Utilização de Data Mining no Cadastro Territorial Multifinalitário

Prof. Sergio Ricardo Ribas Sass [1]
Prof. Dr. Amilton Amorim [2]
Profª. 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
sergio.sass@ifms.edu.br

[1,2,3] Universidade Estadual Paulista - UNESP
Departamento de Cartografia
Rua Roberto Simonsen, 305
19060-900 – Presidente Prudente - SP
amorim@fct.unesp.br
glaucia@comp.uems.br

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 dessas 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).* Esse artigo discute brevemente a tecnologia de *DM, e,* aliado ao CTM, mostra os resultados de um experimento preliminar para a cidade de Ribeirão dos Índios-SP.

Palavras chaves: Banco de Dados, tomada de decisão, Cadastro Territorial Multifinalitário, *Data Warehouse, 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 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 article briefly discusses the technology of the DM, and, allied to the MC shows the results of a preliminary experiment for the town of Ribeirão dos Índios, SP.

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

1 Introdução

O CTM é uma ferramenta que vem auxiliando a gestão pública no planejamento urbano agilizando o processo de desenvolvimento. Porém, a estrutura de uma base cadastral de um CTM necessita de acompanhamento e dedicação para que seus objetivos pretendidos sejam alcançados. Acontece que, com a evolução do CTM, a quantidade de dados armazenados e a heterogeneidade desses dados causaram uma grande complexidade na obtenção de informações adequadas e úteis para o auxílio do gestor.

O Sistema de Informação territorial (SIT) que dá suporte à existência do CTM, por ter como base de dados um Sistema Gerenciador de Banco de Dados (SGBD) tradicional, não consegue tratar esses valiosos dados na busca por conhecimento.

A tecnologia de *DM* aparece para sanar esse problema visando o tratamento adequado desses dados e buscando a extração de conhecimento para auxílio no processo decisório. Neste artigo busca-se apresentar as características do *DM*, além de verificar sua importância para a gestão pública aliando-o ao CTM.

2 Revisão da Literatura

O processo de *DM*, com aplicação de algoritmos específicos para extração de padrões, vem de encontro à necessidade de resolver os problema decorrente da era digital: a sobrecarga de dados (Fayyad; Piatetsky-Shapiro; Smyth, 1996).

Para isso é necessário aliar a tecnologia de Banco de Dados com a tecnologia de Sistemas Inteligentes. Precisa ficar claro aqui que o processo de tomada de decisão não faz parte da tecnologia de Banco de Dados e sim faz uso dela. Na verdade são várias tecnologias que usam os Bancos de Dados para dar suporte a Sistemas Inteligentes como os sistemas de apoio à decisão por exemplo. São elas: *DW*, *Data Mart*, Depósitos de Dados Operacionais, *On-Line Analytical Processing (OLAP)*, Bancos de Dados Multidimensionais, *DM*, entre outros (Date, 2000).

As próximas subseções apresentam uma breve discussão sobre a tecnologia de *DM,* seus conceitos, tecnologias relacionadas, suas etapas e funcionalidades.

2.1 Extração de Conhecimento

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 *Knowledge Discovery Database (KDD)* (Rezende, 2005). *DM* entra como parte particular desse processo com objetivo de aplicação de algoritmos específicos para extração de padrões, como mostra a Figura 1(Fayyad; Piatetsky-Shapiro; Smyth, 1996a).

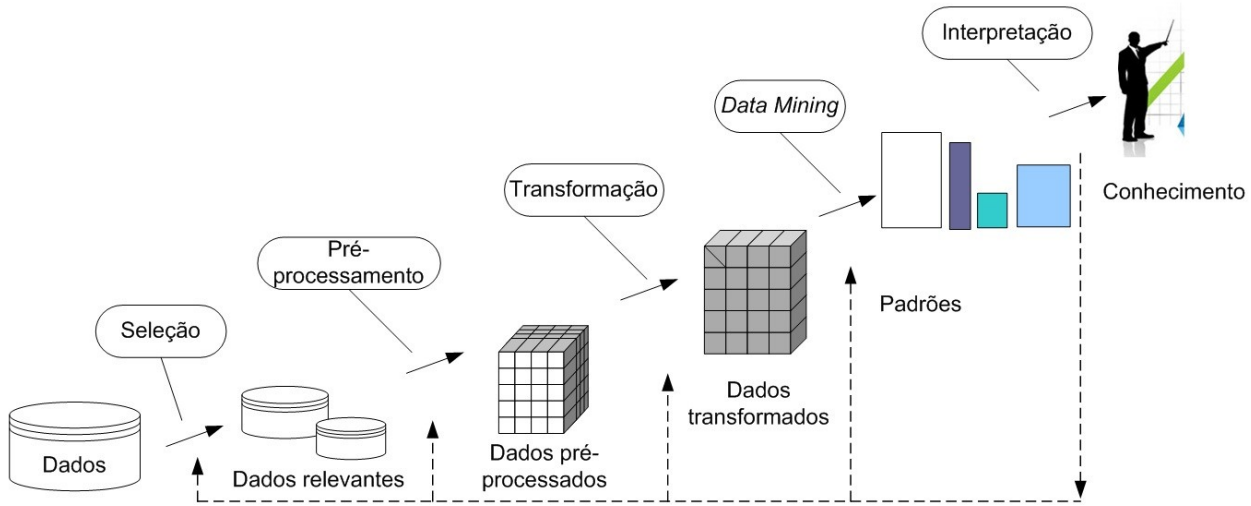

Figura 1 : Uma visão geral das etapas que compõe o processo KDD
Fonte: Adaptado de (Fayyad; Piatetsky-Shapiro; Smyth, 1996a)

De acordo com Han, Kamber e Pei (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 - remove ruídos e inconsistências de dados;
3. Transformação - os dados são apropriadamente transformados para mineração através da realização de operações de agregação, por exemplo;
4. *DM* - processo essencial no qual 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.

2.2. Tecnologias Relacionadas

Para definir conceitos de *DM* é preciso entender e pontuar duas tecnologias envolvidas no seu uso. São elas: Banco de Dados e *DW*.

Segundo Date (2000), "Um banco de dados é uma coleção de dados persistentes utilizada pelos sistemas de aplicação de uma determinada empresa." Persistentes no sentido de que, a partir da hora que um dado for armazenado dentro do Banco de Dados, ele persistirá dentro de sua base até uma solicitação de edição ou exclusão desse dado.

Elmasri, Navathe (2005) acrescentam que "banco de dados é um conjunto de dados relacionados", e que, "os Bancos de Dados tradicionais são transacionais", ou seja, operam com processos rotineiros de transações.

O acúmulo desses dados em Bancos de Dados tradicionais, ou seja, todos separados em tabelas relacionadas entre si, podendo chegar a dezenas e até centenas de tabelas, faz com que empresas desperdicem valiosas informações que poderiam ser geradas com ajudas de novas tecnologias. Esses Bancos de Dados são limitados a processamentos *on-line* de transações e consultas, ou seja, inclusões, atualizações, exclusões e consultas. Com isso, se limitam ao processamento de dados de somente pequenas partes do Banco de Dados(Elmasri; Navathe, 2005).

Essa limitação ocorre em virtude da linguagem de consulta dos Bancos de Dados tradicionais comportar operações da álgebra relacional, permitindo somente seleção de linhas e colunas de dados e informações relacionadas por junção entre atributos comuns. O *DW* por sua vez, possibilita a mesma visualização em múltiplas dimensões dando suporte ainda a execução de diversas ferramentas tecnológicas como o *DM*, por exemplo. São otimizados para recuperar dados e não processar transações (Elmasri; Navathe, 2005).

Inmon (2005) define o termo *DW* como "um conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo de apoio às decisões gerenciais" .

"O *DW* é um tipo especial de banco de dados." (Date, 2000). Vieram suprir a necessidade de fornecer uma origem de dados única e consistente para apoio à decisão (Date, 2000). São diferentes dos Bancos de Dados tradicionais em estrutura, funcionalidade, desempenho e propósito (Elmasri; Navathe, 2005).

O *DW* é orientado a assuntos e atende a interesses específicos, organizados e resumidos por temas, como vendas, *marketing*, finanças, distribuição e transporte. A preocupação é com os dados e não com os processos que os modificam. (Rob; Coronel, 2011) (Machado, 2000). Portanto, a proposta do *DW* é auxiliar a tomada de decisão sustentada por dados.

2.3. *DM*

A maioria das definições de *DM* utilizada pelos autores foi elaborada por Fayyad Piatetsky-Shapiro e 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). Essa estrutura adequada se dá através da criação do *DW*.

A Figura 2 mostra as etapas do processo de *DM* adotado por Rezende (2005), e também adotado nesse artigo, onde os termos *KDD* e *DM* são tratados com mesmo significado ou seja, referenciando o processo de extrair conhecimento a partir de dados. A justificativa para essa adoção está no sentido de que todo processo anterior ao *DM*, como mostra a Figura 1, é feito dentro da etapa de pré-processamento definido por ela.

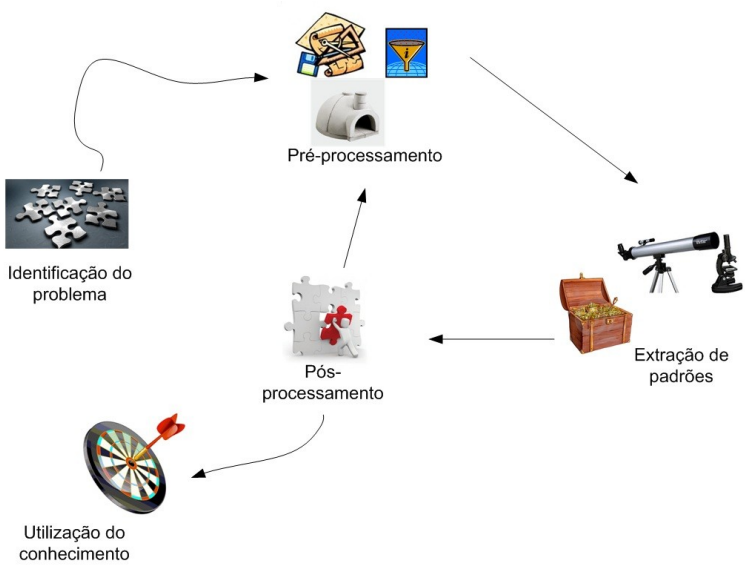

Figura 2 : Etapas do processo de *DM*
Fonte: Adaptado de (Rezende, 2005)

As etapa do processo de DM são (Rezende, 2005):

- 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 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, *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.
- 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.
- Pós-Processamento - 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.

Pela Figura 2, 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.

2.3.1 Técnicas e tarefas de *DM*

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 familiaridade com a técnica a ser utilizada. 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, onde 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.
-

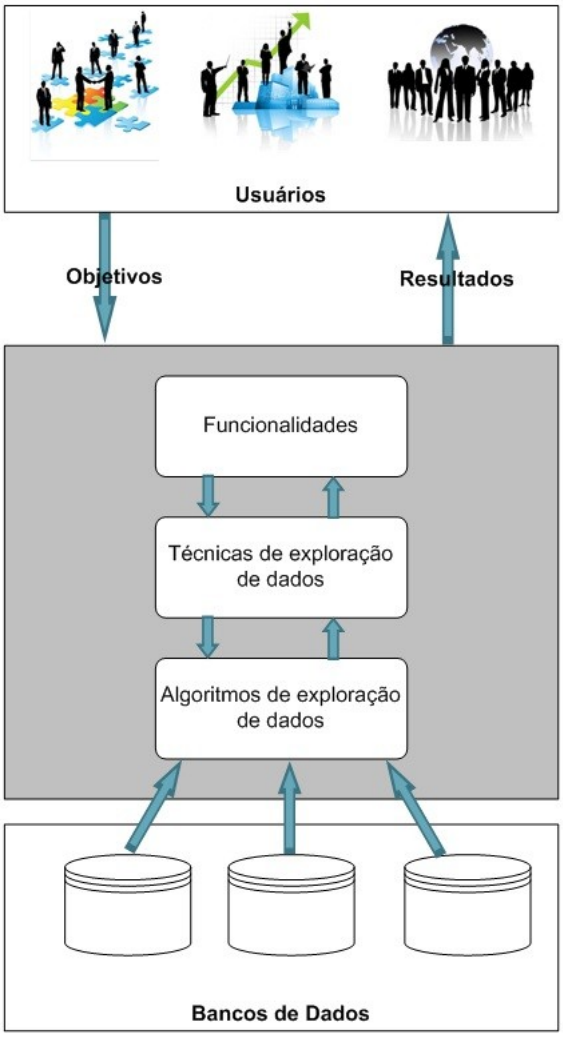

Figura 3 : Interatividade entre as técnicas e tarefas de *DM*

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 3 mostra as interações entre, técnicas, tarefas (funcionalidades) e algoritmos.

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, 2001);
- Classificação, Regressão, *Clusering,* Sumarização, Modelos de Dependência, Escolha e Detecção de Desvios (Fayyad; ; Piatetsky-Shapiro; Smyth, 1996a);
- 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 4.

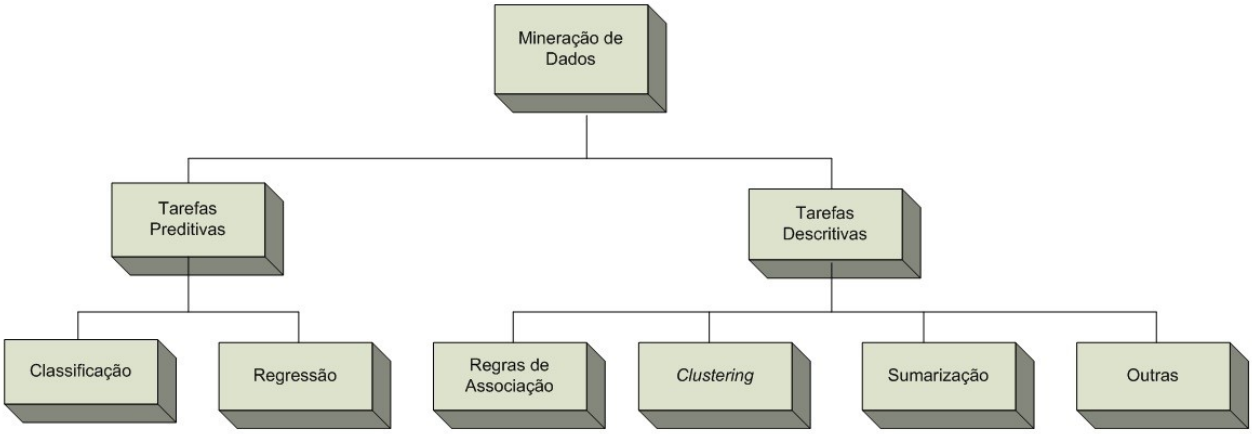

Figura 4 : Tarefas de *DM*
Fonte: Adaptado de (Rezende, 2005)

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).

3. Estudo de Caso: Aplicação de *DM* para o CTM de Ribeirão dos Índios - SP

Para demonstrar a utilização do *DM* com CTM foram utilizados bases de dados de dois levantamentos cadastrais realizados em Ribeirão dos Índios – SP nos anos de 2004 (Amorim, Souza e Dalaqua, 2004) e 2010 (Malaman e Amorim, 2010).
Seguindo os passos de Rezende(2005), primeiramente o domínio do problema foi identificado, ou seja, buscou-se conhecer se o aumento da renda familiar estava diretamente ligado ao aumento do padrão construtivo e ao aumento da área construída.

Na segunda etapa, pré-processamento, criou-se uma única fonte de dados em virtude dessas bases estarem em fontes heterogêneas. O Banco de dados escolhido para essa unificação foi o *PostgreeSQL.* Esses dados foram transformados e diversas *views*[1] foram criadas selecionando somente os campos relativos ao domínio do problema, renda, padrão construtivo e área construída. Esses registros foram filtrados de maneira que foram mantidos somente os que estavam corretamente preenchidos, processo conhecido como limpeza. Após todo esse processo, os dados foram normalizados para uso adequado do algoritmo de *DM* e são mostrados nas figuras 5 e 6.

1*view* é uma relação que não armazena dados, composta dinamicamente por uma consulta que é previamente analisada e otimizada

	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 5 : 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 6 : *Dados normalizados de renda familiar e padrão construtivo*

O software Matlab versão 7 foi utilizado na etapa de extração de padrões, através da técnica de Redes Neurais Artificias e da tarefa de *Clustering,* para geração dos resultados gráficos.

Na Figura 7 foram utilizados informações sobre renda familiar e padrão construtivo. São identificados as relações entre a mudança de renda e os investimentos em padrão construtivo nos anos de 2004 e 2010. Com a visualização pode-se observar alguns padrões dessa relação.

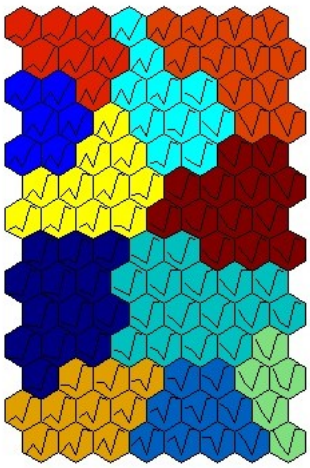

Figura 7 : *Clustering* da relação entre renda familiar e padrão construtivo.

Na Figura 8, é a relação entre renda familiar e tamanho da área construída nos anos de 2004 e 2010.

Essas figuras geradas mostram um agrupamento de pequenos gráficos relativos à oscilação da renda familiar em conjunto com o padrão construtivo e com a área construída, separados por uma coluna de “zeros” para uma melhor leitura dos mesmos. Em virtude da quantidade de registros fazer com que a figura gerada ficasse grande, optou-se pela geração de uma figura menor eliminando parte dos imóveis que mantinham padrões de comportamento idênticos. Esse processo de eliminação foi feito pelo próprio algoritmo.

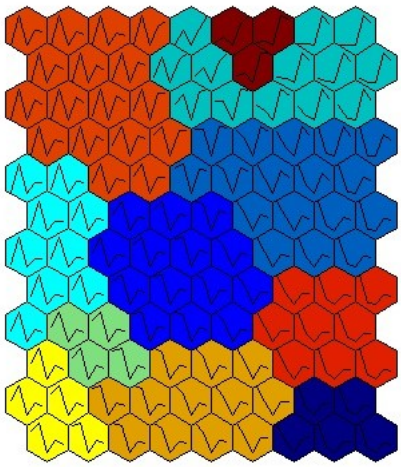

Figura 8 : *Clustering* da relação entre renda familiar e área construída.

As cores simbolizam os grupos parecidos, imóveis que tiveram comportamentos similares de renda com padrão construtivo e renda com área construída.

Na última etapa do processo, pós-processamento, os resultados são interpretados e validados pelo gestor. Por meio dessas figuras, apesar de simples, padrões foram identificados e possuem conhecimentos que podem ser utilizados de maneiras a apoiar a gestão pública

7. Conclusões

Esse trabalho de *DM* com CTM, ainda se encontra em fase preliminar de desenvolvimento. O algoritmo ainda passará por refinamentos buscando melhoria na qualidade dos resultados gerados. Porém, a partir desses primeiros resultados, conclui-se que não existe um padrão criterioso entre aumento de renda familiar e aumento de padrão construtivo ou mesmo aumento na área construída, mas pode-se dizer que o contrário é bem sustentado com os resultados, ou seja, a grande maioria que aumentaram o padrão construtivo e área construída tiveram um aumento da renda familiar.

A extração de conhecimento em grandes bases de dados utilizando *DM* pode trazer recompensas valiosas em diversos setores, auxiliando gestores no processo de tomada de decisão. Os conhecimentos extraídos pode ser incorporados aos critérios utilizados pelo gestor na elaboração do planejamento estratégico do desenvolvimento urbano.

Entretanto, vale salientar que para cada objetivo desejado devem-se aplicar tarefas e técnicas específicas para se conseguir qualidade nos resultados esperados e que para que isso seja realmente possível, é necessário que os dados a serem carregados na base do CTM sejam levantados com a máxima seriedade, para que os resultados possam ser interpretados e transformados em aliados do gestor.
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.

Neste artigo foram apresentados conceitos, técnicas e funcionalidades do processo de *DM* e um estudo de caso preliminar da utilização do *DM* com o CTM.

8 Referências

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.

Berry, M. J. A.; Linoff G. *Data Mining Techniques for Marketing, Sales and Customer Support.* John Wiley & Sons, Inc. 1997.

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. 1715-1725, Maringá, 2002.

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

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. *The KDD Process for Extracting Useful Knowledge from Volumes of Data*. Communications of the ACM , Volume 39, Number 11. 1996a.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. *From DM to Knowledge Discovery in Databases*. AI Magazine, Volume 17, Number 3. 1996b.

Han, J.; Kamber, M.; Pei, J. *DM: Concepts and Techniques.* 2ª ed. Morgan Kaufmann Publisher, 2005.

Harrison, T. H. *Intranet Data Warehouse*. São Paulo: Berkeley Brasil 1998.

Inomn, W. H. *Building the data warehouse*. 4ª ed. Indianapolis: Wiley Publishing, 2005.

Machado, F. N. R. *Projeto de data warehouse: Uma Visão Multidimensional.* São Paulo: Érica, 2000.

Malamam, 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.

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

Rob, P.; Coronel, C. *Sistemas de banco de dados: projeto, implementação e gerenciamento.* Tradução: All Tasks. Revisão Técnica: Ana Paula Appel. 8ª Ed. norte-americana. São Paulo: Cengage Learning, 2011.

Agradecimentos
À Instituto Federal de Mato Grosso do Sul – IFMS unidade de Nova Andradina - MS pelo apóio a esse trabalho. Ao Programa de Pós-Graduação em Ciências Cartográficas da FCT/UNESP de Presidente Prudente -SP, pelos laboratórios e apoio para o desenvolvimento da pesquisa.
Ao Grupo de Pesquisa em Análise e Representação de Dados Espaciais – GARDE.