Revista Brasileira de Cartografia (2014) N^0 66/6: 1215-1230
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto
ISSN: 1808-0936

APLICAÇÃO DE TÉCNICAS DE DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS CADASTRAIS PARA AUXILIAR NO PROCESSO DE TOMADA DE DECISÃO

Knowledge Discovery Techniques Applied in Databases Cadastre to Assist in the Decision-Making Process

Sergio Ricardo Ribas Sass[1], Amilton Amorim[2], Milton Hirokazu Shimabukuro[2] & Glaucia Gabriel Sass[3]

[1]Instituto Federal de Mato Grosso do Sul - IFMS
Curso de Tecnologia em Análise e Desenvolvimento de Sistemas
Fazenda Sta. Bárbara s/n – Caixa Postal 144 – 79750-000 – Nova Andradina – MS - Brasil
sergio.sass@ifms.edu.br

[2]Universidade Estadual Paulista - UNESP
Departamento de Cartografia/Departamento de Matemática e Computação
Programa de Pós-Graduação em Ciências Cartográficas
Rua Roberto Simonsen, 305 – 19060-900 – Presidente Prudente – SP - Brasil
amorim@fct.unesp.br
miltonhs@fct.unesp.br

[3]Universidade Estadual de Mato Grosso do Sul - UEMS
Curso de Ciência da Computação
Cidade Universitária – Caixa Postal 351 – 79804-970 – Dourados – MS - Brasil
glaucia@comp.uems.br

Recebido em 15 de Março, 2013/ Aceito em 04 de Setembro, 2013
Received on March 15, 2013/ Accepted on September 04, 2013

RESUMO

Base 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 nessas bases, os gestores passaram a depender não só dos dados, mas também de informações e 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 informações desses dados, tecnologias e técnicas computacionais se tornam grandes aliadas, como Repositório de Dados (Data Warehouse DW) e Mineração de Dados (Data Mining DM). Esse artigo discute brevemente a tecnologia de DM dentro do processo de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in database KDD), e, aliado ao CTM, mostra os resultados de um experimento a partir de dados coletados e armazenados em diferentes anos para a cidade de Ribeirão dos Índios-SP.

Palavras chaves: KDD, Tomada de Decisão, Cadastro Territorial Multifinalitário, Repositório de Dados, Mineração de Dados.

ABSTRACT

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 datas but also information and knowledge extracted from these data to help in the decision-making process. The Multipurpose Cadastre is the tool that manages the public organization's data, and along with it, to extract information of these data, technologies and techniques become great allies, such as the Data Warehouse (DW) and the Data Mining (DM). This article briefly discusses the technology of the DM in the process of Knowledge Discovery in database KDD, and, allied to the Multipurpose Cadastre, shows the results of an experiment based on data collected and stored in different years for the town of Ribeirão dos Índios, SP.

Keywords: KDD, Decision-making, Multipurpose Cadastre, Data Warehouse, Data Mining.

1. INTRODUÇÃO

Com o aumento da demanda pela informação, o processo de informatização da sociedade e o rápido desenvolvimento de ferramentas de coleta e armazenamento de dados, a quantidade de dados coletados e acumulados vem crescendo muito rapidamente nos últimos anos. Associado a isso, a crescente necessidade por informações privilegiadas, voltadas para decisões estratégicas tem despertado o interesse em descobrir novos conhecimentos intrínsecos nas bases de dados.

Essa demanda por conhecimento não se dá somente em organizações privadas. O Gestor Público toma decisões por meio de informações resultantes da tabulação dos dados contidos em suas bases de dados. Uma das ferramentas que dá suporte a essas decisões é o Cadastro Territorial Multifinalitário (CTM).

O CTM é uma ferramenta que vem auxiliando a gestão pública no planejamento territorial acompanhando o processo de desenvolvimento urbano. Porém, a estrutura de um Banco de Dados Cadastral de um CTM necessita de acompanhamento e dedicação para que seus objetivos sejam alcançados. Entretanto, com a evolução do CTM e consequentemente o aumento de seus objetivos, a quantidade e a heterogeneidade de dados armazenados aumentaram em decorrência da interoperabilidade entre os mesmos, causando uma grande complexidade na extração de informações adequadas e úteis para o auxílio do gestor (MORAES; BASTOS, 2012).

Para tratar esses problemas, e disponibilizar informações e novos conhecimentos ao gestor público, existem técnicas computacionais que podem auxiliar os especialistas na análise automatizada de dados e extração de informação útil a partir de grandes bases de dados Dentre elas, uma que se destaca por ter etapas bem definidas na composição de seu processo é a técnica de KDD (*Knowledge Discovery in Database*), sendo que a etapa responsável por buscar novas informações é a Mineração de Dados (*Data Minig - DM*) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Os resultados obtidos por essas técnicas, depois de interpretados e validados por especialistas do domínio da aplicação, podem transformar-se em conhecimento novos.

O objetivo desse trabalho é, por meio de um estudo de caso, demonstrar a execução das etapas do *KDD* tendo como ponto de partida três bases de dados heterogêneas referentes a três levantamentos cadastrais realizados na cidade de Ribeirão dos Índios – SP , buscando gerar novos conhecimentos para o gestor público. Justifica-se a utilização do método por cada levantamento estar armazenado em bases heterogêneas e pela quantidade de dados da pesquisa cadastral aumentar de um levantamento para outro.

2. REVISÃO DA LITERATURA

O CTM é uma ferramenta que 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 uma determinada região. Nessas análises, os dados são obtidos, geralmente, por meio de censos e levantamentos cadastrais específicos (MALAMAN; AMORIM, 2010).

O processo de *KDD*, que, entre outros objetivos, busca extrair padrões em bases de

dados, vem ao encontro da necessidade de se resolver um dos problemas decorrentes da era digital: como obter vantagens estratégicas com o atual montante de dados armazenados nas bases organizacionais (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Sendo assim, uma das soluções seria aliar a tecnologia de Banco de Dados com a tecnologia de Sistemas de Apoio à Decisão. Para isso, faz-se necessário esclarecer que o processo de tomada de decisão não faz parte da tecnologia de Banco de Dados, mas faz uso da mesma. Na verdade são várias as tecnologias que utilizam os Bancos de Dados para dar suporte a esses tipos de Sistemas, são elas: *Data Warehouse DW*, *Data Mart*, Depósitos de Dados Operacionais, *On-Line Analytical Processing (OLAP)*, Bancos de Dados Multidimensionais, *Data Mining DM*, entre outras (DATE, 2000).

As próximas subseções apresentam uma breve discussão sobre CTM, a tecnologia de *KDD*, seus conceitos, tecnologias relacionadas, suas etapas e funcionalidades. Por fim é apresentado o resultado de um experimento utilizando a tecnologia de *DM* aplicada ao CTM.

2.1 Cadastro Territorial Multifinalitário

Os primeiros Sistemas Cadastrais não tinham a visão multidisciplinar, 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).

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 no ano de 1996. A partir desses eventos, fica 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 multidisciplinar 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 e ambiental ao 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 nova 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 1 mostra como foi a evolução das funções do Cadastro.

2.1.1 Estrutura do Cadastro

O CTM, também definido como Sistema de Informação territorial, difere de outros sistemas territoriais por ser baseado em parcelas. Consiste de textos e mapas e são ligados por um identificador único, como mostra a Figura 2. Esses dados são coletados, armazenados e referenciados (DALE; MCLAUGHLIN, 1990).

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

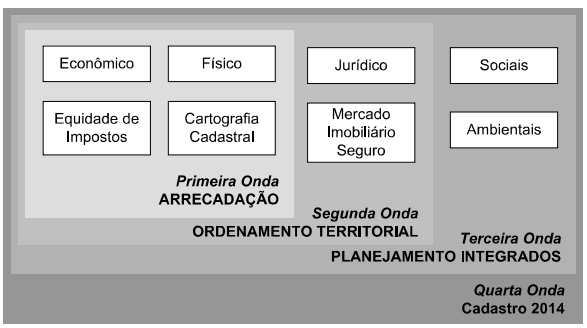

Fig 1 - Evolução das visões do Cadastro. Fonte: Erba (2005).

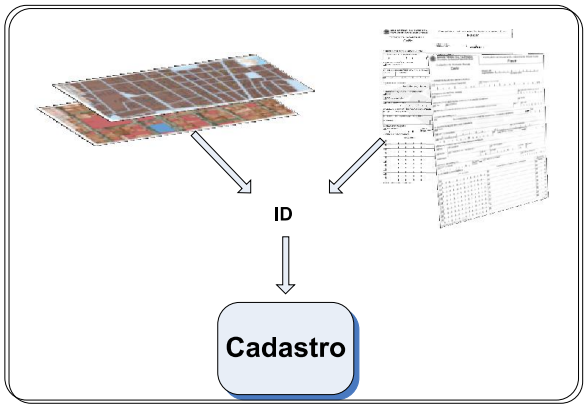

Fig 2 - Estrutura do CTM.

escritório.

Na etapa de planejamento, um diagnóstico da situação atual é realizado bem como um estudo sobre a viabilidade técnica e financeira da implantação do Cadastro. Na etapa de campo é feita a Cartografia e a coleta dos dados sobre as parcelas e seus proprietários. A coleta de dados é feita *in locu* para o preenchimento do Boletim de Informações Cadastrais (BIC). O BIC, armazena os dados socioeconômicos e de caracterização do imóvel. Por meio desses dados coletados, a base cadastral informatizada é carregada. Parte de um BIC é mostrado na Figura 3.

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

De acordo com Pelegrina (2008), a escolha dos campos que serão armazenados e os dados cadastrais são considerados 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.

A integração entre Cadastro Multifinalitário e Sistemas de Informações Geográficas (SIG) se mostra como uma tecnologia de importante auxílio no processo decisório. Porém como na maioria dos projetos de SIG implantados nas prefeituras brasileiras, os bancos de dados não foram preparados para tal integração, ainda

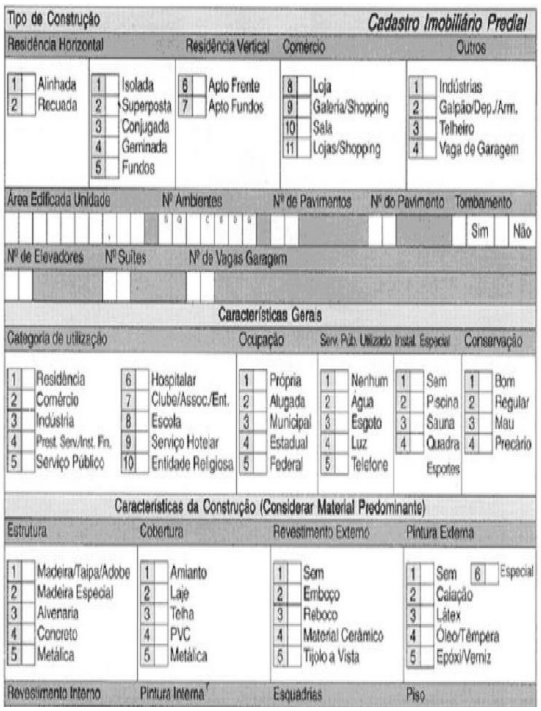

Fig 3 - Exemplo de um BIC convencional.

precisa-se crescer muito nessa área (AMORIM; SOUZA; DALAQUA, 2004).

2.2 Sistemas de Apoio à Decisão

Sistemas de Apoio a Decisão auxiliam na análise de informações do negócio. Tem como objetivo ajudar a administração a definir tendências, apontar problemas e tomar decisões inteligentes. O principal objetivo 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 (OLAP) 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. *OLAP* refere-se à análise multidimensional permitindo examinar as informações armazenadas no banco sob diferentes perspectivas. E *DM*, objetivo desse artigo, busca, por meio da execução das etapas do *KDD*, aplicar algoritmos para 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 e espacializadas.

2.2.1 Dado x Informação x 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 O'Brien (2004), "dados, são fatos ou observações crus normalmente sobre fenômenos físicos ou transações de negócios que ainda não foram convertidos em um contexto significativo".

A informação, componente importante no processo decisório, é formada pelo tratamento do dado. É quando: "sua forma é agregada, manipulada e organizada, seu conteúdo é analisado e avaliado e é colocado em um contexto adequado a um usuário humano(O'Brien, 2004).

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

Atualmente diversas técnicas automatizadas permitem contextualizar a informação de forma a proporcionar ao gestor novas maneiras de interpretá-las e validá-las.

2.2.2 KDD

O processo de extração de conhecimento de bases de dados, responsável por analisar, compreender e extrair padrões de grandes volumes de dados é, por muitos autores, denominado de *KDD (Knowledge Discovery in Database)* (REZENDE, 2005). E *Data Mining* entra como parte particular desse processo com o objetivo de aplicação de algoritmos específicos para a extração de padrões, como mostra a Figura 4 (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

De acordo com (HAN; KAMBER; PEI, 2011) esses passos podem ser detalhados da seguinte maneira:

1. Seleção ➔ quais dados relevantes para a tarefa de pré-processamento são retirados do Banco de Dados;
2. Pré-Processamento ou Limpeza ➔ remover ruídos e inconsistências de dados;
3. Transformação ➔ onde os dados são apropriadamente transformados para mineração através da realização de operações de agregação, por exemplo;
4. Data Mining ➔ processo essencial onde métodos inteligentes 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.

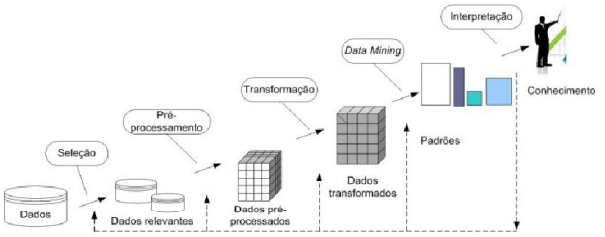

Fig 4 - Uma visão geral das etapas que compõe o processo KDD. Fonte: Adaptado de Fayyad, Piatetsky-Shapiro & Smyth (1996).

2.2.3 Mineração de Dados

Em diversas situações, os conceitos de *KDD* e *DM* se misturam. A maioria das definições de Mineração de Dados utilizada pelos autores foi elaborada por (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996): "Extração de Conhecimento em Bases de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, embutidos nos dados".

Todo processo de *DM* é realizado em função de um domínio específico e dos repositórios de dados inerentes aos mesmos. Para que o *DM* seja executado eficientemente, é necessário que os dados estejam estruturados de forma a serem consultados e analisados adequadamente (REZENDE, 2005). Isso pode ser feito por meio de um *DW,* ou mesmo por meio da unificação temporária e pontual de bases heterogêneas, executando as etapas de Seleção, Pré-processamento e Transformação sem necessariamente criar o *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 grupos:

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 5 mostra as etapas do processo de *DM* adotado por Rezende (2005), e também adotado nos experimentos desse artigo, onde os termos *KDD* e *DM* são tratados com mesmo significado ou seja, referenciando o processo de extrair informações a partir de dados e validá-la como conhecimento. Todo processo anterior ao *DM* (Extração de Padrões), como mostra a Figura 5, é feito dentro da etapa de pré-processamento definido por ela.

- Identificação do problema - detalha-se o domínio da aplicação e define-se os objetivos e metas a serem alcançadas no processo de *DM* .
- Pré-processamento - o processo de *DM* não pode ser aplicado em um banco de dados tradicional, os dados não estão preparados para a aplicação dos algoritmos. 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, *DW*, entre outros;
  - 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;

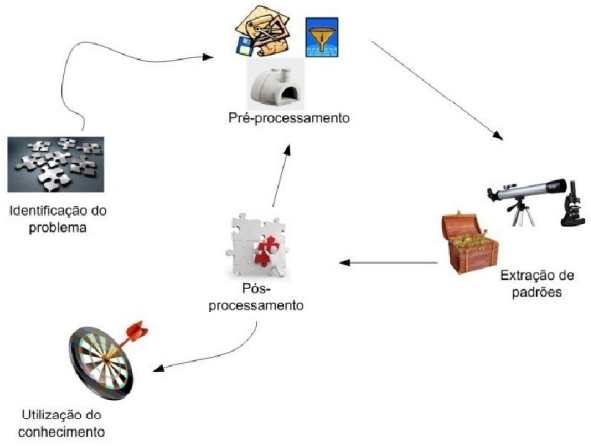

Fig 5 - Etapas do processo de *DM*. Fonte: Adaptado de Rezende (2005).

- Extração de Padrões - etapa direcionada ao cumprimento dos objetivos definidos na identificação do problema. Aqui é realizada a escolha das tarefas de *DM* a serem empregadas e a configuração e execução de uma ou mais técnicas para extração de informação.
- Pós-Processamento – a informação extraída é analisada para verificação de sua relevância. Caso ela 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.

Pela Figura 5, 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.

2.2.3.1 Técnicas e Tarefas de *DM*

As técnicas podem ser consideradas ferramentas utilizadas para atender aos propósitos do *DM*. *N*ão 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 da técnica ao propósito da análise e a familiaridade com a técnica a ser utilizada (Han; Kamber, 2001). Alguns exemplos de técnicas para algumas funcionalidades são descritas abaixo:

- Descoberta de regras de associação – estabelece uma correlação estatística entre atributos de dados e conjunto de dados;
- Árvores de decisão - hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos;
- Raciocínio baseado em casos - baseado no método do vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança;
- Algoritmos Genéticos - métodos gerais de busca e otimização, inspirados na Teoria da Evolução, na qual a cada nova geração, soluções melhores têm mais chance de ter "descendentes" ;
- Redes Neurais Artificiais - modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões.

As tarefas, também chamadas de funcionalidades, são a maneira como os resultados serão apresentados. Técnicas e tarefas são definidas na etapa de extração de padrões. Dependendo da técnica, os algoritmos correspondentes são escolhidos para sua execução (Rezende, 2005). A Figura 6 mostra as interações entre, técnicas, tarefas (funcionalidades) e algoritmos.

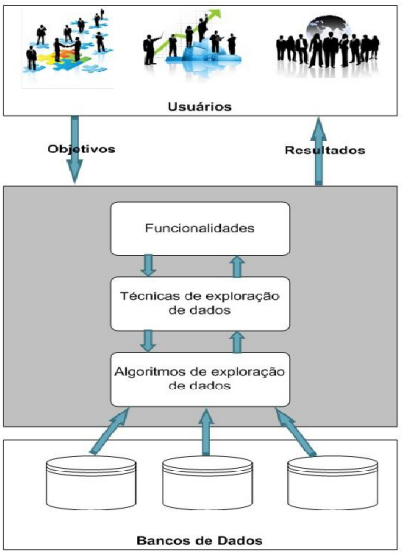

Fig 6 - Interatividade entre as técnicas e tarefas de *DM*.

- Muitos autores definem uma quantidade diferenciada de tarefas para *DM*, uns mais, outros menos em suas definições, como mostrado a seguir:
- Previsão, Identificação, Classificação e Otimização (Elmasri; Navathe, 1999);
- Descrição e Predição (Han; Kamber, 2001);
- Classificação, Regressão, *Clusering,* Sumarização, Modelos de Dependência, Escolha e Detecção de Desvios (Fayyad; ; Piatetsky-Shapiro; Smyth, 1996);
- Classificação, Regressão, Regras de Associação, Sumarização, *Clustering* e Outras (Rezende, 2005);
- Classificação, Regressão, Associação, *Clustering* e Sumarização (Dias, 2002);

Porém, a maioria deles concorda que essas tarefas sejam classificadas em 2 grandes grupos, como mostra Rezende (2005) na Figura 7.

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,

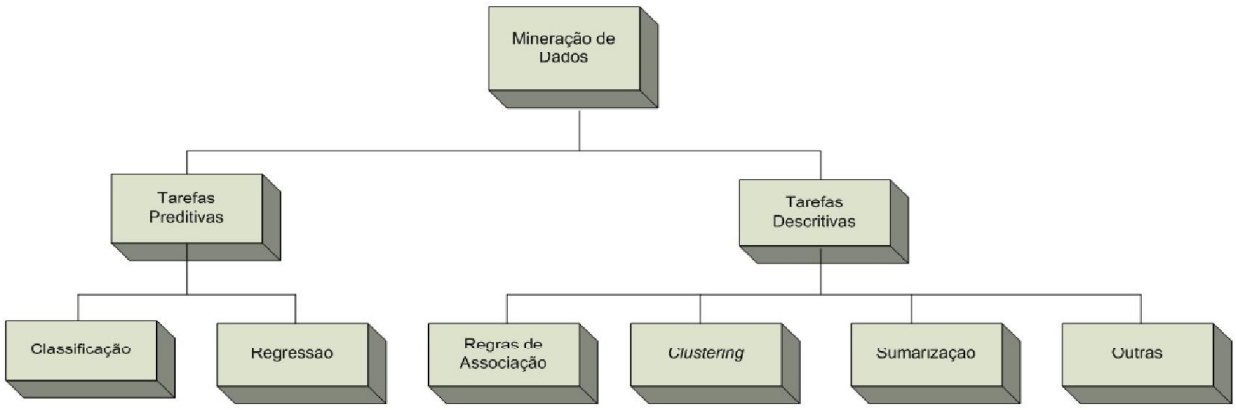

Fig 7 - Tarefas de *DM.* Fonte: Adaptado de Rezende (2005).

2005).

2.2.4 Técnicas de Visualização de Dados

` Existem diversas maneiras de representar a informação extraída 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 da informação permite a representação de dados em formas gráficas ou mesmo espacializadas, permitindo ao usuário utilizar sua percepção visual para otimizar o processo de interpretação desses resultados (KEIM, 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 visual dos dados) visa à inserção do ser humano como parte essencial do processo, aplicando suas habilidades de percepção para a análise dos resultados gerados a partir 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 exploratória de dados.

Para este trabalho, foram escolhidas duas ferramentas computacionais capazes de gerar esse tipo de resultado visual aliado a algoritmos de mineração de dados, são elas: Matlab (http://www.mathworks.com) e SODAS (http://www.ceremade.dauphine.fr/SODAS/).

3. ESTUDO DE CASO: APLICAÇÃO DO PROCESSO DE *KDD/DM* PARA O CTM DE RIBEIRÃO DOS ÍNDIOS – SP

O levantamento cadastral, não acontece todos os anos, porém, quando é executado, uma grande quantidade de dados é coletada e inserida nos Bancos 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 (*Structure Query language)*.

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

A prefeitura municipal de Ribeirão dos Índios vem fazendo desde 1996 em conjunto com a Faculdade de Ciências e Tecnologias da Universidade Estadual Paulista (FCT/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 possui 2187 habitantes em uma área de 197 km^2. Está localizada a oeste do estado de São Paulo, como mostra a Figura8.

3.1 Levantamentos Cadastrais

Ao todo foram realizados quatro levantamentos cadastrais, sendo eles nos anos de 1996, 2004, 2010 e 2012. Somente no 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 sistemas de bancos de dados distintos (1996 – Dbasae, 2004 – Access, 2010 – PostgreSQL) 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, 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.

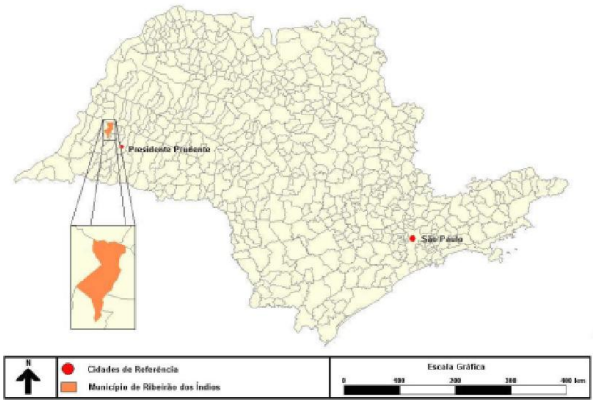

Fig 8 - Localização do Município de Ribeirão dos Índios (DINIZ 2004).

Apesar de existirem 4 levantamentos, o estudo de caso trabalhará somente com os últimos três, 2004, 2010 e 2012. O levantamento de 1996 foi descartado desse trabalho pois percebeu-se que muitos dados estavam sem respostas, o que poderia atrapalhar na qualidade dos resultados finais.

3.2 Etapas do processo de *KDD/DM*

A partir da escolha dos três levantamentos para o processo, as etapas de DM propostas por Rezende (2005) começaram a ser executadas.

3.2.1 Definição do domínio do problema

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, dois questionamentos foram submetidos ao processo:

Buscar uma relação da renda familiar com os seguintes dados: padrão construtivo, área construída e educação, para os levantamentos de 2004 e 2010;

Acompanhar a evolução de uma determinada patologia visando encontrar um padrão nos resultados, para os levantamentos de 2004, 2010 e 2012;

3.2.2 Pré-Processamento

Nessa fase inicia-se o trabalho de preparação das bases de dados para a extração de padrões. Os levantamentos foram unificados em um SGBD, formando uma única fonte de dados.

Após isso, precisou-se realizar a adequação dos dados unificados, pois alguns problemas foram encontrados, como por exemplo, quantidades de atributos diferentes de uma base para outra e até mesmo possibilidades de respostas diferentes para uma mesma pergunta de um levantamento para outro.

Em seguida, começa o processo de limpeza dos dados, pois como o processo de armazenamento foi feito de forma manual, alguns erros de digitação foram encontrados e alguns questionamentos com respostas em branco também. Esses registros foram excluídos da base unificada.

Finalizou-se essa etapa de duas maneiras: a primeira com a seleção somente dos dados referentes ao domínio do problema escolhido

e a segunda com a inserção de alguns campos textuais para que ferramentas de mineração distintas pudessem ser utilizadas.

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 alguns grupos classificatórios e esses valores foram inseridos dentro da tabela virtual como novos atributos.

Normalmente, não se executam essas etapas de pré-processamento quando se utiliza um SIG Cadastral para obtenção desses resultados, mesmo porque, na grande maioria dos casos, os dados são atualizados dentro dos Bancos Cadastrais, perdendo assim os dados anteriores. Isso fortalece a decisão de utilizar a técnica proposta neste trabalho uma vez que num processo convencional (consulta por SQL), os resultados seriam limitados e demonstrados de forma tabular, o que dificultaria muito a decisão de definir o que espacializar.

3.2.3 *DM* – Extração de Padrões e Pós-Processamento

Nessa fase, duas ferramentas foram utilizadas para extração de padrões, Matlab e SODAS. A ferramenta Matlab é 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 ferramenta SODAS, é *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 Neurais 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* (KOHONEM, 2001).

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 é mostrada em forma de um conjunto de "casulos" formando uma "colmeia". Dentro de cada "casulo", que representa um registro analisado, um gráfico linear processado por meio dos dados de entrada é construído. Os grupos homogêneos serão separados por cores definidas pela própria função.

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 a análise de conjunto de dados complexos a fim de "descobrir conhecimento" de tais dados. "Descobrir 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.

Como resultado da análise do primeiro caso, foram geradas duas colmeias contendo agrupamentos de comportamento dos dados referentes às parcelas. A Figura 09 mostra o resultado da análise entre a renda familiar e a área construída, e a Figura 10 mostra o resultado da análise entre a renda familiar e o padrão construtivo.

Iniciando a fase de pós-processamento do primeiro caso, analisou-se o primeiro resultado mostrado na Figura 10, 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 espacializada para uma melhor interpretação. Utilizando mais duas ferramentas computacionais (gvSIG e a extensão PostGIS do PostgreSQL) gerou-se o mapa dessa constatação, que é mostrado na Figura 11.

Por ser uma situação que aparentemente não ocorreria, uma pesquisa foi feita dentro da

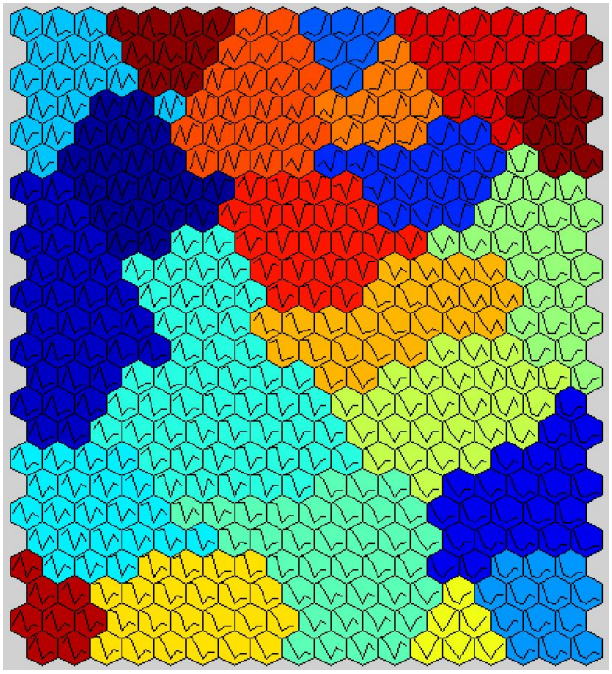

Fig 9 - Representação gráfica em colméia referente à análise entre renda familiar e área construída.

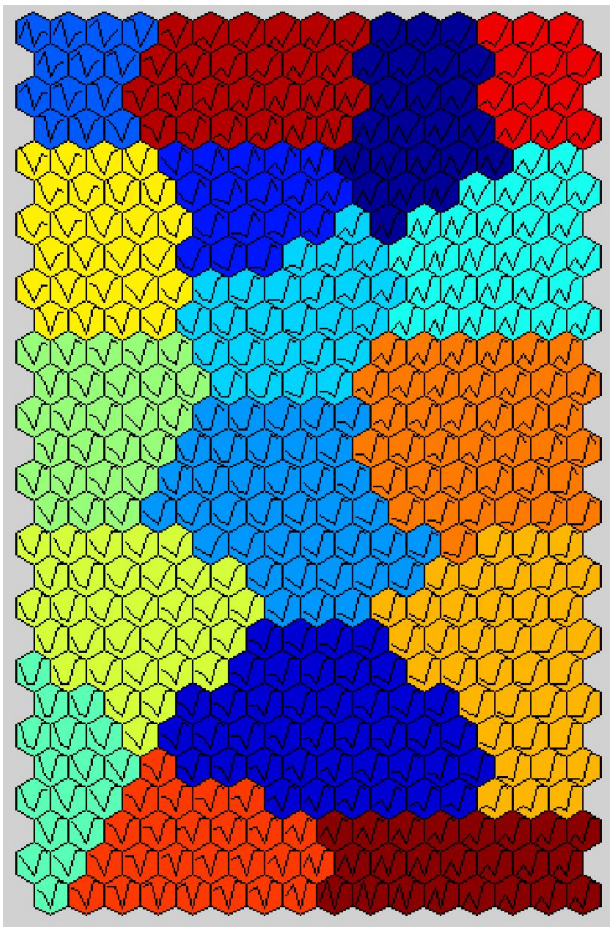

Fig 10 - Representação gráfica em colméia referente à análise entre renda familiar e padrão construtivo.

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. A espacialização desses dados permite definir exatamente onde os cadastradores deverão voltar para buscar os dados corretos e atualizar o Banco de Dados Cadastral.

Na segunda ferramenta, além dos dados anteriores, acrescentou-se dados sobre educação. Procurou saber se uma mudança na renda familiar também poderia influenciar diretamente na escolhe da escola para os filhos, ou seja, particular ou pública. 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 12. A legenda dessa figura representa as classes de oscilação da renda familiar definidas como o objeto simbólico e os eixos do gráfico representam a oscilação da área construída, do padrão construtivo, de membros matriculados na educação pública e particular.

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 espacializada. O resultado é mostrado na Figura 13.

O resultado obtido pelo SODAS mostra uma visão diferente da obtida pelo Matlab. No SODAS o especialista pode criar agrupamentos e verificar por meio de gráficos de barras se ocorre algum padrão de comportamento nos mesmos , enquanto no Matlab são analisados todos os registros e o resultado é demonstrado por meio de gráficos lineares. Porém, ambos os resultados podem ser mais facilmente interpretados e espacializados de acordo com o interesse do gestor.

No segundo caso (referente à evolução da patologia de hipertensão arterial apresentada nos levantamentos de 2004, 2010 e 2012), somente o Matlab foi usado na Mineração. Os dados da

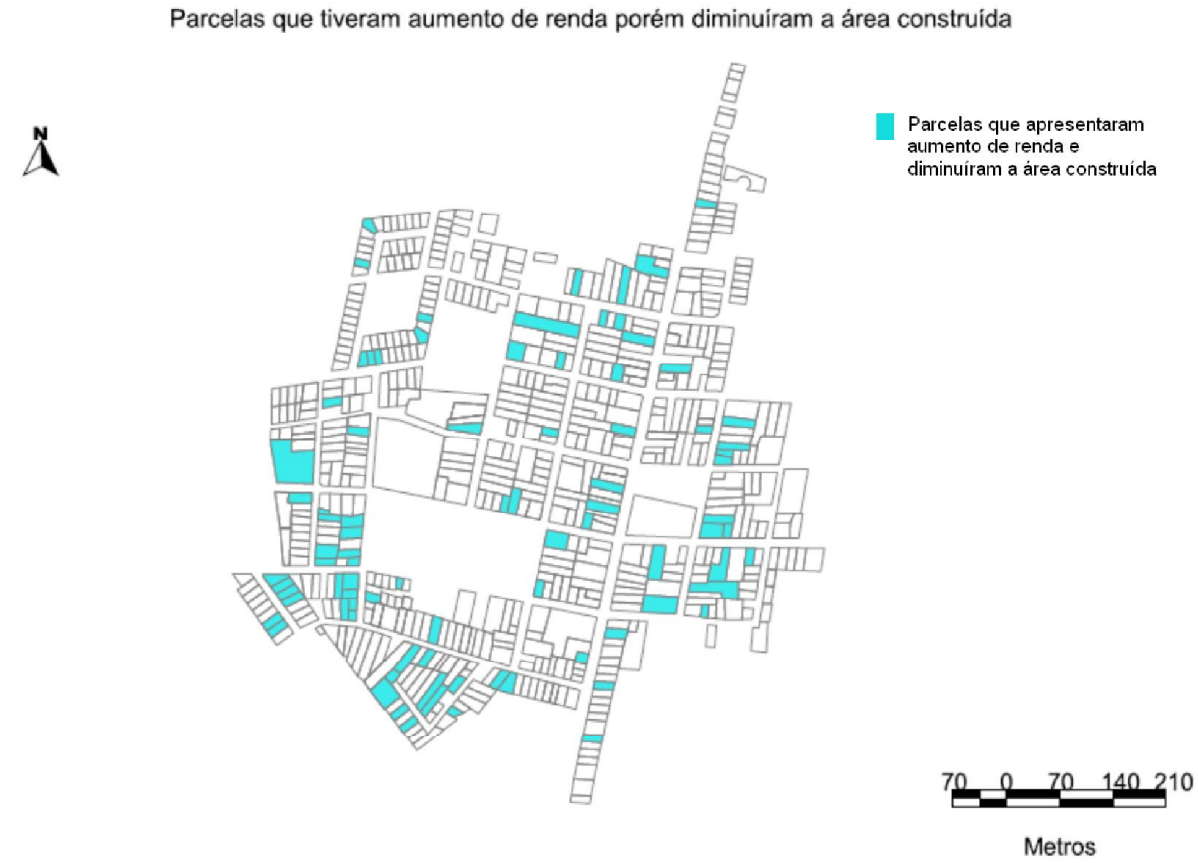

Fig 11 - Espacialização que mostra as parcelas que aumentaram a renda familiar e diminuíram a área construída.

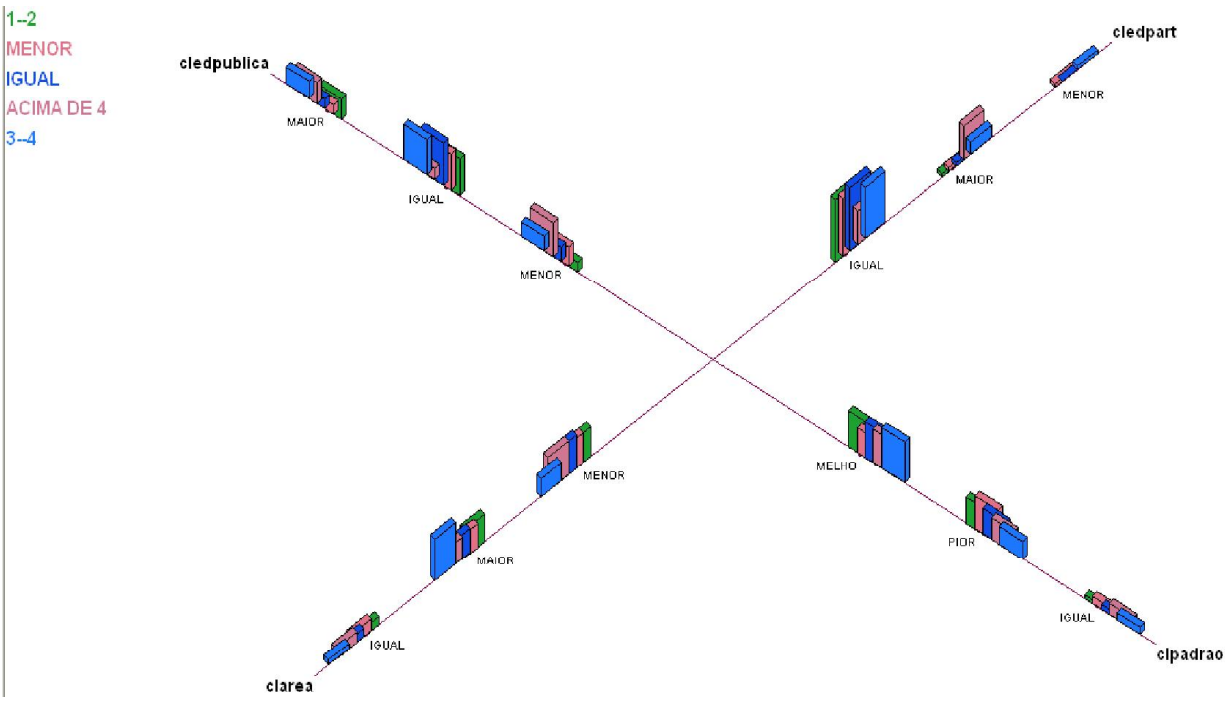

Fig 12 - Resultado da análise de dados feita pelo SODAS envolvendo as classes inseridas na tabela (cledpublica – parcelas com indivíduos em escolas públicas, cledpart – parcelas com indivíduos em escolas particulares, clarea –tamanho da área construída, clpadrao – mudanças no padrão construtivo) agrupadas pela renda familiar.

Fig 13 - Espacialização que mostra as parcelas que tiveram aumento de renda familiar entre 3 e 4 salários mínimos e investiram em padrão construtivo.

tabela virtual foram exportados para o Microsoft Excel e carregados para processamento do algoritmo resultando na representação gráfica em colmeia mostrado na Figura 14.

Apesar do conjunto de dados ter sido pequeno para esse tipo de Mineração, algumas classes foram identificadas e analisadas. e foi escolhida 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 15.

Percebe-se no mapa que não ocorreu nenhum padrão referente à localização das parcelas. Isso não descredencia o resultado visto que para o gestor é importante perceber que em seu município, esse aumento de quantidade de hipertensos por parcela não está acontecendo em um único bairro, e sim, são casos esporádicos que podem ser tratados com os postos de saúdes existentes ou mesmo com um programa de visita domiciliar de assistentes de saúde.

4. CONCLUSÕES

A importância da descoberta 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 atualmente na

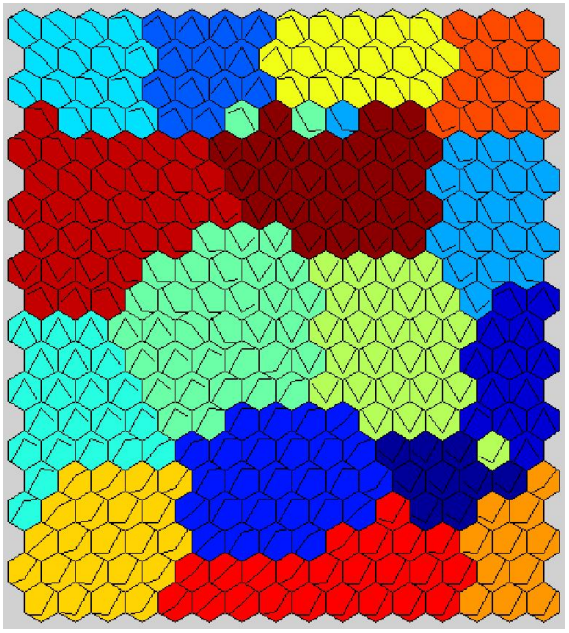

Fig 14 - Representação gráfica em colméia referente à análise da evolução da patologia hipertensão arterial.

Fig. 15 - Espacialização que mostra as parcelas que tiveram aumento acima de um caso na patologia hipertensão arterial em 2012.

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

A descoberta 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, isto é, do histórico. (AMORIM; MALAMAN; SASS, 2013) discutem a necessidade e a importância da atualização cadastral sem a perda do histórico de dados dos Bancos de Dados Cadastrais nas atividades de planejamento urbano.

As inovações tecnológicas aparecem como aliadas nesse processo pois periodicamente novas ferramentas surgem para manipulação de 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 pelos Gestores Públicos, visto que, no estudo de caso apresentado, os levantamentos foram armazenados em bases heterogêneas, precisando que todas as etapas da *DM* fossem executadas. Ainda assim, após a unificação e limpeza da base de dados, o resultado de uma consulta SQL traria informações tabuladas, não proporcionando ao gestor uma facilidade na interpretação dos mesmos.

KDD e *DM* se apresentam nesse contexto para trazer resultados de diferentes formatos, facilitando e agilizando a interpretação e avaliação do tomador de decisão.

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

A presença do profissional especialista no domínio da aplicação 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.

O estudo de caso envolveu um Banco de Dados Cadastral pouco volumoso de um município pequeno, fato esse que fere um pouco o rigor das experiências requeridas pelas técnicas utilizadas.. Entretanto foi possível observar a validade da aplicação do processo de KDD e do uso de técnicas de visualização, bem como as vantagens em criar e manter organizado e atualizado um banco de dados com dados históricos. Os resultados mostraram, também, que os benefícios se aplicam a vários setores além do tributário, colaborando de forma ampla com os gestores públicos.

Dessa forma pode-se concluir que este trabalho pode servir como parâmetro de comparação e/ou ponto de partida para outros trabalhos, principalmente, no que diz respeito à necessidade da aplicação de novos conceitos de modelagem de Banco de Dados Cadastral para aplicação das técnicas de *KDD* visando o aprimoramento do processo de tomada de decisão nas administrações municipais.

REFERÊNCIAS

AMORIM, A.; MALAMAN, C. S.; SASS, G. G. A modernização dos processos de atualização cadastral e as análises temporais. **Revista Brasileira de Cartografia**, Nº 65/2, 2013, p. 375-382.

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**, Nº 56/1, 2004, p. 47-54.

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, p. 265.

DATE, C. J. **Introdução a Sistemas de Banco de Dados.** Tradução: Vandenberg Dantas de Souza. 7ª ed. americana. Rio de Janeiro: Campus, 2000, p.9;599;611.

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, p. 3-8.

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, p. 626-627.

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, p. 17-20. 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 fev. 2013.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, Volume 17, n.3, p. 1-6, 1996.

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 fev. 2013.

FREITAS, C. M. D. S.; CHUBACHI, O. M.; LUZZARDI, P. R. G.; CAVA, R. A. **Introdução a visualização de informações**. Revista de

informática teórica e prática, Volume 8, Número 2, 2001, p. 143-158.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques.** 2ª ed. Morgan Kaufmann Publisher, 2005, p. 7, 285.

KEIM, D. A. **Information Visualization and Visual Data Mining.** IEEE Transactions on Visualization and Computer Graphics. Vol. 7, Number 1, 2002, p. 100 – 106.

KOHONEN, T. Self-Organinzing Maps**. Proceedings of the IEEE**, Vol. 78, Number 9, 1990. P. 1464 – 1477.

LOCH, C.; ERBA, D. A. **Cadastro técnico multifinalitário:** rural e urbano. Cambridge, MA: Lincoln Institute of Land Policy, 2007, p. 104-112.

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, p. 2.

O'BRIEN, James A. **Sistemas de Informação e as decisões gerenciais na era da Internet.** 2 ed. São Paulo: Saraiva, 2004, p.12-13.

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. Disponível em: http://www.ufpe.br/cgtg/SIMGEOII_CD/Organizado/cad/029.pdf. Acesso em: 10/02/2013. p 1-4.

REZENDE, S. O. **Sistemas Inteligentes – Fundamentos e Aplicações.** Barueri: Manole, 2005, p. 309 – 311.

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.

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.