chave para vinculação ao Banco de Dados;
- Dados de localização: dados que tratam da localização do imóvel (logradouro, número, bairro, CEP);

- Dados do proprietário: Nome, RG, CPF etc;
- Dados da construção: aspectos internos (piso, forro, revestimento interno) e externos da edificação (cobertura, pintura, revestimento externo, estrutura);
- Dados do terreno: dados sobre a topografia, ocupação, forma etc;
- Dados de serviços urbanos: água, esgoto, energia elétrica, limpeza pública, telefonia etc;
- Dados sócioeconômicos: dados sobre saúde, educação, emprego etc;

Deve ficar claro que cada município tem seus gestores e particularidades distintas, portanto cada BIC pode conter alguns campos diferentes um dos outros. Um exemplo de formulário convencional é mostrado na Figura 5 e um formulário para leitora ótica de marcas na Figura 6.

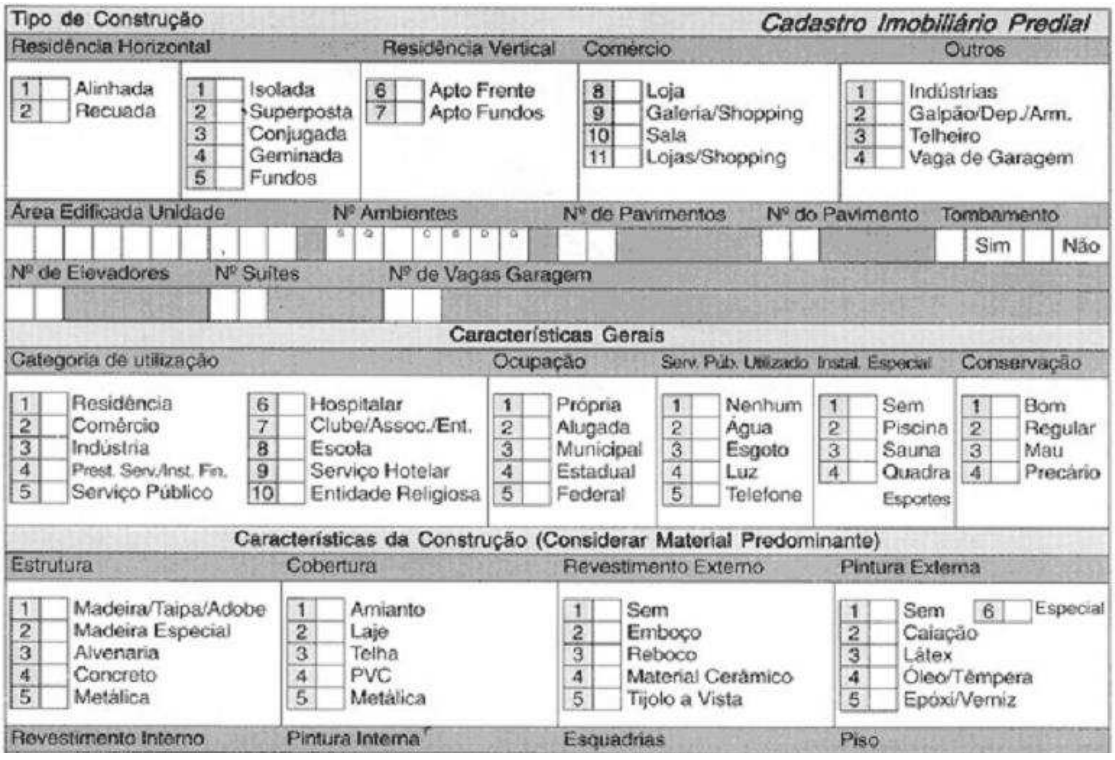


Figura 5: Exemplo de BIC convencional

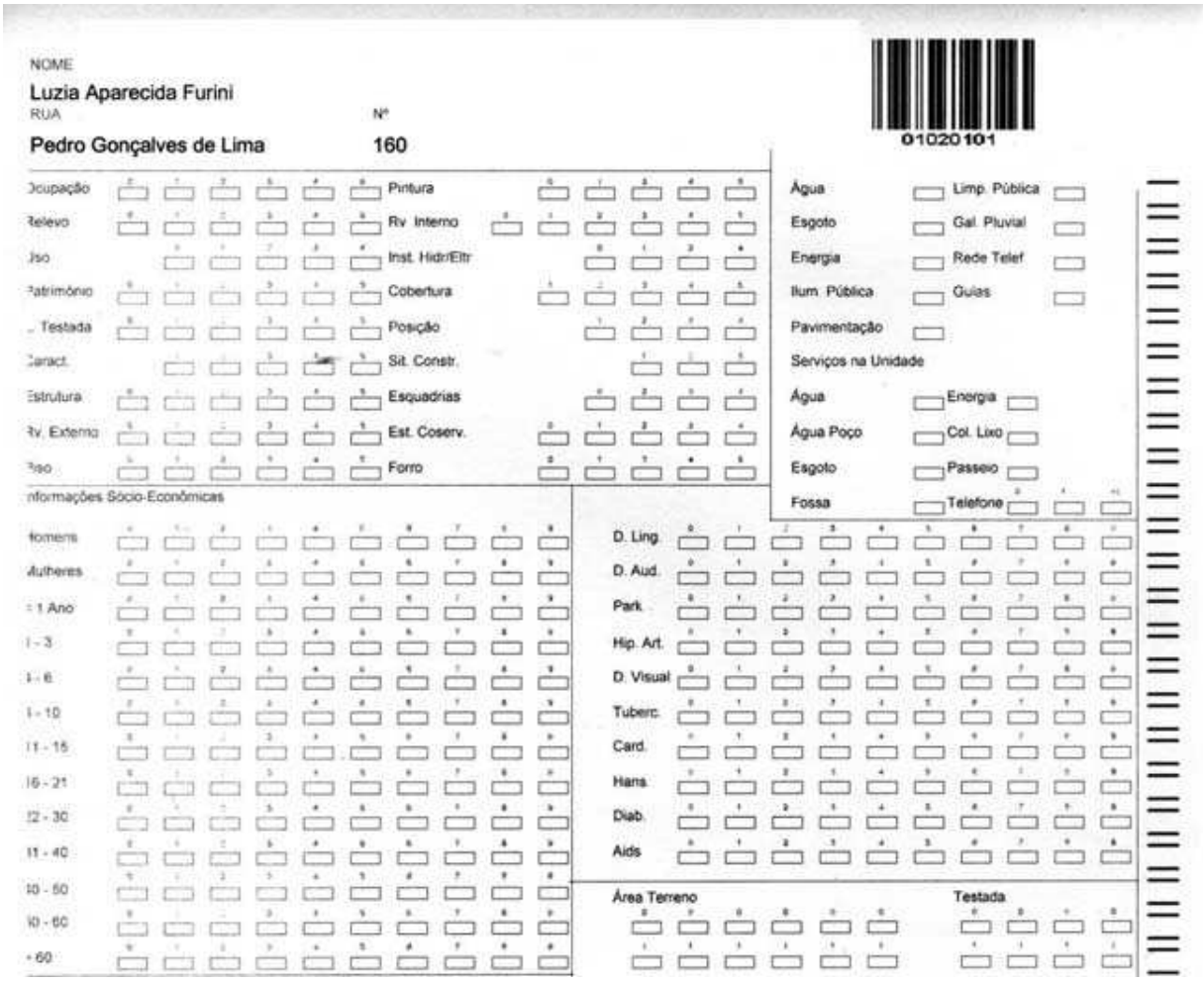


Figura 6: Exemplo de BIC para leitora ótica

De acordo com Pelegrina (2008), a escolha dos campos que serão armazenados os dados cadastrais é considerada como um importante procedimento prévio à organização de um Cadastro Territorial Multifinalitário. Ele esclarece que “um Cadastro eficaz e consistente começa pela concepção correta do BIC”.

Outro fator não menos importante é a concepção e a construção do banco de dados. A quantidade de variáveis envolvidas no processo deixa claro que o desenvolvimento do modelo de dados deve ser bem criterioso para não apresentar anomalias durante seu uso.

2.2 *Sistema Gerenciador de Banco de Dados (SGBD)*

Armazenar dados está presente em nosso cotidiano há muito tempo, sejam armazenamentos manuais ou informatizados. A evolução tecnológica proporcionou a

informatização do armazenamento de dados agilizando o processo de coleta e manipulação dos mesmos.

Porém, de acordo com Date(2000), vários problemas foram detectados nos primeiros modelos utilizados no processo de armazenamento dos dados, como:

- Inconsistência e redundância de dados – os dados eram mantidos em arquivos diferentes e programas eram escritos para o acesso a esses dados. Novos programadores resultavam em novos programas que precisavam de arquivos de dados com formatos diferentes para acessá-los. Isso resultava em duplicidade e informações desencontradas.
- Dificuldade de acesso aos dados – Caso algum relatório não fosse previsto na construção dos programas, o acesso a uma determinada informação se conseguia somente com a construção de um novo programa.
- Isolamento de dados – os dados eram dispersos em diferentes arquivos que podiam apresentar diferentes formatos, dificultando a escrita de novos programas.
- Integridade de dados – As restrições eram mantidas por meio dos programas, caso alguma restrição sofresse alteração, os programas precisavam ser alterados;
- Problema de atomicidade – as transações executadas sobre os dados necessitam que sejam executadas por completo ou desfeitas por completo. As transações não eram atômicas.
- Anomalias no acesso concorrente – O acesso concorrente aos dados não eram controlados corretamente podendo ocasionar inconsistência nos mesmo
- Segurança – Programas de aplicação eram inseridos no sistema como um todo dificultando o controle de segurança.

Ainda de acordo com Date (2005), todas essas desvantagens eram conhecidas nos modelos de dados chamados sistemas de arquivos. Com o surgimento do conceito de Sistema Gerenciador de Banco de Dados (SGBD), Figura 7, novas evoluções, como os modelos de dados hierárquico e de rede tentaram durante algum tempo solucionar esses problemas, mas a metodologia utilizada mostrou que também continham algumas anomalias e precisavam ser corrigidos.

O modelo de SGBD que revolucionou o mercado e até hoje é usado amplamente em sistemas comerciais é conhecido como modelo relacional. Esse modelo resolveu grande parte

dos problemas apresentados por separar o armazenamento físico dos dados de sua representação conceitual e prover uma fundamentação matemática para os bancos de dados (ELMASRI, 2005).

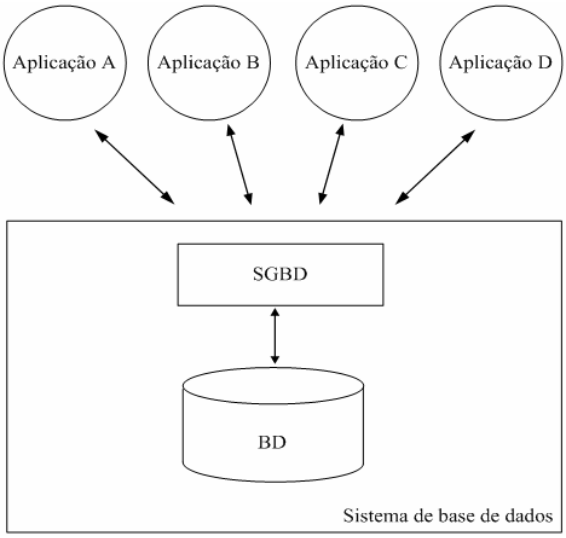


Figura 7: Representação simplificada de um Sistema de Banco de Dados (ELMASRI, 2005)

2.2.1 SGBD Relacional

Em 1970, um matemático britânico, pesquisador da IBM, Edgar Frank Codd publicou um artigo chamado "*Relational Model of Data for Large Shared Data Banks"* no qual aplicou conceitos de um ramo da matemática chamado álgebra relacional para resolver os problemas de armazenamento de grandes quantidades de dados (KROENKE, 1998).

O modelo relacional sugere a representação do mundo real por meio de um conjunto de tabelas bidimensionais (linhas X colunas), também chamadas de relações, para representar tanto os dados como o relacionamento entre eles, como mostra a Figura 8. Cada tabela possui colunas denominadas de atributos ou campos, e linhas que são chamadas de registros ou tuplas. Além disso, propõe também um conjunto de operações para manipulação dessas tabelas.

Date (2000) coloca que um sistema de banco de dados baseado no esquema relacional possui:

1. Aspecto Estrutural – os dados são percebidos pelos usuários como tabelas;
2. Aspecto de integridade – é realizado por meio das restrições impostas pelo modelo, são elas:

a. Restrição de domínio ➔ O valor de cada atributo deve ser um valor atômico do domínio daquele atributo ou um valor nulo. Por atômico entende-se que cada valor no domínio é indivisível no que diz respeito ao modelo relacional. Um domínio nada mais é do que um tipo de dado especificado para cada atributo (ELMASRI; NAVATHE, 2005);
b. Restrição de chave ➔ Cada relação do banco de dados é composta por um conjunto de tuplas, na qual, todas as tuplas dessa relação devem ser distintas, isto é, duas tuplas não podem ter a mesma combinação de valores para todos os seus atributos. Essa restrição é obedecida pelo conceito de chave primária, que, por definição, é o conjunto de um ou mais atributos que, tomados coletivamente, permite identificar de maneira única um registro dentro de uma relação. Nenhum valor de chave primária pode ser nulo (ELMASRI; NAVATHE, 2005) (SILBERSCHATZ, 2006).
c. Restrição de integridade referencial ➔É usada para manter a consistência entre as tuplas que se relacionam em duas relações. Essa restrição é obedecida pelo conceito de chave estrangeira, que, por definição é o conjunto de um ou mais atributos que faz referência por meio da chave primária a outra relação e necessita satisfazer a duas regras básicas: a chave estrangeira deve referenciar sempre uma tupla existente na relação de origem por meio da chave primária e ter o mesmo domínio desta, ou pode ser nula (ELMASRI; NAVATHE, 2005).

3. Aspecto manipulativo – essas tabelas possuem operadores que possibilitam sua manipulação para propósito de busca de dados, como por exemplo: restrição, projeção e junção.

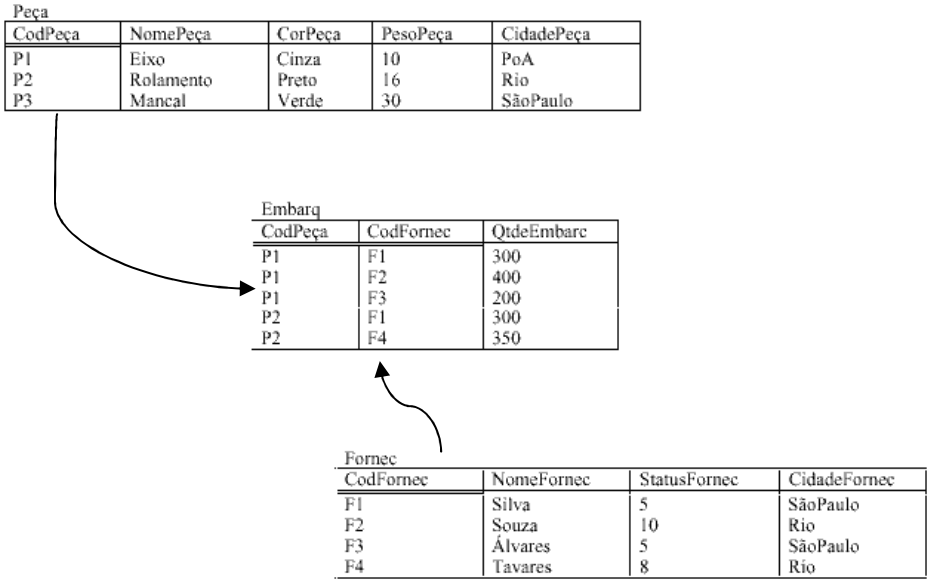


Figura 8: Exemplo de tabelas do modelo relacional

Baseada nos comandos da álgebra relacional e na teoria do pesquisador Edgar Frank Codd, uma linguagem de consulta foi desenvolvida e adotada como padrão na manipulação de dados em SGBD relacionais. Inicialmente chamada de SEQUEL, a Linguagem de Consulta Estruturada (Structure Query Language – SQL) foi concebida e desenvolvida pela IBM (DATE, 2000).

Além dos comandos básicos vindos da álgebra relacional, como UNIÃO, IINTERSEÇÃO, DIFERENÇA e PRODUTO CARTESIANO, novos comandos foram implementados e inseridos para que a SQL se tornasse padrão entre todos os SGBDs relacionais. O Instituto Nacional Americano de Padrões (American National Standards Institute – ANSI) é o comitê responsável pela padronização dessa linguagem (RAMALHO, 1999).

Ainda de acordo com Ramalho (1999), SQL pode ser dividida em 3 partes:

1. DDL (Linguagem de definição de dados) – responsável pelos comandos de criação da estrutura e do próprio banco de dados. Exemplo: *Create Table, Create Index, Create View*.
2. DML (Linguagem de manipulação de dados) – responsável pelos comandos que executam consultas e alterações nos dados. Exemplo: *Select, Update, Delete.*
3. DCL (Linguagem de controle de dados) – comandos responsáveis pela segurança dos dados. Exemplo: *Grant, Revoke*.

Pouco depois do surgimento do modelo relacional foi apresentado o modelo entidade-relacionamento (ER) para projetos de banco de dados. Proposto pelo Dr. Peter Chen em 1976, possibilita ao projetista concentrar-se apenas na utilização dos dados, sem se preocupar em considerar armazenamento e eficiência (CHEN, 1990).

O modelo de entidade-relacionamento representado pelo diagrama (E-R) quando criado, atendia tanto o modelo relacional como os modelos de dados hierárquicos e de rede, porém com a prosperidade do modelo relacional, ficou cada vez mais forte essa união entre o modelo relacional e o modelo entidade-relacionamento (CHEN, 1990).

Com todas essas características, os SGBDs relacionais se tornaram importantes aliados nas aplicações de *softwares* para análise exploratória de dados. Em virtude da utilização de SGBDs em grande parte serem para processamento de dados operacionais, sugere-se a criação de um repositório separado, denominado Data Warehouse(DW), considerado a fonte de dados ideal para qualquer outro tipo de ferramenta analítica. Possuir um DW não é condição obrigatória para isso, mas diminui o caminho a ser percorrido, uma vez que grande parte do tempo que deveria ser gasto durante as etapas de processamento (seleção, preparação e limpeza) dos dados é reduzido drasticamente.

2.3 *Apoio a Decisão*

Sistemas de Apoio a Decisão são sistemas que auxiliam na análise de informações do negócio. Tem como objetivo ajudar a administração a definir tendências, apontar problemas e tomar decisões inteligentes. A ideia básica é coletar dados operacionais do negócio e reduzi-los a uma forma que possam ser usados para análise do comportamento do negócio e modificar seu andamento de maneira inteligente (DATE, 2000).

Esses sistemas estão tradicionalmente associados a três tecnologias: *Data Warehouse (DW), On-Line Analytical Processing (OLTP) e Data Mining (DM)*. Um *DW* é considerado um repositório único, limpo, integrado e orientado por assunto que permite o armazenamento de informações relevante para a tomada de decisão. *OLTP* realiza uma análise multidimensional permitindo examinar as informações armazenadas no banco sob diferentes perspectivas. E *DM*, objetivo desse trabalho, busca, por meio da execução de algumas etapas, aplicar algoritmos de exploração de dados na identificação de padrões, modelos, relacionamentos etc. (SANTOS; RAMOS, 2006).

O resultado do processo de *DM* pode ser apresentado utilizando técnicas de visualização, geralmente interativa, visando auxiliar a análise e compreensão de um conjunto de dados por meio de representações gráficas.

2.3.1 Dado, Informação e Conhecimento

A evolução na manipulação dos dados, gerando informações e mais recentemente, conhecimentos, tem se destacado como fator de competitividade em diferentes tipos de organização. O gerenciamento desses recursos informacionais subsidia várias atividades melhorando o planejamento estratégico e o processo de tomada de decisão na organização (FREITAS, 2001).

Segundo Côrtes (2007), “dados, são sucessões de fatos brutos, que não foram organizados, processados, relacionados, avaliados ou interpretados, representando apenas partes isoladas de eventos, situações ou ocorrências”.

A informação, componente importante no processo decisório, é formada pelo tratamento do dado. Quando esses dados passam por algum tipo de relacionamento, análise, interpretação ou classificação, gera-se a informação (CÔRTES, 2007).

Porém, um novo componente foi inserido, a contextualização da informação. Quando a informação gerada é introduzida em um determinado contexto, gera-se o conhecimento.

2.3.2 KDD

De acordo com Fayyad Piatetsky-Shapiro e Smyth, (1996a): “Extração de Conhecimento em Bases de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, embutidos nos dados”.

Segundo Cortês (2003), o processo de construção do conhecimento envolve a transformação dos dados em informação e conhecimento propriamente dito. Os dados são a matéria bruta para este processo, guardam os aspectos dos fenômenos que estão sendo estudados. A informação é resultado de um processamento executado nesses dados e o conhecimento é um conjunto de argumentos e explicações que interpretam o conjunto de informações.

O processo de descoberta de conhecimento em bases de dados, responsável por analisar, compreender e extrair padrões de grandes volumes de dados é, por muitos autores

denominado de *Knowledge Discovery in Database (KDD)* (REZENDE, 2005). A Figura 9 mostra as etapas que compõe o processo de KDD.

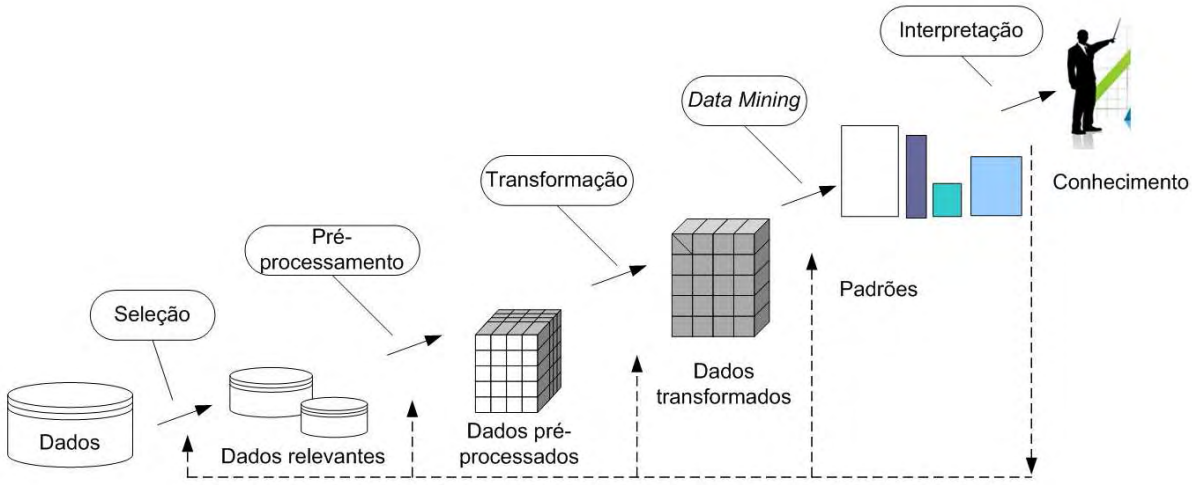


Figura 9: Uma visão geral das etapas que compõe o processo KDD
Fonte: Adaptado de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a)

De acordo com Han, Kamber e Pei (2005) esses passos podem ser detalhados da seguinte maneira:

1. Seleção ➔ quais dados relevantes para a tarefa de pré-processamento são selecionados do Banco de Dados;
2. Pré-Processamento ou Limpeza ➔ remover ruídos e inconsistências de dados;
3. Transformação ➔ na qual os dados são transformados de forma adequada para mineração;
4. *DM* ➔ processo essencial no qual algoritmos são aplicados seguindo certa ordem para extrair padrões de dados;
5. Interpretação ➔ identificação e interpretação de padrões válidos, bem como sua apresentação para tomada de decisão.

2.3.3 Data Mining

A etapa de *DM*, considerada a etapa mais importante do processo de *KDD* tem como objetivo a aplicação de algoritmos específicos para extração de padrões (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a).

Todo processo de *DM* é realizado em função de um domínio específico e dos repositórios de dados referentes a esses domínios. Para que o *DM* seja executado

eficientemente, é necessário que os dados estejam estruturados de forma a serem consultados e analisados adequadamente (REZENDE, 2005). Como foi dito anteriormente, essa estrutura adequada se dá através da criação do *DW*.

Ainda de acordo com Rezende (2005) outro componente importante é a interação entre as diversas classes de usuários existentes na execução do processo. Esses usuários podem ser divididos em três partes:

1. Especialistas do domínio ➔ usuário com amplo conhecimento do domínio da aplicação e que fornece apoio à execução do processo;
2. Analista ➔ usuário especialista e responsável pelo processo de extração de conhecimento. Conhece profundamente as etapas do processo;
3. Usuário Final ➔ Representa os analistas de negócio, ou seja, os atores que usam o resultado do *DM* para tomar decisões no ambiente empresarial. Esse usuário não precisa ter conhecimento aprofundado das etapas do processo.

A Figura 10 mostra as etapas do processo de *DM* adotado por Rezende (2005), e em seguida a descrição de cada uma dessas etapas. A princípio pode parecer que o processo é idêntico mudando somente a nomenclatura das etapas. Porém Rezende (2005) insere a característica de reprocessamento, ou seja, caso o resultado do pós-processamento não alcance objetivos relevantes, esse resultado pode novamente ser submetido à etapa de pré-processamento e assim por diante, até alcançar resultados de relevância ao gestor.

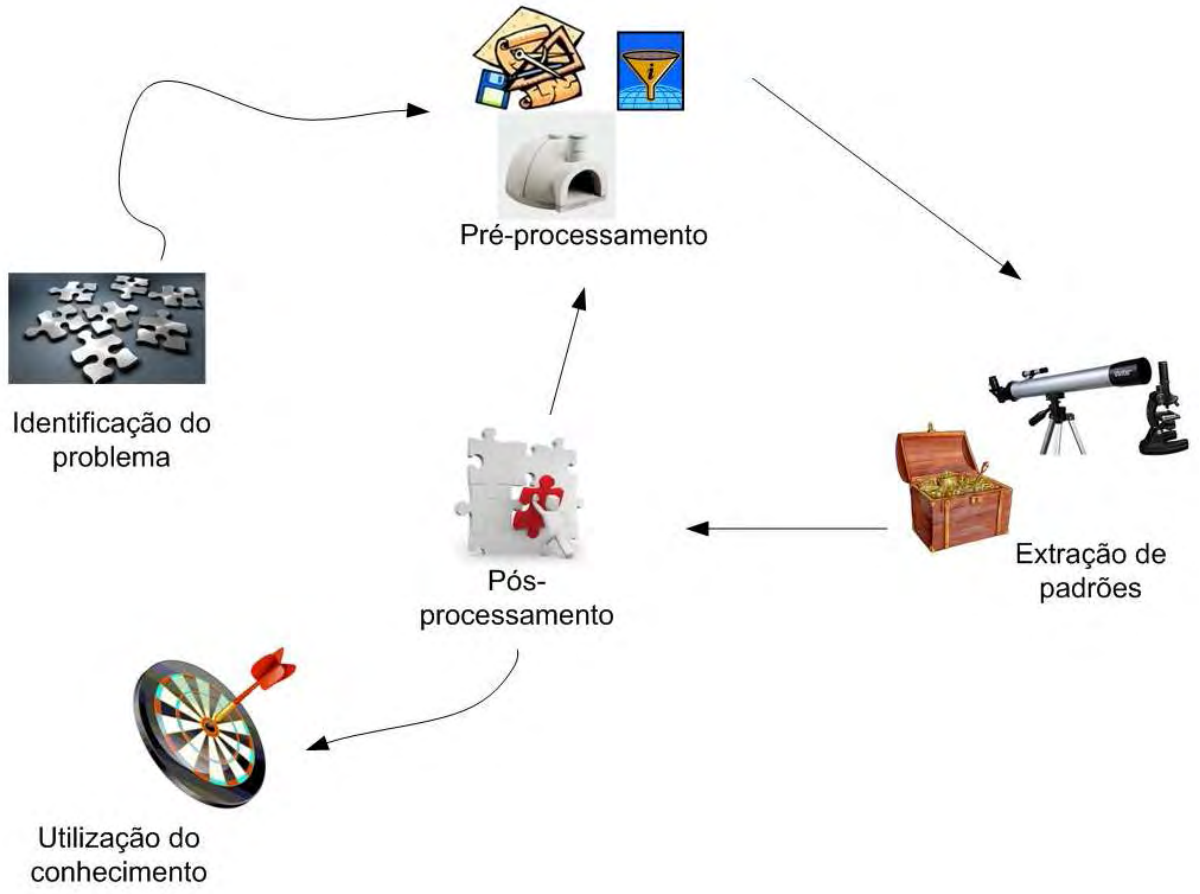


Figura 10: Etapas do processo de DM
Fonte: Adaptado de (REZENDE, 2005)

- Identificação do problema ➔ Nessa etapa detalha-se o domínio da aplicação e definem-se os objetivos e metas a serem alcançadas no processo de *DM*. Algumas questões precisam ser respondidas, tais como:
  - Quais as principais metas do processo?
  - Quais critérios de desempenho são importantes?
  - O conhecimento extraído deve ser compreensível pelos humanos ou esse resultado pode servir de repositório para um novo processo?
  - Qual deve ser a relação entre simplicidade e precisão nos resultados obtidos?
- Pré-processamento ➔ O processo de *DM* não pode ser aplicado em um Banco de Dados comum, os dados não estão preparados para a aplicação dos algoritmos, podendo causar problemas de instabilidade no SGBD. É necessária a aplicação de métodos para tratamento desses dados:
  - Extração e Integração: Unificação dos dados, formando uma única fonte de dados já que eles podem ser encontrados em diversas fontes heterogêneas

como textos, planilhas, Bases de Dados diversas, entre outros. Geralmente essa unificação se dá por meio da criação de um *DW*;

- Transformação: Adequar os dados unificados para serem utilizados nos algoritmos de extração de padrões. Essas transformações são extremamente importantes no caso de aplicações que envolvam séries temporais, como predições de crescimento populacional;
- Limpeza: Mesmo transformados, esses dados foram armazenados muitas vezes de forma manual, ou seja, através da digitação de um usuário final. Com isso, há grande chance de existir ruídos e inconsistências nesse preenchimento. A limpeza objetiva eliminar esses ruídos e inconsistências;
- Seleção e redução de dados: Algumas vezes podem existir certas restrições que inviabilizam o processo em todo repositório. É o caso do espaço em memória disponível e do tempo de processamento. Quando isso acontece, sugere-se uma redução nos dados antes de iniciar a busca por padrões.

- Extração de Padrões ➔ Etapa direcionada ao cumprimento dos objetivos definidos na identificação do problema. Aqui é realizada a escolha das tarefas a serem empregadas e a configuração e execução de uma ou mais técnicas para extração de conhecimento.

  As técnicas podem ser consideradas ferramentas utilizadas para atender aos propósitos do *DM*. De acordo com Harrison (1998) não existe uma técnica que resolva todos os problemas de *DM*. Cada propósito exige uma técnica determinada que por sua vez, tem vantagens e desvantagens na sua aplicação. Para facilitar a escolha, leva-se em conta primeiramente a adequação ao problema e a familiaridade com a técnica escolhida.

  As tarefas, também chamadas de funcionalidades, são a maneira como os resultados serão apresentados. Dependendo da técnica, os algoritmos correspondentes são escolhidos para sua execução. A Figura 11 mostra as interações entre, técnicas, tarefas (funcionalidades) e algoritmos.

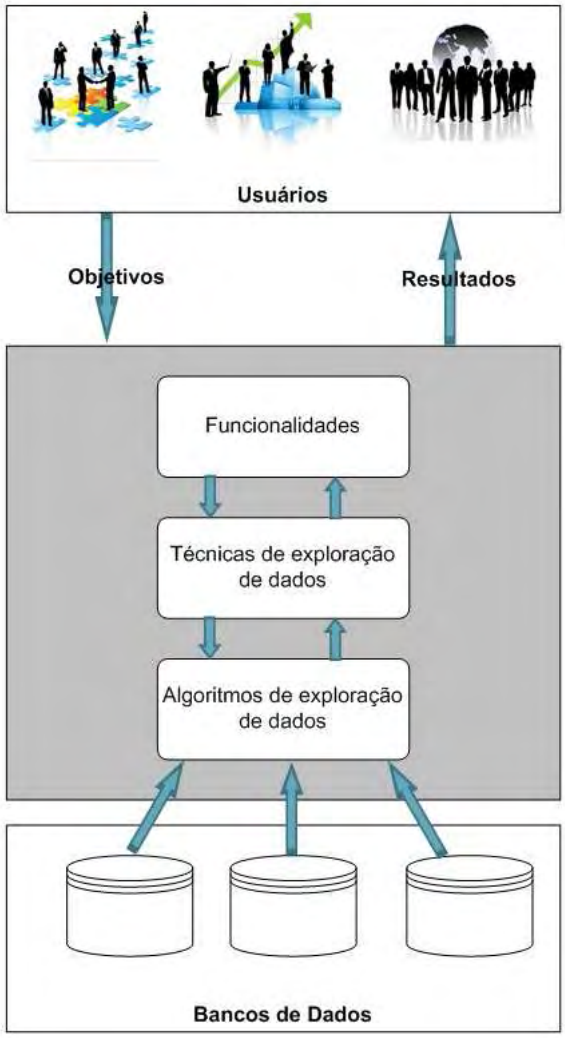


Figura 11: Interatividade entre as tarefas, técnicas e algoritmos de DM.

- Pós-Processamento ➔ Nessa etapa, o conhecimento extraído é analisado para verificação de sua relevância. Caso ele não seja de interesse do usuário ou não cumpra com os objetivos propostos, o processo de extração pode ser repetido, ajustando-se os parâmetros ou melhorando o processo de escolha dos dados para obter resultados que possam ser interpretados com mais qualidade.

O conhecimento do domínio auxilia em todas as outras etapas do processo: no pré-processamento, ajudando na escolha do melhor conjunto de dados para se extrair os padrões; na extração de padrões, ajudando na escolha de um critério de preferência entre os modelos gerados ou mesmo na geração de um conhecimento inicial a ser fornecido como entrada do algoritmo de mineração; no pós-processamento, ajudando a avaliar os padrões extraídos com a execução dos algoritmos; e na utilização do conhecimento, sabendo aproveitar os resultados obtidos na tomada de decisão.

Pela Figura 10, é possível perceber que a etapa de pré-processamento é realizada antes da etapa de extração de padrões, porém, em virtude do processo ser iterativo, algumas atividades de pré-processamento podem ser realizadas novamente após a análise dos padrões encontrados.

Segundo Dias (2002), a aplicação de técnicas de *DM* já está presente em várias áreas, como por exemplo:

- Marketing ➔ Aplicadas com o objetivo de descobrir preferências do consumidor bem como padrões de compra. Com o resultado, procura-se realizar marketing direto de produtos e ofertas promocionais de acordo com o perfil do consumidor;
- Detecção de Fraudes ➔ Alguns tipos de fraudes não precisam de *DM* para serem encontradas porém alguns padrões necessitam de uma análise mais criteriosa, como exemplo a previsão de um inadimplente no pagamento de empréstimo ou mesmo a detecção de padrões de consumo propensos à fraudes em consumidores de energia elétrica;
- Medicina ➔ Buscar padrões de novas doenças, identificação de terapias médicas de sucesso para diferentes doenças, buscar categorizar comportamento de pacientes para previsão de visitas;
- Ciência ➔ Encontrar padrões em estruturas moleculares, dados genéticos, mudanças climáticas;
- Transporte ➔ Determinar escalas de distribuição, analisando padrões de carga entre distribuidores;
- Banco ➔ Análise de padrões no uso de cartões de crédito de maneira fraudulenta, agrupamento de clientes por padrões de gastos;
- Gestão pública ➔ Auxiliar o gestor público com resultados gráficos ou até demonstrados em mapas da caracterização de edificações, crescimento de bairros num período de tempo, previsão de crescimento populacional etc.

2.3.3.1 Tarefas de DM

Muitos autores definem uma quantidade diferenciada de tarefas para *DM*, uns mais, outros menos em suas definições, como mostrado a seguir:

- Classificação, Estimação, Predição, Afinidade em grupos, Agrupamentos (*Clustering*) e Descrição (BERRY; LINOFF, 1997);

- Previsão, Identificação, Classificação e Otimização (ELMASRI; NAVATHE, 1999);
- Descrição e Predição (HAN; KAMBER, 2005);
- Classificação, Regressão, Regras de Associação, Sumarização, *Clustering* e Outras (REZENDE, 2005);
- Classificação, Regressão, Associação, *Clustering* e Sumarização (DIAS, 2002);
- Classificação, Regressão, *Clustering*, Sumarização, Modelagem de Dependências, Análise de Links e Análise Sequencial (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b).

Porém, a maioria deles concorda que essas tarefas sejam classificadas em 2 grandes grupos, como mostra Rezende (2005) na Figura 12.

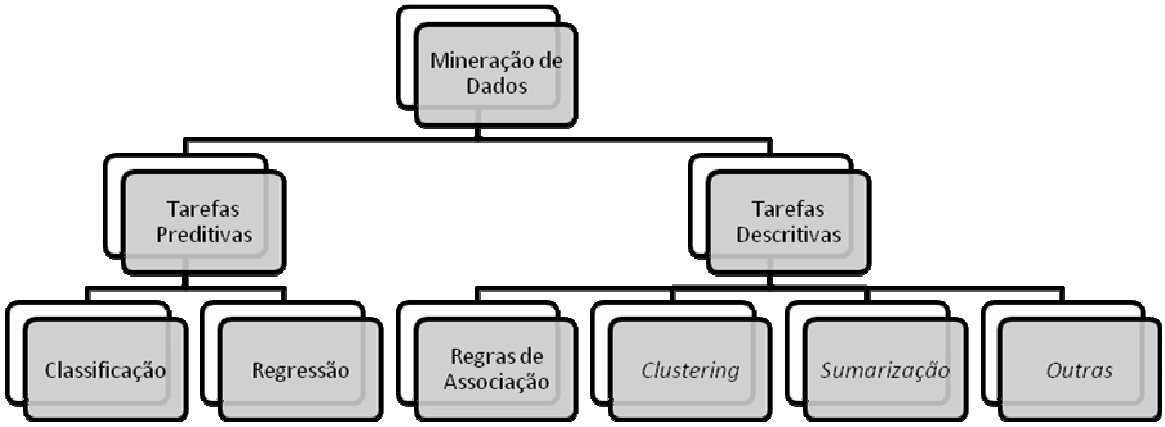


Figura 12: Tarefas de DM
Fonte: Adaptado de (REZENDE, 2005, p. 318)

As tarefas preditivas envolvem atributos de um conjunto de dados para prever o valor futuro de uma variável meta, visando principalmente à tomada de decisão. Já as Tarefas Descritivas procuram padrões interpretáveis pelos humanos, visando o suporte à tomada de decisão (REZENDE, 2005). Essas tarefas são subdivididas em:

- Classificação➔ Categorização de dados em classes. Objetiva descobrir relacionamentos entre um atributo meta, e um conjunto de atributos de previsão. Como exemplo, tem-se: Classificação de pedidos de crédito, Esclarecimento de pedidos de seguro fraudulento e Identificação da melhor forma de tratamento de um paciente (DIAS, 2002). A Figura 13 mostra um exemplo de como seria a saída com a utilização da tarefa de classificação.

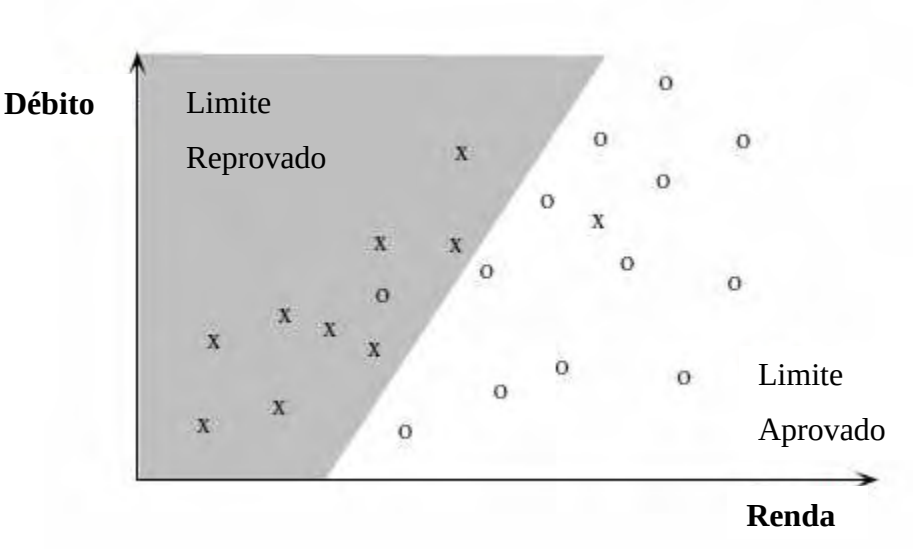


Figura 13: Uma classificação linear simples para um conjunto de dados sobre limites de empréstimos. Adaptado de: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a).

- Regressão ➔ Conceitualmente similar à tarefa de classificação, diferenciando somente na variável meta, que aqui, passa a ser contínua. Como exemplo, tem-se: Estimar o número de filhos ou a renda total de uma família, Estimar o valor em tempo de vida de um cliente, Estimar a probabilidade de que um paciente morrerá baseando-se nos resultados de diagnósticos médicos e Prever a demanda de um consumidor para um novo produto (DIAS, 2002). A Figura 14 mostra um exemplo de como seria a saída com a utilização da tarefa de regressão.

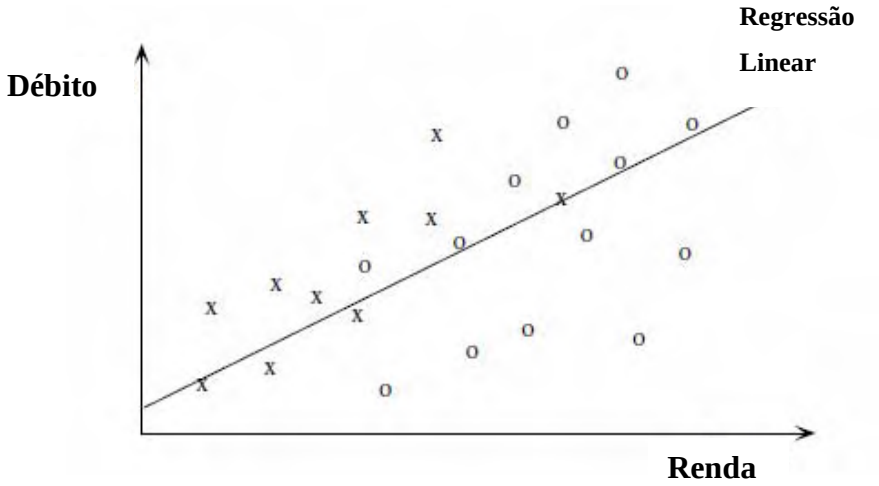


Figura 14: Uma regressão linear simples para um conjunto de dados de empréstimos. Adaptado de : (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a).

- Regras de Associação ➔ De maneira geral, usada para determinar quais itens tendem a serem adquiridos juntos em uma mesma transação. Como exemplo, tem-se: Determinar quais produtos costumam ser colocados juntos em um carrinho de supermercado (DIAS, 2002).
- Segmentação ou *Clustering* ➔ Separa grupos, a princípio heterogêneos em subgrupos mais homogêneos. Como exemplo, tem-se: Agrupar clientes por região de país, Agrupar clientes com comportamento de compra similar (DIAS, 2002). A Figura 15 mostra um exemplo de como seria a saída com a utilização da tarefa de *Clustering*.

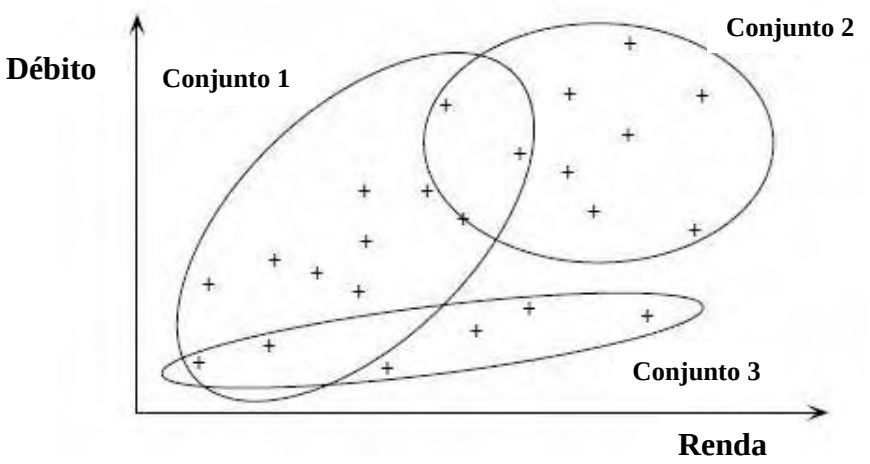


Figura 15: Um Cluster simples para um conjunto de dados de empréstimos separados em 3 grupos. Adaptado de: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a).

- Sumarização ➔ Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados. Como exemplo, tem-se: Tabular o significado e os desvios padrão para todos os itens de dados (DIAS, 2002).
- Modelagem de Dependência ➔ Descreve as dependências mais significativas entre as variáveis. Existem dois níveis de modelos: Estrutural e Quantitativo. O modelo Estrutural especifica quais variáveis são dependentes localmente. Já o Quantitativo especifica o peso das dependências usando algumas escalas numéricas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a).
- Análise de Links ➔ Consiste em extrair correlações dos atributos de um conjunto de dados. Inicia-se com dados que podem ser representados como uma rede e infere conhecimento útil a partir dos nós e links da rede (REZENDE, 2005).

- Análise Seqüencial ➔ Auxilia a detecção de ocorrência de padrões sequenciais num fluxo de identificação de categorias descrevendo as relações (REZENDE, 2005).

As técnicas com seus algoritmos específicos são escolhidas de acordo com as tarefas desejadas do seu *DM*. Alguns exemplos de técnicas para algumas tarefas são mostradas no Quadro 1.

Técnicas	**Descrição**	**Tarefas**	**Exemplos de Algoritmos**
Descoberta de Regras de Associação	Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados	Associação	Apriori, AprioriTid, AprioriHybrid, AIS, SETM e DHP
Árvores de Decisão	Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos	Classificação Regressão	CART, CHAID, C5.0, Quest, ID-3, SLIQ, SPRINT
Raciocínio Baseado em Caso	Baseado no método do vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança	Classificação Segmentação	BIRCH, CLARANS, CLIQUE
Algoritmos Genéticos	Métodos gerais de busca e otimização, inspirados na Teoria da Evolução, onde a cada nova geração, soluções melhores têm mais chance de ter "descendentes"	Classificação Segmentação	Algoritmo Genético Simples, Genitor, CHC, Algoritmo de Hillis, GA-Nuggets, GA-PVMINER
Redes Neurais Artificiais	Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões	Classificação Segmentação	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB

Quadro 1: Exemplos de técnicas e algoritmos para algumas tarefas de *DM*.
Fonte: (DIAS, 2002)

2.3.4 Técnicas de Visualização de Dados

Existem diversas maneiras de representar o conhecimento extraído das bases de dados. Essas representações ajudam a melhorar a compreensão e a interpretação dos resultados gerados pelo processo de Mineração de Dados. Combinando algumas técnicas computacionais com o processo de Mineração de Dados, a visualização de informações permite a apresentação de dados em formas gráficas permitindo ao usuário utilizar sua percepção visual para otimizar o processo de interpretação desses resultados (KEIN, 2002).

A visualização, nos últimos anos, vem se destacando e recebendo fortes contribuições de diversas áreas científicas, como as ciências da computação, psicologia, semiótica, cartografia, artes, entre outras. Sendo assim, sua utilização se torna pertinente em várias aplicações, mas visando sempre um objetivo: utilização da metáfora visual para a representação da estrutura e dos relacionamentos entre os dados (VANDE, 2005).

De acordo com Keim (2002), a exploração de dados combinados com recursos visuais (exploração de dados visuais) visa a inserção do ser humano como parte essencial do processo, aplicando suas habilidades de percepções para a análise de grandes conjuntos de dados disponíveis atualmente.

Assim, ferramentas computacionais capazes de gerar e apresentar resultados de análise de dados por meio visual podem dar apoio aos utilizadores em todo processo de análise exploratório de dados.

Para essa dissertação, foram escolhidas duas ferramentas computacionais capazes de gerar esse tipo de resultado visual aliado a algoritmos de mineração de dados, são elas: Matlab[8] e SODAS[9].

[8] http://www.mathworks.com
[9] http://www.ceremade.dauphine.fr/SODAS/

3 ESTUDO DE CASO – APLICAÇÃO DE ABORDAGENS DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS DE CADASTRO TERRITORIAL MULTIFINALITÁRIO DA CIDADE DE RIBEIRÃO DOS ÍNDIOS - SP

O levantamento cadastral, como o censo demográfico do IBGE, não acontece todos os anos. Porém, quando é levantado, uma grande quantidade de dados é coletada e inserida nas bases de dados cadastrais para sua atualização. A cada levantamento, novos dados podem ser inseridos no BIC dependendo da necessidade do gestor.

Com essa grande quantidade de dados armazenada de tempos em tempos, o processo de extração de informações novas e potencialmente úteis se torna uma tarefa complexa levando-se em consideração a estrutura das bases existentes nas prefeituras brasileiras. Isso acontece basicamente por dois motivos: essas bases não estão preparadas para armazenamento de dados históricos, e a quantidade de variáveis envolvidas é muito grande, dificultando a obtenção de resultados por meio de consultas SQL.

A aplicação de abordagens de descoberta de conhecimento (*KDD*) em bases de dados cadastrais visa preparar os dados para serem analisados, analisá-los e por fim interpretá-los para validação. Essa metodologia envolve várias etapas que serão descritas nas subseções seguintes utilizando-se como área de estudo o município de Ribeirão dos Índios, no Oeste do Estado de São Paulo.

A prefeitura municipal de Ribeirão dos Índios vem fazendo desde 1996 em conjunto com a Universidade Estadual Paulista (UNESP) de Presidente Prudente, levantamentos cadastrais com características multifinalitárias buscando um melhor acompanhamento do desenvolvimento territorial do município.

O município, de acordo com o levantamento do IBGE 2007[10] possui 2187 habitantes em uma área de 197 km^2. Está localizada a oeste do Estado de São Paulo, como mostra a Figura 16.

[10] Disponível em: http://www.ibge.gov.br/cidadesat/painel/painel.php?codmun=354323#, acesso em 30/10/2012.

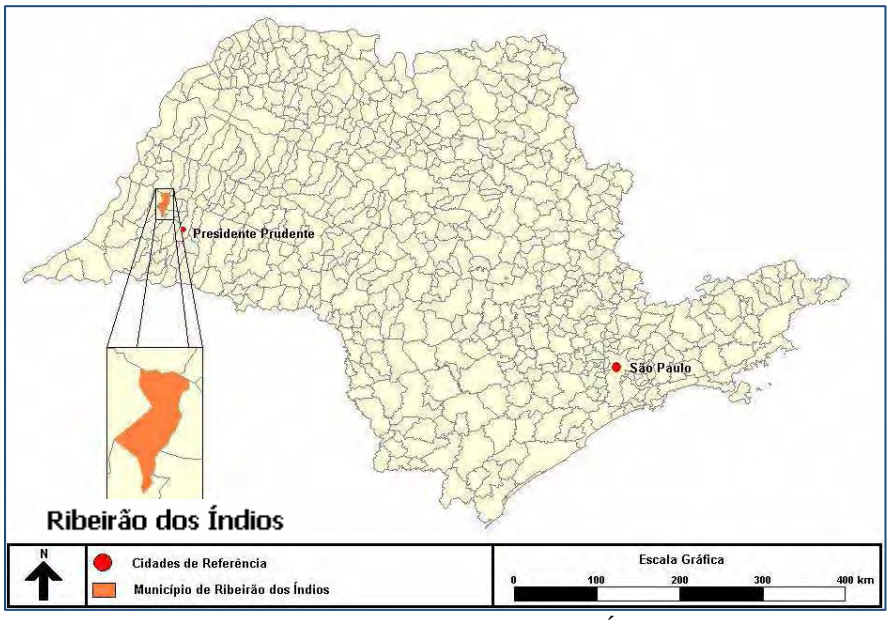


Figura 16: Localização do Município de Ribeirão dos Índios (DINIZ 2004).

3.1 *Levantamentos Cadastrais*

Ao todo foram realizados quatro levantamentos cadastrais, sendo eles nos anos de 1996, 2004, 2010 e 2012. Somente o modelo de dados especificado em 2012 foi previsto o armazenamento de dados históricos, porém não se pode simplesmente abandonar o que foi coletado nos anos anteriores, justamente porque este fato reflete a realidade quando se diz respeito à Sistemas Cadastrais.

As bases anteriores a 2012 estão armazenadas em bancos distintos e precisaram ser preparadas para a obtenção de conhecimentos estratégicos. A construção de um *DW* seria uma solução ideal para esses dados, porém a realidade vista nas prefeituras mostra a dificuldade de realizar tal investimento. Portanto, apesar de estar destacado no texto a importância de um *DW* para a Mineração, optou-se por demonstrar o processo de *KDD* sem a utilização do *DW*.

Toda a preparação dos dados objetivou a unificação das bases e a criação de tabelas virtuais contendo domínios específicos para a execução de ferramentas de extração de informações.

A seguir são apresentados relatos das coletas de 1996, 2004 e 2010 para uma contextualização dos dados armazenados.

3.2 *Conversão e Unificação das bases de dados*

O processo de conversão e unificação das bases de 1996, 2004, 2010 e 2012 teve início com um estudo detalhado da estrutura desses dados. Esse estudo permitiu identificar várias situações de heterogeneidade semântica e estrutural dos dados.

O SGBD PostgreSQL foi escolhido para receber os dados das coletas anteriores a 2012 (2012 já foi projetado para armazenamento no SGBD PostgreSQL) por 2 motivos: ser uma ferramenta considerada *software livre* e ser integrado com ferramentas capazes de espacializar dados oriundos das tabelas relacionais.

Em 1996, começaram os trabalhos no referido município. A base gráfica foi criada a partir de um levantamento topográfico que deu origem ao mapeamento referenciado ao Sistema Geodésico Brasileiro – SGB, uma vez que a Rede de Referência Cadastral Municipal foi implantada de acordo com a NBR – 14166, permitindo que fosse executado o primeiro levantamento cadastral.

Os dados cadastrais coletados em 1996 foram armazenados no banco de dados DBASE. Nesse primeiro levantamento o BIC ainda era manual, porém com a evolução do projeto de CTM no referido município, uma leitora ótica para leitura de BIC foi adquirida e nos próximos levantamentos, os BICs já foram preparados para serem lidos por essa leitora.

A conversão começou com os dados de 1996 sendo carregados no aplicativo DATABASE DESKTOP que está integrado com o BORLAND DELPHI para serem copiados e enviados para o MICROSOFT EXCEL. Após isso, os dados passaram por um processo de correção onde os campos que recebiam respostas do tipo “S” e “N” foram substituídos por valores numéricos “1” e “0” para que ficassem compatíveis com as outras coletas. Em seguida foram importados pelo aplicativo MICROSOFT ACCESS e convertidos para o banco de dados PostgreSQL pelo aplicativo MS ACCESS TO POSTGRES CONVERSION[11].

Em 2004, novo levantamento foi feito, a base gráfica do município foi atualizada e o levantamento cadastral já foi preparado para ser armazenado no MICROSOFT ACCESS e o BIC desenvolvido para ser lido por uma leitora ótica (AMORIM; SOUZA; DALAQUA, 2004). A leitora produz um arquivo texto (.txt) que, após salvo, é enviado ao MICROSOFT EXCEL e importado pelo MICROSOFT ACCESS. As Figuras 17 e 18 mostram respectivamente a leitora e o Diagrama de Entidade Relacionamento construído para o banco.

[11] Aplicato disponível em: http://www.bullzip.com/products/a2p/info.php

Figura 17: Leitora ótica utilizada desde 2004 para leitura dos BICs (AMORIM; SOUZA; DALAQUA, 2004)

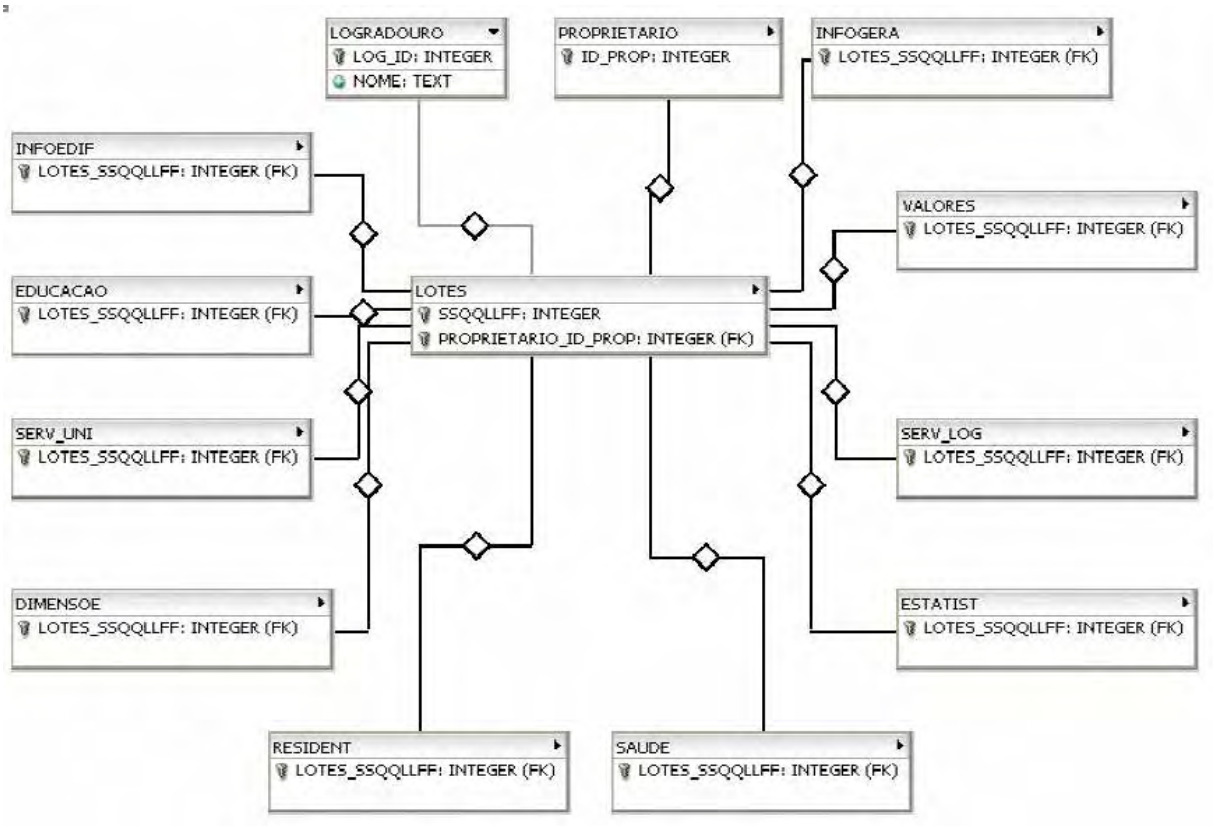


Figura 18: Diagrama de Entidade Relacionamento para o banco de dados de 2004 (AMORIM; SOUZA; DALAQUA, 2004).

Visto que em 2004 os dados já foram armazenados no MICROSOFT ACCESS, foi feito o mesmo processo descrito anteriormente, utilizou-se o aplicativo MICROSOFT

ACCESS TO POSTRGES CONVERSION na conversão dos dados para o banco PostgreSQL.

No levantamento de 1996 a base de dados foi desenvolvida sem um modelo de dados prévio, simplesmente tabelas foram criadas para armazenar os dados coletados. No levantamento de 2004, um planejamento em relação à base foi realizado, porém ocorrendo ainda diversas anomalias na modelagem. Somente em 2010 essas anomalias foram corrigidas.

Um dos problemas foi detectado nas tabelas que contem os dados referentes às informações da edificação. Para cada levantamento novas características referentes a edificação foram inseridas ao BIC e o valor da pontuação dessas características foram alteradas nos levantamentos. Para não comprometer o processo de conversão, e por se tratar de uma análise histórica dos dados, as novas características foram descartadas, mantendo somente as originais e toda pontuação passou por um processo de normalização de valores.

Outro problema detectado foi a geometria da base gráfica, que até então não estava conectada diretamente ao banco de dados. Dentro do aplicativo AUTOCAD, cada geometria dos lotes recebia um código, no caso a chave primária da tabela lote (ssqqllff) e por meio de uma interface SIG com arquitetura DUAL, o desenho era ligado aos dados e conseguia-se a espacialização.

O modelo de arquitetura DUAL foi um modelo apresentado como tentativa de um melhor gerenciamento de dados geográficos, fazendo a integração dos dados operacionais e espaciais por meio de identificadores comuns, porém, armazenando em bases distintas. Essa arquitetura, de acordo com Ferreira (2005), compromete a garantia de integridade entre as partes geométrica e descritiva da representação do objeto geográfico. Essa dificuldade se dá em virtude de existir a possibilidade de atualização dos dados descritivos sem que a estrutura de dados geográficos tenham conhecimento desse fato. Na Figura 19 é mostrado o modelo de arquitetura DUAL.

Figura 19: Modelo de Arquitetura DUAL
Fonte; (FERREIRA, 2005)

Ainda de acordo com Ferreira (2005) a arquitetura integrada unificou os dados descritivos e espaciais tudo em um SGBD. Como principal vantagem, está a utilização dos recursos do SBGD para controle e manipulação de objetos espaciais. Com isso a manutenção da integridade entre os dados espaciais e descritivos é feita pelo SGBD. A Figura 20 mostra o modelo da arquitetura INTEGRADA.

Figura 20: Modelo de arquitetura INTEGRADA
Fonte; (FERREIRA, 2005)

Em 2010 um novo levantamento cadastral foi feito e a base gráfica foi conectada ao banco de dados por meio de uma arquitetura integrada. A modelagem do banco desse ano já foi preparada para solucionar os problemas de modelagem e projeto anteriormente mencionados, com isso novas tabelas foram adicionadas ao modelo. Parte do Diagrama de Entidade Relacionamento mostrando essa solução é mostrado na Figura 21. Apesar de ter que fazer todo o processo de conversão de arquivo texto para o MICROSOFT EXCEL e depois para o MICROSOFT ACCESS, o banco já foi criado no SGBD PostgreSQL para receber os dados dessa coleta.

O levantamento cadastral de 2012 vem com uma proposta de modelagem de banco de dados diferente. Apesar de ainda estar em desenvolvimento em outros projetos de pesquisa, o objetivo a partir desse ano é preparar o banco de dados para armazenamento de dados históricos buscando conhecer a “dinâmica” da expansão e das transformações ocorridas com as parcelas e com os dados temáticos ligados a ela. A proposta é o desenvolvimento de um modelo espaço-temporal.

Um modelo espaço-temporal é composto por três componentes: atributo, espaço e tempo que ajudam a responder questões do tipo: O que? Onde? e Quando? Cada um desses componentes determina uma categoria de dimensão ao longo da qual os valores são medidos (WORBOYS, 1995).

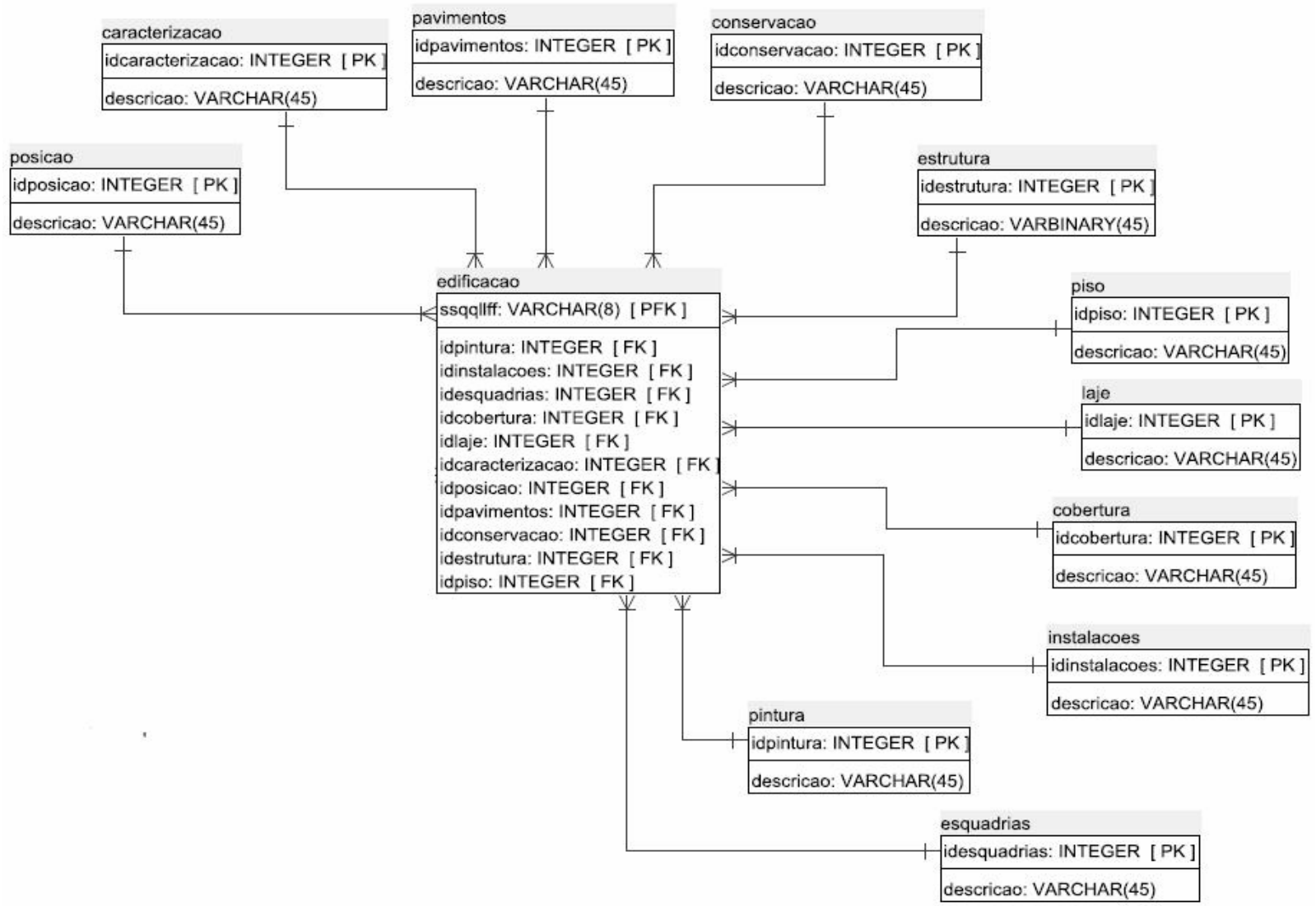


Figura 21: Parte do Diagrama de Entidade Relacionamento do banco de 2010

3.3 *Data Mining*

Após as devidas conversões, conseguiu-se armazenar todas as coletas no SGBD PostgreSQL. Optou-se por trabalhar com uma única base de dados contendo todas as tabelas dos quatro levantamentos. Porém guardou-se também uma cópia de cada base separadamente para possíveis utilizações futuras. Após uma análise dos dados unificados, percebeu-se a ausência de várias informações nas tabelas do levantamento de 1996. Esse problema foi esclarecido com a explicação de que, por se tratar do primeiro levantamento, vários moradores ficaram receosos em fornecer determinadas informações como por exemplo renda familiar. Com isso, para esse estudo de caso, apesar das tabelas do levantamento de 1996 estarem juntas na base unificada, optou-se por trabalhar somente com os levantamentos de 2004, 2010 e 2012.

3.3.1 Definição do domínio do problema

A partir daí começa o processo de *DM* proposto por Resende (2005). Primeiramente foi necessário delimitar um domínio de problema a ser analisado. Foram levantados alguns questionamentos cujas respostas possivelmente seriam de interesse do gestor municipal. Para esse estudo de caso, três questionamentos foram submetidos ao processo:

1. Buscar uma relação da renda familiar com os seguintes dados: padrão construtivo, área construída e educação, para os levantamentos de 2004 e 2010;
2. Acompanhar a evolução de uma determinada patologia visando encontrar um padrão nos resultados, para os levantamentos de 2004, 2010 e 2012;
3. Procurar alguma ligação entre a evolução de patologias, relacionada ao aumento da faixa etária da população, para os levantamentos de 2004 e 2012;

3.3.2 Pré-Processamento

3.3.2.1 Primeiro Caso

Para conseguir obter resultados neste primeiro caso, o primeiro passo foi separar os dados de acordo com o domínio do problema levantado, ou seja, o histórico de renda familiar,

histórico do padrão construtivo, histórico do tamanho da edificação e quantidade de estudantes em escolas públicas e particulares.

O fato das bases de dados terem sido todas convertidas para o SGBD PostgreSQL não implicou em códigos SQL simples para a seleção desses dados. Em virtude de cada levantamento ter passado por melhorias e correções, dados de levantamentos diferentes possuíam tipos de valores diferentes, valores de referência diferentes e precisavam ser padronizados para poderem ser utilizados nas análises.

Alguns ajustes tiveram que ser feitos para que tabelas virtuais fossem criadas com valores compatíveis, etapa conhecida como pré-processamento. Um exemplo disso foi a pontuação, referente ao somatório das características do padrão construtivo, de cada edificação estar com valores diferentes nos 3 levantamentos, com isso precisou-se normalizar esses valores para que eles representassem totalizações equivalentes. Com isso, para obter uma comparação histórica, optou-se por utilizar somente os itens presentes em todos os levantamentos. Os quadros a seguir mostram os itens pontuados nos levantamentos de 2004 e 2010 e os itens utilizados para análise, consecutivamente.

BD2004	**Valor**	**BD2010**	**Valor**
ESTRUTURA	5	ESTRUTURA	5
		PAVIMENTOS	4
VER_EXTERNO	5		
PISO	5	PISO	5
FORRO	5	LAJE/FORRO	5
REVESTIMNENTO INTERNO	5	REVESTIMNENTO INTERNO	5
PINTURA	5	PINTURA	5
INSTALAÇÃO HIDRÁULICA/ELÉTRICA	4		
		INSTALAÇÕES REDIAIS	5
COBERTURA	5	COBERTURA	4
SITUAÇÃO DA CONSTRUÇÃO	2		
ESQUADRIAS	4	ESQUADRIAS	4
RODAPES	2		
ESTADO DE CONSERVAÇÃO	4	ESTADO DE CONSERVAÇÃO	3
Total	51		47

Quadro 2: Atributos da edificação coletados em 2004 e 2010

BD2004	Valor	BD2010	Valor
"ESTRUTURA"	5	"EstruturaIdEstrutura"	5
"PISO"	5	"PisoIdPiso"	5
"FORRO"	5	"LajeIdLaje"	5
"REVEST_INTERNO"	5	"RevestimentoIdRevestimento"	5
"PINTURA"	5	"PinturaIdPintura"	5
"COBERTURA"	5	"CoberturaIdCobertura"	4
"ESQUADRIAS"	4	"EsquadriasIdEsquadrias"	4
"EST_CONSERVACAO"	4	"ConservacaoIdConservacao"	3
Total	38		36

Quadro 3: Atributos da edificação de 2004 e 2010 utilizados para análise histórica

Para se chegar a uma tabela virtual que contivesse os dados relativos ao domínio desse questionamento levantado, alguns passos foram executados, são eles:

1. Gerar o somatório dos pontos para o padrão construtivo de 2004 e 2010 somente das características escolhidas, bem como normalizar esses valores para que ambos os levantamentos tivessem totalizações equiparadas;
2. Tendo como referência o levantamento mais atual, foram eliminados os registros de 2004 que não foram levantados também em 2010. Aqui também foram eliminados registros com valores nulos.
3. Para os dados da área construída, o levantamento de 2004 precisou ser convertido de valores decimais para inteiros visto que o levantamento de 2010 estava com valores textuais impossibilitando sua conversão para valores decimais.
4. Outro passo foi a unificação da área construída com a área de dependência que em 2010 estavam em campos separados.
5. Para a renda familiar bastou a unificação dos dados de 2004 e 2010 comparando os dois levantamentos para exclusão de registros nulos e dos que não constavam em ambos levantamentos.
6. Para os dados de educação foram somados a quantidade de estudantes do ensino fundamental, ensino médio e ensino superior tanto da Instituição Pública como da Particular.
7. Tanto o identificador dos lotes bem como o campo que define a geometria de cada registro, foram mantidos os do levantamento de 2010.

8. Finalizou-se essa tarefa de duas maneiras: a primeira somente com o resultado desses passos, unificando tudo dentro de uma tabela virtual como mostra a Figura 22 e a segunda com a inserção de alguns campos textuais para que ferramentas de mineração distintas pudessem ser utilizadas, como mostra a Figura 23.

Os campos textuais inseridos foram para o processamento dos dados na ferramenta SODAS que analisa objetos simbólicos classificando-os de acordo com classes identificadas. Como os dados a serem analisados e processados são formados somente por valores numéricos, foram criados aleatoriamente grupos classificatórios e esses valores foram inseridos dentro da tabela virtual como novos atributos.

Para melhor entendimento do resultado da etapa de exploração dos dados tem-se:

- cledpublica: classe criada para identificar as parcelas com relação aos dados da educação pública. Essa classe é composta por valores resultantes da comparação entre a quantidade de pessoas que estudavam em escolas públicas em 2004 e 2010;
- cledpart: classe criada para identificar as parcelas com relação aos dados da educação particular. Essa classe é composta por valores resultantes da comparação entre a quantidade de pessoas que estudavam em escolas particulares em 2004 e 2010;
- clrenda: classe criada para identificar as parcelas com relação aos dados da renda familiar. Essa classe é composta por valores resultantes da comparação entre a renda familiar das parcelas em 2004 e 2010;
- clpadrão: classe criada para identificar as parcelas com relação aos dados do padrão construtivo. Essa classe é composta por valores resultantes da comparação entre os valores do padrão construtivo em 2004 e 2010;
- clarea: classe criada para identificar as parcelas com relação aos dados da área construída. Essa classe é composta por valores resultantes da comparação entre as metragens da área construída em 2004 e 2010;

Para a classificação do padrão construtivo, três classes foram criadas comparando a evolução de 2010 em relação a 2004:

1. Parcelas que mantiveram o padrão construtivo, foi inserido o valor “IGUAL” ao seu registro;
2. Parcelas que melhoraram o padrão construtivo, foi inserido o valor “MELHOR” ao seu registro;

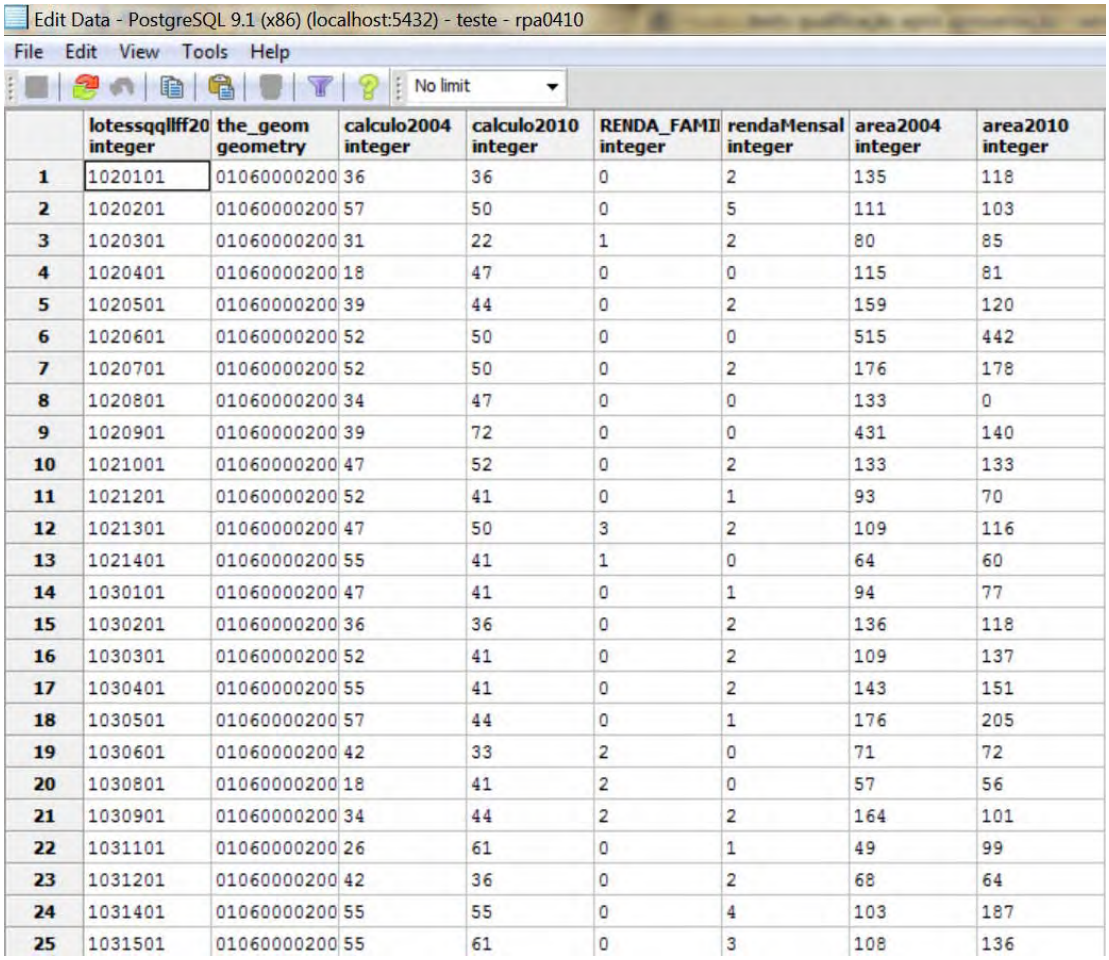
3. Parcelas que pioraram o padrão construtivo, foi inserido o valor “MENOR” ao seu registro;

O mesmo foi feito para os registros relativos à área construída.

Para a renda familiar, porém, mais classes foram criadas, pois a oscilação em salários mínimos dos registros, comparando 2004 a 2010, varia de valores menores a um salário mínimo (denominado pelo valor “0”) até 5 salários mínimos (denominado pelo valor “5”). Para esses registros foram criadas as seguintes classes:

1. Parcelas que mantiveram a renda familiar, foi inserido o valor “IGUAL” ao seu registro;
2. Parcelas que diminuíram a renda familiar, foi inserido o valor “MENOR” ao seu registro;
3. Parcelas que aumentaram a renda familiar entre 1 e 2 salários mínimos, foi inserido o valor “1--2” ao seu registro;
4. Parcelas que aumentaram a renda familiar entre 3 e 4 salários mínimos, foi inserido o valor “3--4” ao seu registro;
5. Parcelas que aumentaram a renda familiar acima de 4 salários mínimos, foi inserido o valor “ACIMA DE 4” ao seu registro;

A oscilação de renda familiar menor que 1 salário mínimo não foi levado em consideração.



Edit Data - PostgreSQL 9.1 (x86) (localhost:5432) - teste - rpa0410

File Edit View Tools Help

No limit

	lotessqqlfff20 integer	the_geom geometry	calculo2004 integer	calculo2010 integer	RENDA_FAMI integer	rendaMensal integer	area2004 integer	area2010 integer
1	1020101	01060000200	36	36	0	2	135	118
2	1020201	01060000200	57	50	0	5	111	103
3	1020301	01060000200	31	22	1	2	80	85
4	1020401	01060000200	18	47	0	0	115	81
5	1020501	01060000200	39	44	0	2	159	120
6	1020601	01060000200	52	50	0	0	515	442
7	1020701	01060000200	52	50	0	2	176	178
8	1020801	01060000200	34	47	0	0	133	0
9	1020901	01060000200	39	72	0	0	431	140
10	1021001	01060000200	47	52	0	2	133	133
11	1021201	01060000200	52	41	0	1	93	70
12	1021301	01060000200	47	50	3	2	109	116
13	1021401	01060000200	55	41	1	0	64	60
14	1030101	01060000200	47	41	0	1	94	77
15	1030201	01060000200	36	36	0	2	136	118
16	1030301	01060000200	52	41	0	2	109	137
17	1030401	01060000200	55	41	0	2	143	151
18	1030501	01060000200	57	44	0	1	176	205
19	1030601	01060000200	42	33	2	0	71	72
20	1030801	01060000200	18	41	2	0	57	56
21	1030901	01060000200	34	44	2	2	164	101
22	1031101	01060000200	26	61	0	1	49	99
23	1031201	01060000200	42	36	0	2	68	64
24	1031401	01060000200	55	55	0	4	103	187
25	1031501	01060000200	55	61	0	3	108	136

Figura 22: Tabela virtual criada contendo os dados relativos ao domínio do questionamento levantado

	lotessqqllff20 integer	the_geom geometry	padrao2004 integer	padrao2010 integer	renda2004 integer	renda2010 integer	area2004 integer	area2010 integer	edpublica200 integer	edpart2004 integer	edpublica201 integer	edpart2010 integer	clrenda character var	clpadrao character var
1	1020501	01060000200	39	44	0	2	159	120	1	0	0	0	1--2	MELHOR
2	1021301	01060000200	47	50	3	2	109	116	2	0	1	0	MENOR	MELHOR
3	1021401	01060000200	55	41	1	0	64	60	2	0	0	0	MENOR	PIOR
4	1020401	01060000200	18	47	0	0	115	81	0	0	0	0	IGUAL	MELHOR
5	1020601	01060000200	52	50	0	0	515	442	0	0	0	0	IGUAL	PIOR
6	1020801	01060000200	34	47	0	0	133	1	0	0	0	0	IGUAL	MELHOR
7	1020901	01060000200	39	72	0	0	431	140	0	0	0	0	IGUAL	MELHOR
8	1021201	01060000200	52	41	0	1	93	70	2	0	2	0	1--2	PIOR
9	1050701	01060000200	52	58	7	0	160	83	0	0	0	0	MENOR	MELHOR
10	1051401	01060000200	36	36	3	1	166	123	0	0	0	0	MENOR	IGUAL
11	1060501	01060000200	57	52	2	1	141	107	0	0	0	0	MENOR	PIOR
12	4062501	01060000200	21	30	2	1	26	15	0	0	0	0	MENOR	MELHOR
13	5010301	01060000200	39	41	1	1	69	91	0	0	0	0	IGUAL	MELHOR
14	5010601	01060000200	55	50	3	0	94	150	0	0	0	0	MENOR	PIOR
15	5011301	01060000200	50	52	2	2	87	87	0	0	0	0	IGUAL	MELHOR
16	5012001	01060000200	42	52	1	1	56	56	0	0	0	0	IGUAL	MELHOR
17	1021001	01060000200	47	52	0	2	133	133	1	0	2	0	1--2	MELHOR
18	1040701	01060000200	13	30	0	2	55	49	0	0	1	0	1--2	MELHOR
19	1042101	01060000200	52	38	2	2	70	71	0	0	1	0	IGUAL	PIOR
20	1050901	01060000200	52	47	3	3	122	119	0	0	2	0	IGUAL	PIOR
21	1020201	01060000200	57	50	0	5	111	103	2	0	0	2	ACIMA DE 4	PIOR

Figura 23: Tabela virtual com a inserção de campos classificatórios para execução da ferramenta SODAS

3.3.2.2 Segundo Caso

Para o segundo caso, tomou-se como referência a patologia hipertensão arterial e o levantamento de 2012, portanto criou-se uma tabela virtual com o identificador das parcelas do levantamento de 2012 e a união dos dados dessa patologia nos levantamentos de 2004 e 2010 que tinham sido coletados também em 2012. Parte da tabela gerada é mostrada na Figura 24.

	ssqqllffdiagm character var	hart2004 double precis	hart2010 integer	hart2012 integer
1	03071401	0	1	1
2	02071901	1	0	0
3	01040301	1	0	1
4	01031001	3	1	2
5	03030201	2	1	1
6	03010401	1	1	0
7	02080701	0	2	0
8	01070301	1	0	1
9	01021201	0	1	0
10	01080601	0	1	0
11	02061201	3	0	0
12	02041001	0	0	1
13	03020201	1	1	0
14	04050801	0	0	2
15	04071401	0	1	1
16	03060301	0	3	2
17	02070401	1	1	0
18	05020401	1	0	1
19	01051001	1	1	1
20	01090601	0	1	0
21	04071502	0	0	1
22	02071601	2	0	1

Figura 24: Tabela virtual gerada da evolução da patologia hipertensão arterial

3.3.2.3 Terceiro Caso

Nesse último caso estudado, três patologias foram analisadas em relação à faixa etária dos moradores: hipertensão arterial, cardiopatia e depressão. A faixa etária analisada foi a que contempla moradores acima de 45 anos. Essa escolha se deu em virtude do BIC de 2004, que contemplava como última faixa etária, moradores acima de 45 anos. Com isso uma somatória teve que ser feita nos dados do levantamento de 2012 que contem em seu BIC moradores de 46 a 60 anos e acima de 60 anos. Para este caso, a tabela foi gerada para ser processada pelo SODAS (*software* francês de mineração de dados). Como no primeiro caso, alguns atributos classificatórios precisaram ser inseridos na tabela gerada para ser processada corretamente.

Nesse caso foram trabalhados somente com três classes: “MENOR”, “IGUAL” e “MAIOR”. Parte dessa tabela é mostrada na Figura 25.

Para melhor entendimento do resultado da etapa de exploração dos dados tem-se a descrição das classes criadas:

- clfaixa: classe criada para identificar as parcelas com relação aos dados de faixa etária. Essa classe é composta por valores resultantes da comparação entre a quantidade de pessoas acima de 45 anos em 2004 e 2012;

- cldepressao: classe criada para identificar as parcelas com relação aos dados de depressão. Essa classe é composta por valores resultantes da comparação entre a quantidade de pessoas com a patologia de depressão em 2004 e 2012;

- clcardiopatia: classe criada para identificar as parcelas com relação aos dados de cardiopatia. Essa classe é composta por valores resultantes da comparação entre a quantidade de pessoas com a patologia de cardiopatia em 2004 e 2012;

- clhiper: classe criada para identificar as parcelas com relação aos dados de hipertensão arterial. Essa classe é composta por valores resultantes da comparação entre a quantidade de pessoas com a patologia de hipertensão arterial em 2004 e 2012;

	ssqqllffdiagm character var	acima45anos integer	depressao04 integer	cardiopata04 integer	hiper04 integer	acima45anos double precis	depressao12 double precis	cardiopata12 double precis	hiper12 double precis	clfaixa character var	cldepressao character var	clcardiopatia character var	clhiper character var
1	03051601	0	0	0	1	1	0	0	0	MAIOR	IGUAL	IGUAL	MENOR
2	05021001	0	0	0	1	1	0	0	0	MAIOR	IGUAL	IGUAL	MENOR
3	04051001	1	1	0	1	2	0	0	0	MAIOR	MENOR	IGUAL	MENOR
4	03080501	1	0	0	0	3	0	1	1	MAIOR	IGUAL	MAIOR	MAIOR
5	03020701	1	0	1	0	2	0	0	0	MAIOR	IGUAL	MENOR	IGUAL
6	03072001	1	0	0	0	2	0	1	2	MAIOR	IGUAL	MAIOR	MAIOR
7	01081201	1	0	0	0	2	0	0	0	MAIOR	IGUAL	IGUAL	IGUAL
8	05020601	0	0	0	0	2	0	0	0	MAIOR	IGUAL	IGUAL	IGUAL
9	01060401	0	0	0	0	2	0	0	0	MAIOR	IGUAL	IGUAL	IGUAL
10	01051601	0	0	0	0	2	0	0	0	MAIOR	IGUAL	IGUAL	IGUAL
11	01091701	0	0	0	0	2	0	0	0	MAIOR	IGUAL	IGUAL	IGUAL
12	02051701	0	0	0	0	1	1	0	0	MAIOR	MAIOR	IGUAL	IGUAL
13	01041801	1	0	0	1	2	0	0	1	MAIOR	IGUAL	IGUAL	IGUAL
14	01080501	1	0	0	0	2	1	0	0	MAIOR	MAIOR	IGUAL	IGUAL
15	04050701	0	0	0	0	1	0	0	0	MAIOR	IGUAL	IGUAL	IGUAL
16	04062101	0	0	0	0	1	0	0	0	MAIOR	IGUAL	IGUAL	IGUAL
17	03051901	1	0	0	0	0	0	0	0	MENOR	IGUAL	IGUAL	IGUAL
18	04072501	1	0	0	0	0	0	0	0	MENOR	IGUAL	IGUAL	IGUAL
19	02061201	0	0	0	0	1	0	1	3	MAIOR	IGUAL	MAIOR	MAIOR
20	01071601	0	0	0	0	2	0	0	0	MAIOR	IGUAL	IGUAL	IGUAL
21	01031401	0	0	0	0	1	0	0	1	MAIOR	IGUAL	IGUAL	MAIOR
22	04090701	1	0	0	1	2	0	0	1	MAIOR	IGUAL	IGUAL	IGUAL

Figura 25: Tabela virtual com a inserção de campos classificatórios para execução da ferramenta SODAS

3.3.3 Extração de padrões e pós-processamento

Duas ferramentas foram utilizadas na fase de extração de padrões. A primeira delas é a Matlab[12] considerada uma linguagem de alto nível e um ambiente interativo para computação numérica, visualização e programação. Com ela pode-se analisar dados, desenvolver algoritmo e criar modelos e aplicações. A outra ferramenta foi a SODAS[13], *software* francês de mineração de dados desenvolvido pelo Departamento CEREMADE (*Centre De Recherche en Mathématiques de lá Décision)*, tendo como principais características a análise de dados simbólicas e *clustering.*

No Matlab, o estudo foi baseado em técnicas de Redes Neuraus Artificiais (RNA) cuja tarefa escolhida foi a de Classificação. O algoritmo de teste escolhido foi o de redes de *Kohonen*, mais especificamente Self *Organizing Maps* (*SOM*), também chamado de Mapas Auto-Organizáveis de *Kohonen*.

As *SOM* são redes neuronais com especial interesse nas tarefas de "*clustering*" e visualização. Podem ser usados para projetar uma grande quantidade de informação de elevada dimensão numa dimensão menor, como define Kohonen (2001). São uma classe de Redes Neurais Artificiais onde a aprendizagem ocorre de modo não-supervisionado, podendo descobrir padrões e características relevantes em um conjunto de dados. Possui como

[12] Disponível em: http://www.mathworks.com/products/matlab/. Acessado em 17/12/2012.
[13] Disponível em: http://www.ceremade.dauphine.fr/SODAS/sodas-presentation.htm. Acessado em 17/12/2012.

principal característica a auto-organização, ou seja, realizam classificações por meio dos conteúdos dos dados (sua natureza) e não por classificações previamente informadas.

Sua arquitetura, mostrada na Figura 26, possui duas camadas conectadas: a primeira composta pelos dados de entrada e a segunda camada pelos elementos processados e classificados podendo ser constituída de uma ou mais dimensões.

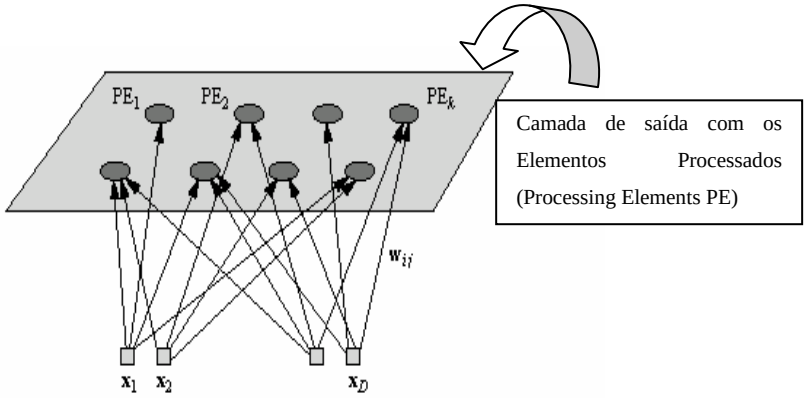


Figura 26: Arquitetura de uma rede *SOM* com saída 2D
Fonte: (KOHONEN, 2001)

As características essenciais do algoritmo são:

- Aproximação do Espaço de Entrada: O *SOM* é capaz de preservar a estrutura do espaço de entrada relativamente bem
- Ordenamento Topológico: Os PEs na saída do *SOM* estão topologicamente ordenados no sentido de que PEs vizinhos correspondem a regiões similares no espaço de entrada
- Manutenção da densidade: Regiões no espaço de entrada com maior densidade de pontos são mapeadas para regiões maiores no espaço de saída

Existem, no Matlab, diversas funções predefinidas de visualização dos mapas de Kohonen. A função utilizada nesse Estudo de Caso foi a *som_plotplane* que mostra um gráfico de linha para cada registro mapeado. Como o objetivo é obter o conhecimento da evolução histórica dos dados, buscando grupos homogêneos, optou-se pela escolha dessa função, pois um gráfico linear demonstra tal evolução.

A visualização será mostrada em forma de um conjunto de “casulos” formando uma “colméia”. Dentro de cada “casulo”, que representa um registro analisado, um gráfico linear processado por meio dos dados de entrada será construído. Os grupos homogêneos serão separados por cores definidas pelo próprio programa.

Já no SODAS, o estudo foi baseado na Análise de Dados Simbólicos(*Symbolic Data Analysis – SDA)*, uma de suas principais características.

Análise de dados simbólicos é um campo relativamente novo que fornece uma série de métodos para análise de conjunto de dados complexos a fim de “extrair conhecimento” de tais dados. “Extrair Conhecimento” significa obter resultados explicativos, por isso “objetos simbólicos” são introduzidos e estudados nesta técnica (DIDAY; MONIQUE, 2008).

O método utilizado dentro da ferramenta SODAS foi o *VIEW*, que é o Visualizador de Dados Simbólicos.

3.3.3.1 Primeiro Caso

Primeiramente separou-se a tabela virtual em 2 partes para analisar as seguintes comparações: influência do histórico de renda familiar considerando o histórico do padrão construtivo e a influência do histórico de renda familiar considerando o histórico da área construída. Essa separação foi feita no *software* Microsoft Excel resultando em duas planilhas distintas como mostram as Figuras 27 e 28. Em seguida os dados foram normalizados linearmente para uma melhor resposta do *software.*

Cada linha da planilha corresponde a um registro inserido dentro da base de dados, ou seja, dados de uma determinada parcela, mas somente os dados a serem minerados. Como são dois contextos a serem analisados dentro de um mesmo gráfico, foi adicionada uma coluna separadora (valores “0”) em cada planilha para conseguir visualizar separadamente a oscilação gráfica em cada parcela.

	A	B	C	D	E
1	renda 2004	renda 2010	coluna separadora	área construída 2004	área construída 2010
2	0	2	0	2	1
3	0	5	0	1	1
4	1	2	0	1	1
5	0	0	0	1	1
6	0	2	0	2	1

Figura 27: Dados normalizados de renda familiar e área construída

	A	B	C	D	E
1	renda 2004	renda 2010	coluna separadora	padrão construtivo 2004	padrão construtivo 2010
2	0	2	0	4	4
3	0	5	0	6	5
4	1	2	0	3	2
5	0	0	0	1	5
6	0	2	0	4	5

Figura 28: normalizados de renda familiar e padrão construtivo

Como resultados da análise, foram gerados duas “colméias” contendo agrupamentos de comportamento dos dados referentes às parcelas. A Figura 29 mostra o resultado da análise entre a renda familiar e a área construída, e a Figura 30 mostra o resultado da análise entre a renda familiar e o padrão construtivo.

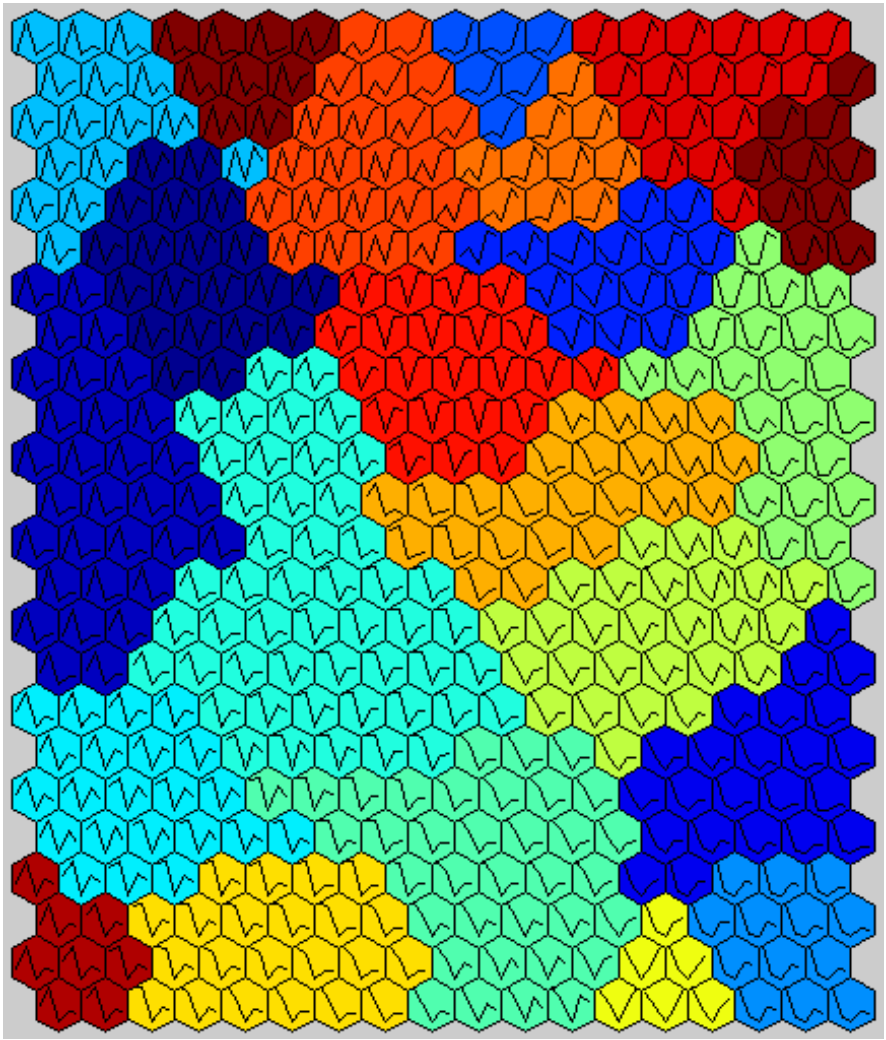


Figura 29: Colmeia de gráficos referente à análise entre renda familiar e área construída

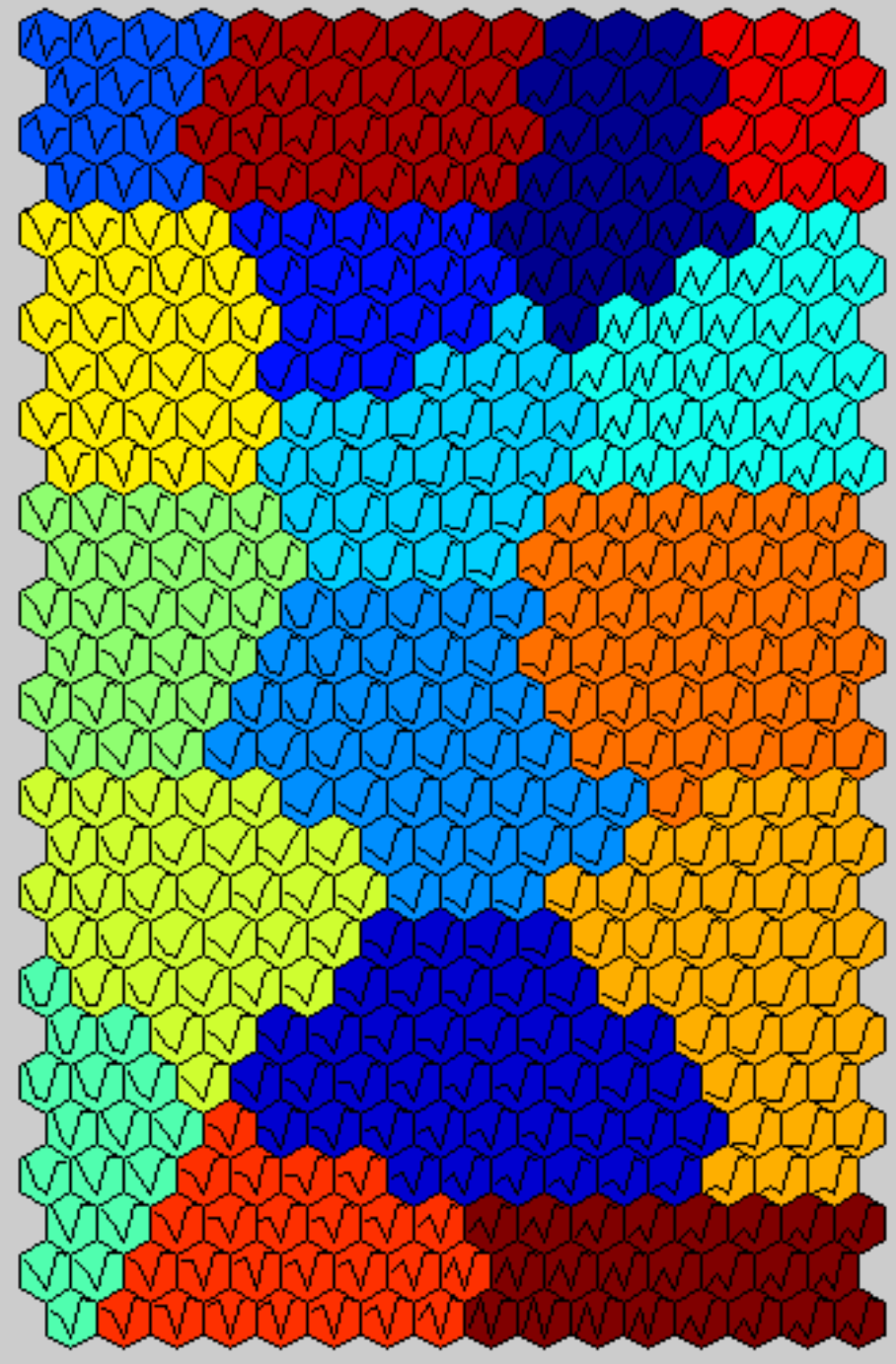


Figura 30: Colmeia de gráficos referente à análise entre renda familiar e padrão construtivo.

Iniciando a fase de pós-processamento, analisou-se o primeiro resultado mostrado na Figura 29, e percebeu-se por meio dos gráficos uma situação atípica, agrupamentos de parcelas que tiveram aumento da renda familiar, porém diminuíram a área construída comparando os levantamentos de 2004 e 2010. Entretanto seria interessante para o gestor que essa informação fosse especializada para uma melhor interpretação. Utilizando mais duas ferramentas computacionais (gvSIG e PostGIS) gerou-se o mapa dessa constatação, que é mostrado na Figura 31.

Por ser uma situação inusitada, uma pesquisa foi feita dentro da base de dados para verificar qual o percentual de parcelas que diminuíram sua área e verificou-se que 55% dessas parcelas diminuíram 10% ou menos em relação ao levantamento cadastral de 2004 significando um possível erro de coleta dos dados. Com isso pode-se verificar que as técnicas de mineração de dados também podem auxiliar na descoberta de erros de coleta de dados. Com a espacialização desses dados fica fácil definir onde os coletores deverão voltar para buscar os dados corretos e atualizar a base cadastral.

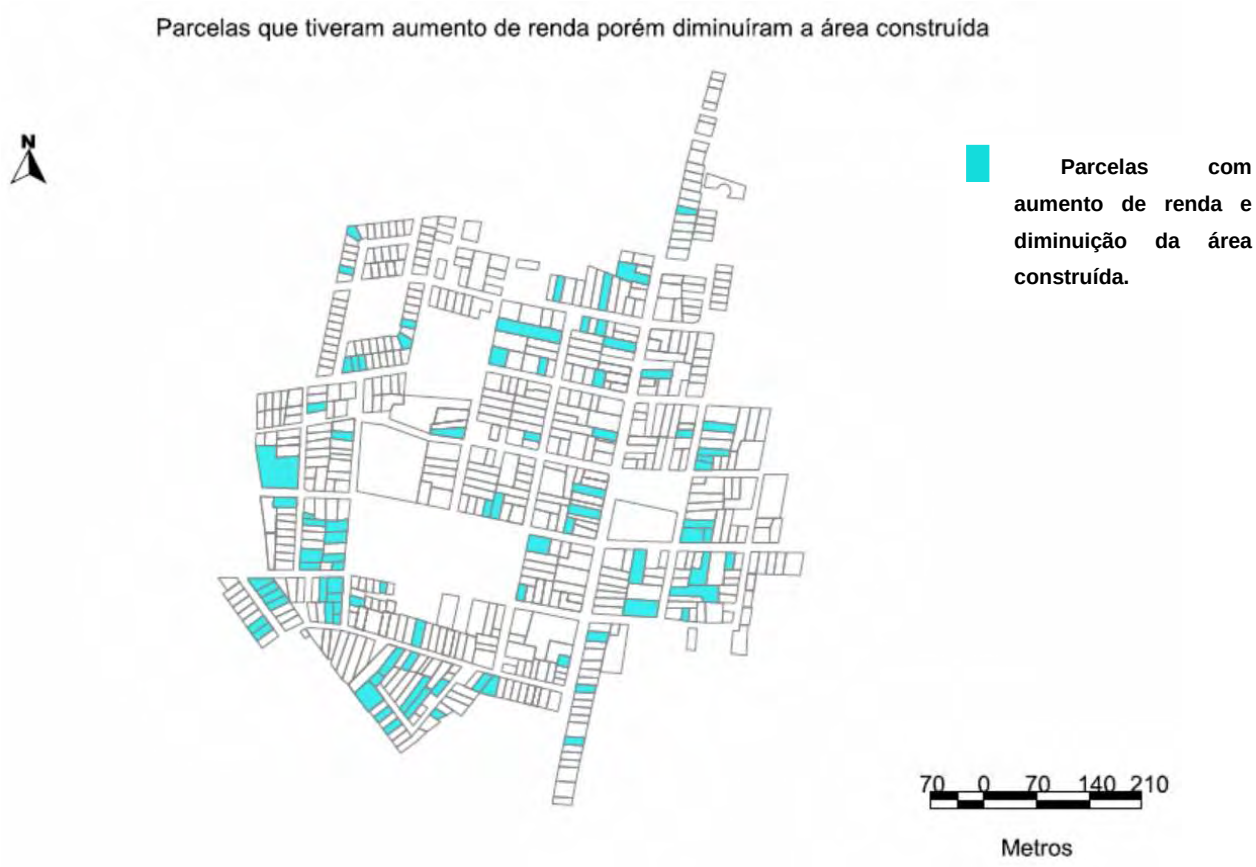


Figura 31: Espacialização que mostra as parcelas que tiveram aumento de renda familiar e uma diminuição da área construída

Na segunda ferramenta, todos os dados processados foram utilizados para a análise. Nesses dados foram inseridos alguns atributos de classificação de valores para uso da ferramenta. O resultado gerado é mostrado na Figura 32. A legenda dessa Figura 32 representa as classes de aumentos salariais definidas como o objeto simbólico e os eixos do gráfico representam a oscilação das outras classes em relação a esse objeto.

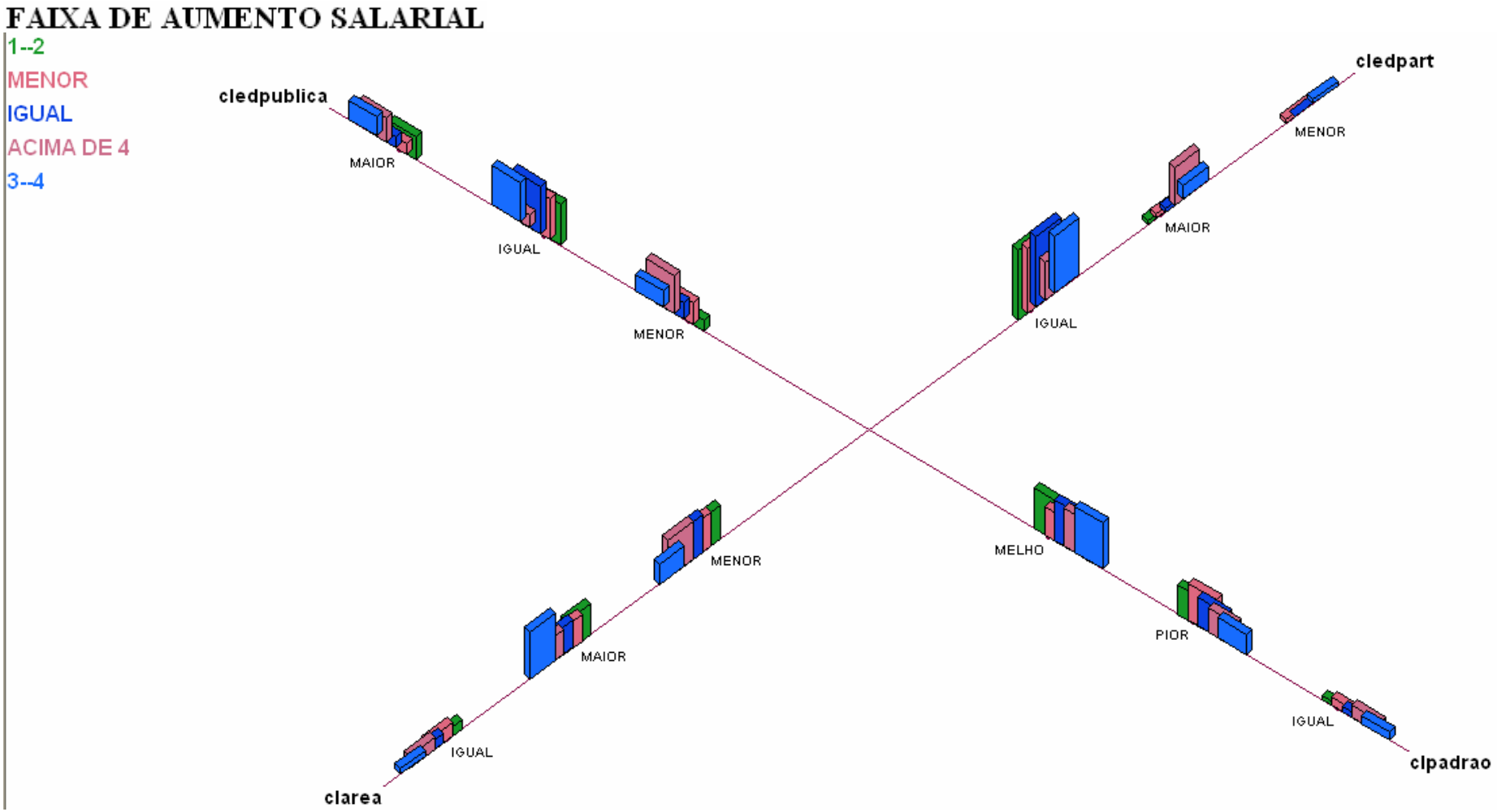


Figura 32: Resultado da análise de dados feita pelo SODAS

A análise de pós-processamento feita neste caso foi em comparação entre a renda familiar e o padrão construtivo. Nota-se neste caso que as parcelas que tiveram um aumento de renda entre 3 e 4 salários mínimos de 2004 para 2010 foram as que mais investiram na melhoria do padrão construtivo. Novamente, entende-se que para o gestor, essa informação precisa ser especializada. O resultado é mostrado na Figura 33.

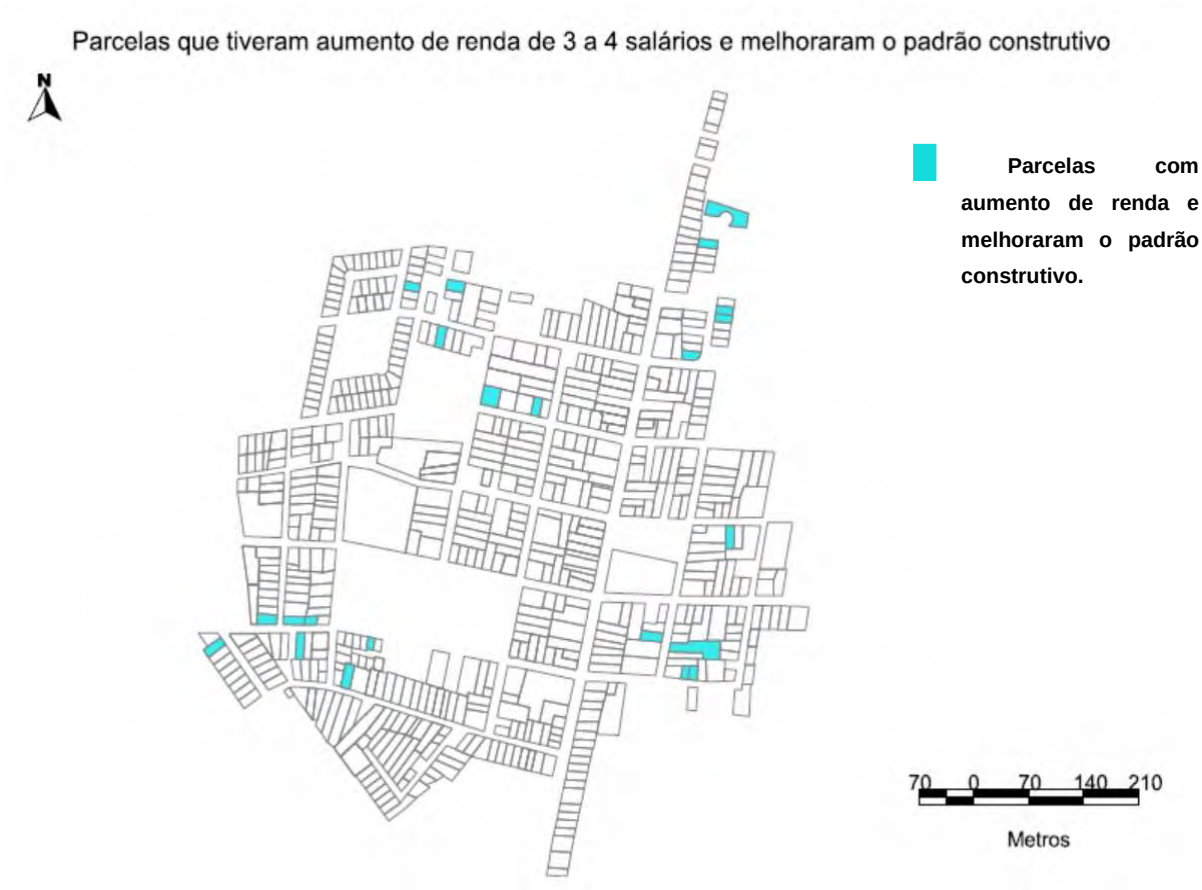


Figura 33: Espacialização que mostra as parcelas que tiveram aumento de renda familiar entre 3 e 4 salários mínimos e investiram em padrão construtivo

3.3.3.2 Segundo Caso

Nesse caso, somente o Matlab foi usado na Mineração. Os dados da tabela virtual foram exportados para o Microsoft Excel e carregados para processamento do algoritmo resultando na colmeia de gráficos mostrado na Figura 34.

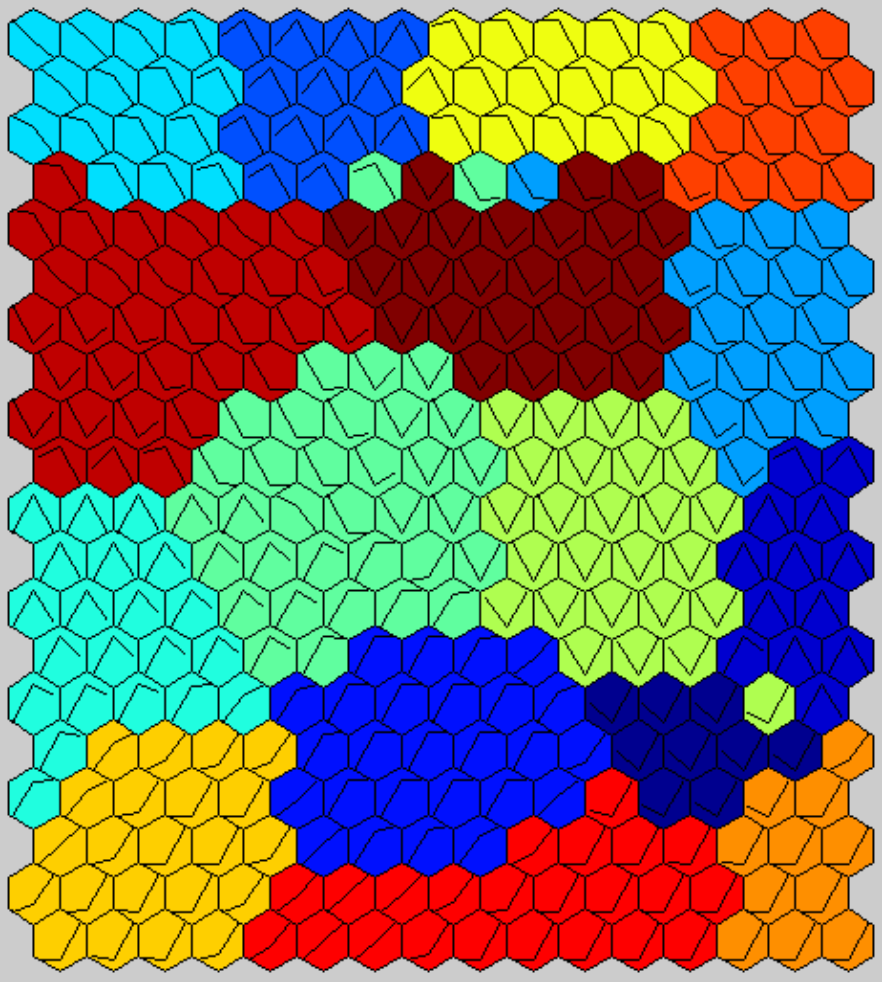


Figura 34: Colmeia de gráficos referente à análise da evolução da patologia hipertensão arterial.

Apesar do conjunto de dados ter sido pequeno para esse tipo de Mineração, algumas classes foram identificadas e analisadas. Analisando a Figura 34, foi escolhido para espacialização a classe de parcelas que em 2004 e 2010 não existiam indícios dessa patologia porém em 2012 ela apareceu com mais de um caso. Essa espacialização é mostrada na Figura 35.

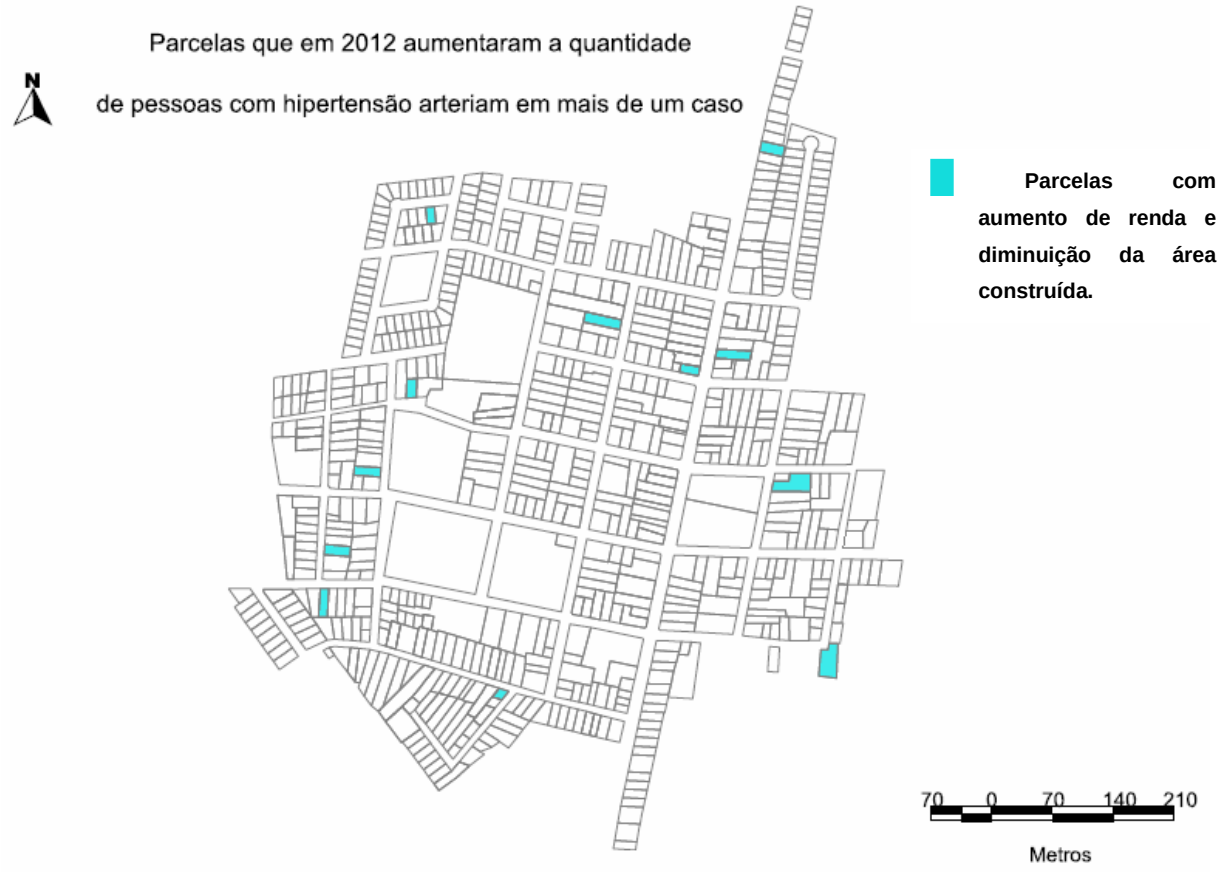


Figura 35: Espacialização que mostra as parcelas que tiveram aumento acima de um caso na patologia hipertensão arterial em 2012

3.3.3.3 Terceiro Caso

No último caso, ao contrário do anterior, somente a ferramenta SODAS foi utilizada para verificar a influência da faixa etária na ocorrência de 3 patologias: Cardiopatia, Depressão e Hipertensão arterial. O resultado gerado é mostrado na Figura 36.

Por meio desse resultado, analisa-se que ocorreu certa estabilidade na ocorrência destas três patologias para a faixa etária acima de 45 anos. Percebe-se que provavelmente alguns falecimentos ou mudanças de localização ocorreram, porém novos integrantes à essa faixa etária surgiram mas nem por esse motivo grandes oscilações ocorreram.

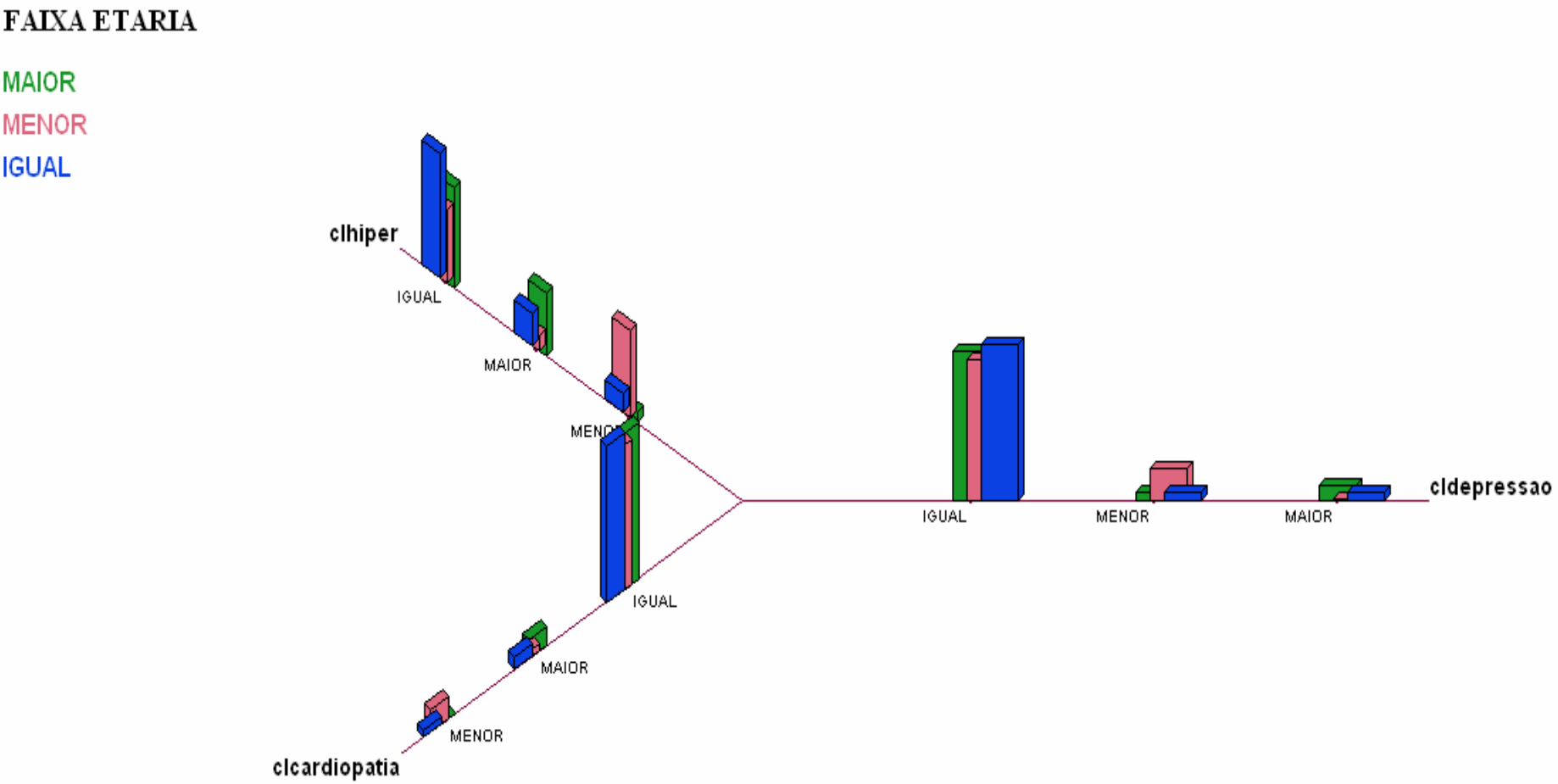


Figura 36: Resultado das análises feita pelo SODAS

4 CONCLUSÕES

A importância da criação e disseminação do conhecimento para qualquer ambiente organizacional é parte importante no processo de tomada de decisão. A Mineração de Dados propõe transformar dados em informação e conhecimento propriamente ditos.

O que se encontra hoje na maioria das prefeituras brasileiras são enormes repositórios de dados (matéria bruta) que guardam características e aspectos do ambiente trabalhado.

A extração de conhecimento em grandes bases de dados utilizando *DM* objetiva buscar informações que é o resultado do processamento executado nesses dados, e gerar conhecimento, que é um conjunto de argumentos e explicações interpretando as informações processadas.

A modelagem de dados também se apresenta com um papel fundamental nesse processo. Não se pode projetar um Banco de Dados observando somente a realidade momentânea, é necessário ter uma visão futura do que os dados armazenados poderão trazer de benefícios aos gestores.

Não se pode ignorar as inovações tecnológicas. Novas ferramentas surgem para manipulação dos dados e precisam ser avaliadas e utilizadas. A busca por melhores resultados nos processamentos dos dados mostra que, somente consultas SQL não alcançam os objetivos buscados por gestores. *KDD* e *DM* se apresentam nesse contexto para trazer resultados de diferentes formatos, facilitando a interpretação e avaliação do tomador de decisão.

Entretanto, vale salientar que para cada objetivo desejado devem-se aplicar tarefas e técnicas específicas para se conseguir qualidade nos resultados esperados.

A presença do profissional também não pode ser descartada, ele participa desde o início como conhecedor do domínio do problema até o final na análise de viabilidade dos resultados.

Outro ponto essencial é a qualidade de dados utilizados na mineração. Para que o resultado possa ser utilizado por profissionais no processo de tomada de decisão, é necessário que os dados sejam coletados e armazenados de maneira precisa. Dados imprecisos podem não descrever corretamente um imóvel.

4.1 *Recomendações para trabalhos futuros*

A partir dos estudos realizados, algumas sugestões de atividades futuras são apresentadas:

- Aplicação de novas tarefas e técnicas de mineração de dados para os dados do Cadastro Territorial Urbano;
- Construção de um modelo de *DW* que consiga integrar os departamentos das prefeituras a fim de gerar um repositório único que possa fornecer por meio de técnicas de *KDD* novos conhecimentos a seus gestores;
- Desenvolver uma pesquisa com especialistas sobre os dados que possivelmente posam ser inseridos no Cadastro Territorial Urbano possibilitando novos resultados para decisões mais elaboradas;

5 REFERÊNCIAS

AMORIM, A.; et al. **A Modernização do cadastro técnico multifinalitário urbano e a influência da evolução tecnológica: uma reflexão sobre o futuro e a multidisciplinaridade do cadastro.** Info GPS/GNSS, Curitiba, p. 46 - 47, 2007.

AMORIM, A. SOUZA, G. H. B; DALAQUA, R. R. **Uma metodologia alternativa para otimização da entrada de dados em sistemas cadastrais**. Revista Brasileira de Cartografia. Rio de Janeiro, V.56, n. 1, p. 47-54. 2004.

AMORIM, A.; SOUZA, G. H. B.; YAMASHITA, M. C. **Cadastro técnico multifinalitário via internet: um importante instrumento de apoio ao planejamento municipal.** Revista Brasileira de Cartografia (Online), v 60/2, p. 119-125, 2008. Disponível em: http://www.rbc.ufrj.br/_2008/60_2_02.htm. Acesso em: 27 set. 2011.

ANTUNES, A. F. B. **Cadastro técnico urbano e rural**. Universidade Federal do Paraná, 2007. Disponível em: http://www.scribd.com/doc/2436511/UFPR-Eng-Cart-ApostilaCadastro2007. Acesso em: 14 de mai. 2010.

BERRY , Michael J. e LINOFF, Gordon. **Data Mining Techniques For Marketing, Sales and Customer Support.** John Wiley & Sons, 1997.

BRASIL. Ministério do Planejamento, Orçamento e Gestão. Secretaria de Gestão. **Programa Nacional de Gestão Pública e Desburocratização – GESPÚBLICA**. Brasília, 2008.

CHEN, P. **Modelagem de Dados.** São Paulo: Makron Books, 1990.

CÔRTES, P. L. **Administração de Sistemas de Informação.** Editora Saraiva, São Paulo, 2007.

DALE, P. F.; MCLAUGHLIN, J. D. **Land information management:** an introduction with special reference to cadastral problems in third world countries. Reprinted (with correction). Oxford. Oxford University Press, 1990.

DATE, C. J. **Introdução a Sistemas de Banco de Dados.** Tradução: Vandenberg Dantas de Souza. 7ª ed. americana. Rio de Janeiro: Campus, 2000.

DIAS, M. M. **Parâmetros na escolha de técnicas e ferramentas de Mineração de Dados.** Acta Scientiarum, v. 24, n. 6, p. 1715, Maringá, 2002.

DIDAY, E. ; MONIQUE N.F. **Symbolic Data Analysis and the SODAS Software.** John Wiley, 2008.

DINIZ, E. A. et AL. **Atualização do Sistema Cadastral da Cidade de Ribeirão dos Índios – SP.** 2004, 124f. Trabalho de Graduação (Graduação em Engenharia Cartográfica). Faculdade de Ciência e Tecnologia, Universidade Estadual Paulista “Júlio de Mesquita Filho”, Presidente Prudente, 2004.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de Banco de Dados.** Revisor: Luis Ricardo de Figueiredo. 4° ed. São Paulo: Pearson Addison Wesley, 2005.

ERBA, D. A. O Cadastro Territorial: presente, passado e futuro. In: ERBA, D. A.; OLIVEIRA, F. L.; LIMA JÚNIOR, P. N. (Org.) **Cadastro multifinalitário como instrumento da política fiscal e urbana.** Rio de Janeiro, 2005. Disponível em: http://www.cidades.gov.br/index.php?option=com_content&view=article&id=547:cadastro-multifinalitario-como-instrumento-de-politica-fiscal-e-urbana&catid=48&Itemid=83. Acesso em: 01 out. 2011.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From *DM* to Knowledge Discovery in Databases.** AI Magazine, Volume 17, Number 3. 1996a.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **The KDD Process for Extracting Useful Knowledge from Volumes of Data.** Comunications of the ACM, Volume 39, Number 11. 1996b.

FERREIRA, K. R. et al. Arquiteturas e linguagens. In: LAENDER A. H. F. et al. **Banco de dados geográfico.** São José dos Campos, Rio de Janeiro, Belo Horizonte: Livro *on-line*, 2005. Disponível em: http://www.dpi.inpe.br/gilberto/livro/bdados/. Acesso em: 22 de set. 2012

FIG – Fedération Internationale de Gèométres. **Statement on the cadastre**. International Federation of Surveyors. Disponível em: http://www.fig.net/commission7/reports/cadastre/statement_on_cadastre.html. Acesso em: 03 mai. 2010.

FREITAS, C. M. D. S. et al. **Introdução a visualização de informações.** Revista de informática teórica e prática, Volume 8, Número 2, 2001, p. 143-158.

GARCIA, R. C.; **O que é preciso saber sobre Cadastro Técnico Multifinalitário.** Biblioteca do Instituto Brasileiro de Administração Municipal (IBAM) 2007.

HAN, J.; KAMBER, M.; PEI, J. ***DM*: Concepts and Techniques.** 2ª ed. Morgan Kaufmann Publisher, 2005.

HARRISON, T. H. **Intranet data warehouse**: ferramentas e técnicas para a utilização do data warehouse na intranet. Berkeley Brasil: São Paulo, 1998.

KEIM, D. A. **Information Visualization and Visual Data Mining.** IEEE Transactions on Visualization and Computer Graphics. Vol. 7, Number 1, 2002.

KOHONEN, T. **Self-Organinzing Maps.** 3rd ed. Springer-Verlag, Berlin, 2001.

KURAHASSI, L. F. **Gestão de energia elétrica – bases para uma política pública municipal.** São Paulo, 2006

LARSSON, G. **Land registration and cadastral systems**. Reprinted. England, UK, Longman Group, 1996

LOCH, C.; ERBA, D. A. **Cadastro técnico multifinalitário:** rural e urbano. Cambridge, MA: Lincoln Institute of Land Policy, 2007.

MALAMAN, C. S.; AMORIM, A. **Utilização do software gvSig no cadastro técnico multifinalitário do município de Ribeirão dos Índios - -SP.** In: Congresso Brasileiro de Cadastro Técnico Multifinalitário – COBRAC, 2010.

MUNIZ, D. P. et al. **Implantação do Cadastro Técnico Multifinalitário em uma área teste.** In:II Congresso Brasileiro de Cadastro Técnico Multifinalitário. Florianópolis – SC, 1996. Anais... Florianópolis – SC, 1996.

O'BRIEN, JAMES A. **Sistemas de Informação e as decisões gerenciais na era da internet.** Tradução: Cid Knipel Moreira. 9ª ed. americana. São Paulo: Saraiva 2003.

OLIVEIRA, F. H. Do cadastro territorial multifinalitário. In: CUNHA, M. P.; ERBA, D. A. (Org). **Manual de apoio – CTM:** Diretrizes para a criação, instituição e atualização do cadastro territorial multifinalitário nos municípios brasileiros. Brasília: Ministério das Cidades, 2010. Disponível em: http://www.cidades.gov.br/images/stories/Arquivos Capacitacao/Capacita%C3%A7%C3%A3o/livro%20diretrizes%20em%20alta.pdf. Acesso em: 27 de out. 2011.

PELEGRINA, et al. **Importância da Análise da Consistência Cadastral Aplicada ao Cadastro Fiscal (Tributário).** II Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação. Recife – PE, 2008. Anais Recife – PE, 2008.

PHILIPS, J. ***Breve histórico do cadastro de imóveis no mundo.*** IRIB em Revista. 317. São Paulo. 2004 p.14-19. ISSN – 1677-437X.

RAMALHO, J. A. A. **SQL: A linguagem dos Bancos de Dados.** São Paulo: Berkley Brasil, 1999.

RAMÃO, Wesley. **Descoberta de Conhecimento Relevante em Banco de Dados sobre Ciência e Tecnologia.** Teses de Doutorado (Pós Graduação em Engenharia de Produção). Universidade Federal de Santa Catarina. Florianópolis, 2002.

REZENDE, S. O. **Sistemas Inteligentes – Fundamentos e Aplicações.** Barueri: Manole, 2005.

SANTOS, M. Y.; RAMOS I. **Business Intelligence: tecnologias da informação na gestão de conhecimento.** FCA Editora de Informática, 2006. ISBN 972-722-405-9. p. 2-10.

SASS, G. G. **Desenvolvimento de um método de Análise Multitemporal para Cadastro Territorial Multifinalitãrio.** Texto de qualificação de Doutorado (Pós Graduação em Ciências Cartográficas) UNESP, 2012.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Sistema de banco de dados.** 5ª ed. São Paulo: Makron Books, 2006.

SOUZA, G. H. B. **Método de modelagem da parcela especial para o cadastro tridimensional.** 2011. 97 f. Tese (Doutorado em Ciências Cartográficas) Faculdade de Ciência e Tecnologia da Universidade Estadual Paulista - UNESP. Presidente Prudente – SP. Disponível em: http://www4.fct.unesp.br/pos/cartografia/docs/teses/t_souza_GHB.pdf. Acesso em: 10 jan. 2012.

VANDE, A. M. **Form Follows Data – The Symbiosis between Design & Information Visualization**. Proceedings of International Conference on Computer-Aided Architectural Design (CAADfutures) pages:31-40, 2005.

WORBOYS, M.F. **GIS: A computing perspective**. London: Taylor and Francis, 1995.