

A Influência da Informação Espacial no Processo de Descoberta do Conhecimento

Jorge Henrique de Castro

UERJ – Mestrado em Engenharia de Computação - Concentração Geomática
Petrobras - Gás e Energia – Planejamento Estratégico
Av. Almirante Barroso, 81 33º andar
23030-003 Rio de Janeiro RJ – Brasil
jorgecastro@petrobras.com.br

Resumo – As organizações têm se mostrado dispostas a armazenar e tratar volumes cada vez maiores de dados referentes a seus produtos e clientes. O foco das análises é estabelecer a descoberta de padrões de comportamento que permitam uma atuação mais eficiente nos respectivos segmentos de mercado. Determinar onde, o que e para quem vender é fundamental chave na sobrevivência da empresa. Com base nisto o emprego de ferramentas e técnicas de avaliação destes padrões cresce substancialmente, porém encontra-se uma lacuna de atuação sobre a informação que permite determinar a correta distribuição da base de clientes no espaço: o cadastro de endereços. Este artigo aborda a relação entre os cadastros de clientes e endereços e o processo de descoberta do conhecimento, a necessidade do emprego de plataformas SIG para a garantia de um bom cadastro de endereços, bem como a consequência das inferências sobre este cadastro e os erros cadastrais de maior incidência.

Palavras chave: base de conhecimento, análise espacial e cadastro.

Abstract - Nowadays the enterprises store and manipulate a big amount of data related to both products and customers. The aim of analysis is to establish pattern discovery that allows to company act more efficiently in different market segment. It is fundamental to the company to know where, what and to who sell their products in order to company survivor. Based on it the use of techniques and tools of pattern evaluation grows substantially. Therefore, there is a lack about the information that allows to determine the correct distribution of customer's spatial database that is the address cadastre. This paper presents the relationship between both customers and client cadastre and Knowledge Discovery in Database, the need of GIS to guarantee a reliable address cadastre, as well as the impacts of inferences in this cadastre and the most common cadastral mistakes.

Keywords: knowledge database, spatial analysis and cadastre.

1 – Introdução

O processo de descoberta do conhecimento, ou melhor, de construção do conhecimento envolve a cadeia de transformação de dados em conhecimento propriamente dito. Os dados, ou a matéria bruta para este processo, são as evidências mais básicas para uma investigação. Guardam os aspectos dos fenômenos que estão sendo estudados. Uma vez coletados registram os acontecimentos com certo grau de precisão e qualidade (Côrtes, 2003).

A informação é o resultado da organização, transformação e/ou análise de dados. É oriunda do processo de tratamento de um conjunto de dados com o intuito de produzir significado. Ressalta-se que isto é válido quando este tratamento é realizado de forma científica, não induzida e obedecendo as regras da teoria quantitativa e qualitativa (Côrtes, 2003).

O conhecimento é um conjunto de argumentos e explicações que interpretam um conjunto de informações. Trata-se de conceitos e argumentos lógicos essencialmente abstratos que interligam e dão significado a fatos concretos (Côrtes, 2003).

O tratamento necessário a este processo demanda o estabelecimento de estruturas para armazenar dados, com processamento e recuperação em alta velocidade; estas estruturas são denominadas sistemas de informações e se utilizam de bancos de dados como elementos de armazenamento de dados e informações. A partir destes bancos serão processadas as atividades necessárias à construção do conhecimento.

Segundo Han (Han, 2002) o processo de descoberta de conhecimento, deste momento em diante intitulado KDD – *Knowledge Discovery in Database* - a partir de bancos de dados consiste de uma seqüência de sete passos, a saber: limpeza de dados, integração de dados, seleção de dados, transformação de dados, mineração de dados, avaliação de padrões e apresentação/assimilação do conhecimento. Na figura 1 apresenta-se esquematicamente este processo

Este artigo aborda a aplicação da informação espacial e a importância da qualidade do cadastro atendo-se aos dois primeiros passos, visto que os seguintes tratam da manipulação do dado em si. Para tal conceitua-se limpeza de dados como a atividade de remoção de dados inconsistentes e fora de padrões, e integração de dados como a possibilidade de integrar várias fontes de dados mantendo a consistência.

2 – O KDD e sua relação com o cadastro urbano e de clientes.

O cadastro urbano, de uma forma resumida, consiste no registro alfanumérico, em um banco de dados, dos atributos da ocupação e uso do solo; bem como dos equipamentos que constituem o patrimônio urbano. Com relação à gestão da ocupação urbana usualmente emprega-se o MUB, ou mapa urbano básico, que segundo Aguiar (Aguiar, 1996) deve ser entendido como uma camada (ou pano de fundo) sobre a qual se apresentam os diversos elementos que condicionam a ocupação do território, onde se desenvolvem as atividades humanas e ocorrem os relacionamentos de ordem social e ambiental; este “pano de fundo” se apresenta em forma de cartas (mapas).

Estas cartas devem seguir os padrões do sistema cartográfico municipal, que segundo a Norma NB 14.166 da ABNT – Associação Brasileira de Normas Técnicas – integram as cartas e plantas que constituem a base cartográfica municipal. Estas deveriam conter basicamente as informações territoriais

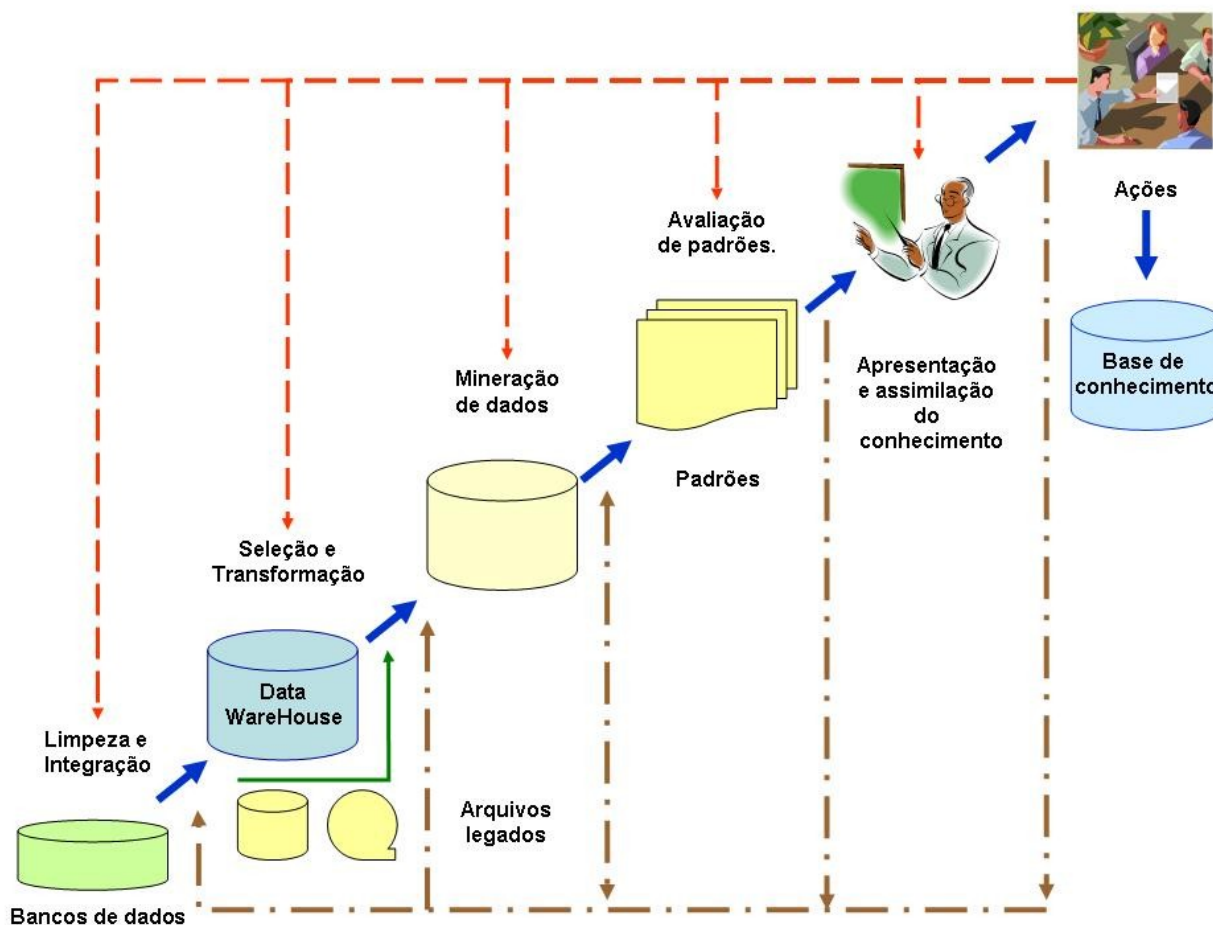


Figura 1: O processo de KDD

necessárias ao desenvolvimento de planos, anteprojetos, projetos, cadastro técnico e imobiliário fiscal, acompanhamento de obras e outras atividades de projeto que necessitem do terreno como referência.

Segundo Aguiar (Aguiar, 1996) o MUB constitui o conjunto de informações gráficas e alfanuméricas referentes à base cartográfica a ser utilizada, em geral relativa às plantas cadastrais, ou seja, os vários planos de informações (feições) que compõem o cadastro urbano e que permitem sua apresentação gráfica. Dentre estes planos estão os relacionados com a constituição de uma base de endereçamento, que será utilizado na abordagem deste artigo.

O MUB normalmente está disponível em plataformas de geoprocessamento, CAD ou ainda em meio analógico (papel) e as informações (exemplo: traçado das ruas, informações sobre os logradouros e suas codificações, endereços, unidades territoriais básicas, posicionamento das propriedades), podem ser obtidas a partir de plantas cadastrais, ou de aerolevantamentos, seguido de levantamentos de campo.

O cadastro de clientes, de forma resumida, consiste em um modelo de dados para criação e manutenção da base de dados de clientes ou consumidores de uma determinada empresa. Neste cadastro serão armazenados os dados referentes ao endereço, informações pessoais, produtos consumidos, tempo de relacionamento, entre outros.

Estes cadastros, com exceção do MUB, normalmente estão armazenados nos sistemas do tipo OLTP – *On Line Transaction Processing* – e estão disponíveis no ambiente de produção com a missão de

suportar as operações diárias da empresa. A sua manutenção normalmente demanda uma estrutura particular para tal. Em geral nas empresas de *utilites* a atividade de manutenção do cadastro dos logradouros é realizada por uma área específica, assim como o de clientes é mantido por uma interface de *contact center*.

Até meados da década de 90 a informação através de cartas era comumente utilizada para validar as operações de manutenção da base de dados de logradouros a partir de comparação entre as bases alfanuméricas e as cartas. Em verdade isto não poderia se caracterizar como um cadastro vivo, visto a desintegração entre o ambiente de manutenção dos meta-dados e o ambiente de gestão da cartografia, fosse esta digital ou analógica. Com a adoção dos sistemas de informações geográficas, principalmente na figura das ferramentas de AM-FM (*Automated Mapping and Facility Management*) e os CAD's estas bases começaram a se integrar constituindo verdadeiras ferramentas de gestão, porém sem a garantia de que permeiem toda a estrutura operacional da empresa.

Apesar de estes cadastros serem da mesma natureza, integrá-los é primordial para garantir a qualidade do cadastro geral e a replicabilidade dos dados ao longo do processo de KDD.

O modelo de dados da base de endereçamento que se está contemplando neste artigo compreende: município, distrito, localidade, bairro, código de logradouro, tipo do logradouro, título do logradouro, nome do logradouro, código de endereçamento postal (CEP), quadra e segmento de logradouro, número de edificação e o complemento das edificações.

O processo de KDD, seja este aplicado a atividades de Marketing, Vendas, Planejamento urbano ou coleta de impostos, dependendo do uso, utiliza como elemento básico para disposição espacial de clientes ou contribuintes, a informação do endereçamento cadastrada na base de clientes. A partir desta orientação é realizada a espacialização do universo de clientes ou contribuintes e realizado o conjunto de análises aplicado ao problema estudado, sejam estas do tipo espacial, de mineração e/ou avaliação de padrões.

É vasta a gama de análises e ferramentas utilizadas no processo de KDD. Em Côrtes (Côrtes, 2003) as análises são classificadas em:

- Análise descritiva – investigação dos dados que busca descrever fatos relevantes e analisar a base de dados como um todo. Divide-se em análise prévia (*outliers*, desvios e histogramas) e descobrimento (classificação, associações, *clustering*, descrição, detecção de seqüências, segmentação, sumarização e visualização, otimização, regularidades e tendências).
- Análise de prognóstico – busca inferir resultados a partir dos padrões encontrados na análise descritiva, ou seja, prognosticar o comportamento de um novo conjunto de dados. São divididas em análises de estimação, predição e classificação.

Côrtes (Côrtes, 2003) salienta que uma vez definido o resultado que se deseja alcançar são definidas as técnicas que devem ser utilizadas para a realização da análise. Por exemplo, a técnica de estimação pode ser feita utilizando-se a regressão linear ou a regressão múltipla. Várias são as ferramentas que podem ser empregadas, mas ressaltam-se as de:

- Visualização - que visa ter uma medida inicial da qualidade dos dados e de onde os padrões possam ser encontrados, assemelhando-se as análises de arranjos espaciais.
- Análise de vizinhança - que visa identificar os conjuntos de registros que estão próximos e que tem similaridade em função de alguma característica nos dados.
- Análises de *clustering* - que visam detectar a existência de diferentes grupos dentro de um determinado conjunto de dados.

Apesar de Côrtes (Côrtes, 2003) não citar a utilização das técnicas de análise espacial como ferramentas aplicadas ao processo de KDD, estas técnicas foram destacadas para deixar clara a semelhança com as técnicas empregadas nas análises espaciais por padrão de pontos e áreas, que são análises típicas de SIG's - Sistema de Informações Geográficas.

São várias as possibilidades de implementação de arquiteturas funcionais para plataformas de manutenção de cadastros urbanos. Este artigo adota como referência a forma mais comum, ou seja, o cadastro alfanumérico e a plataforma digital (SIG ou CAD) associados ao DW (*Data Warehouse*) através de ETL (Estágios de Tratamento e Limpeza) e servindo como início do processo de disseminação de informação baseada na distribuição espacial, como se pode ver na figura 2.

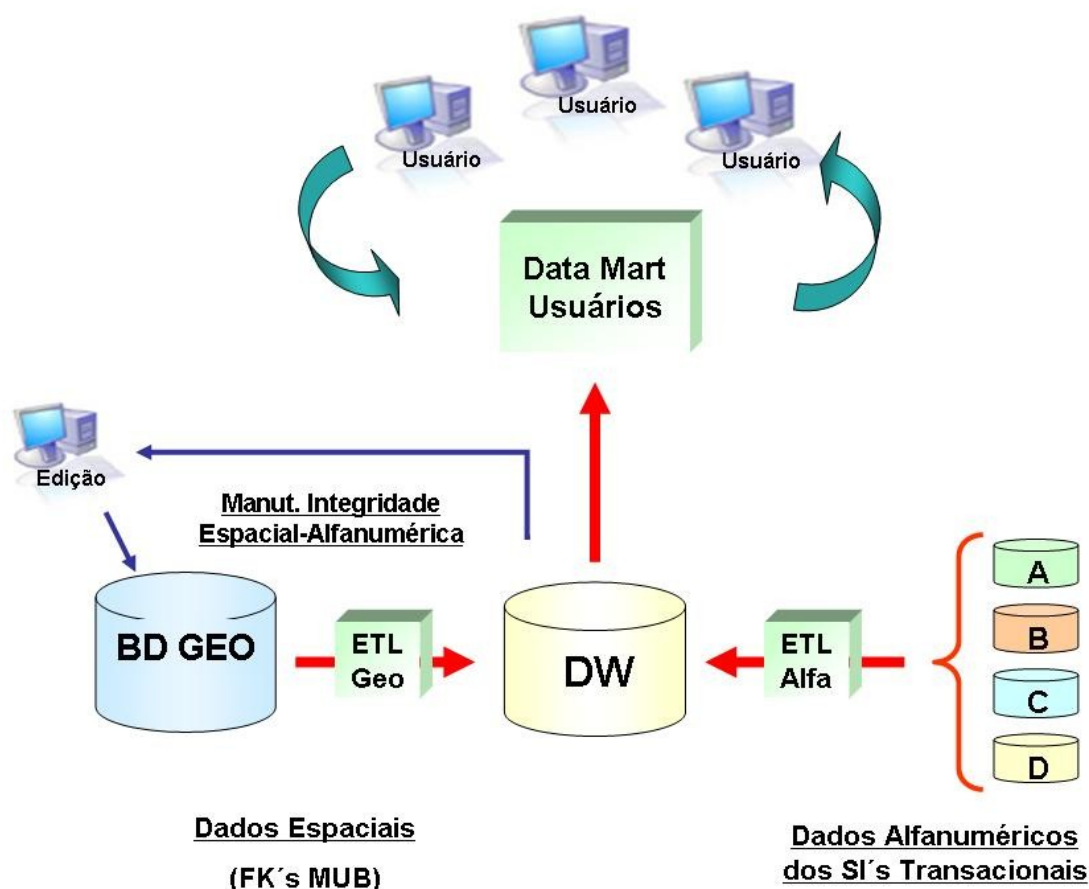


Figura 2: Arquitetura funcional mais comum.

Côrtes (Côrtes, 2003) sugere que as etapas de limpeza e integração só são necessárias quando os dados utilizados no processo de KDD estejam armazenados em vários bancos ou arquivos legados, e que se os dados forem oriundos de um DW pode-se iniciar o processo pela transformação dos mesmos.

Levando-se em consideração que o tipo de dado seja exclusivamente do tipo alfanumérico o armazenamento em um DW demanda que estejam organizados para tal. Entretanto quando o tipo de dado tem uma dimensão mais ampla, como por exemplo, os dados de cadastro e a sua correspondência no mapeamento, não necessariamente o que está armazenado em um DW está certificado.

3 - Propagação de erros

Os dados cadastrais relativos à manutenção da base de endereços e de clientes estão implantados nos sistemas transacionais de uma empresa. Desta forma o que está armazenado neste meta-dados permeia toda a estrutura da empresa.

O cadastro de um cliente demanda a pré-existência de seu endereço, desta forma torna-se necessário identificar primeiro o endereço do mesmo para permitir seu completo cadastramento. Este cadastro pode ser feito a partir de uma base certificada de endereços levantados em campo e armazenados em bancos de dados, com ou sem referência espacial; ou, sem esta certificação.

Para o caso sem certificação do endereço o cadastro do endereço de um cliente pode ser realizado a partir de um banco de dados com parte das informações, por exemplo, logradouros fornecidos por prefeituras ou outros órgãos oficiais; ou ainda, pelo fornecimento total das informações por parte do cliente com posterior confirmação. Neste caso o cadastramento compreenderia três partes:

- Informações oficiais - encontram-se disponíveis no banco de dados de endereçamento, pois são informações conhecidas (logradouros, CEP, bairros e municípios), ou são fornecidas pelos clientes.
- Informações do domicílio - são as informações de numeração e complemento do domicílio do cliente.
- Informações complementares – são as informações que somente com uma referência gráfica são passíveis de definição (quadra, lote e segmento).

O cadastro do endereço completo, neste caso, se dará quando estas três partes forem fornecidas a cadastradas.

A utilização das informações dos clientes de forma direta optando por sua posterior certificação é a maior fonte de erros em um cadastro de endereços. Neste caso o cadastro está sujeito à subjetividade por parte do cliente, sendo que em muitos destes casos a informação é omitida ou alterada por uma simples questão de orgulho do cliente.

O erro introduzido pelo cliente se potencializa quando culmina com a criação de novas referências locais que não as das fontes oficiais. Quando um cliente omite ou adultera a informação de um bairro ou logradouro induz ao mantenedor do cadastro que uma nova referência passou a existir. Esta uma vez criada será replicada para todos os sistemas legados e desta forma tratada e analisada. Por exemplo, se um loteamento é novo e / ou ainda não está cadastrado nos órgãos oficiais não existe uma numeração oficial e sim a identificação dos complementos por quadra e lote. Neste caso o cliente possui como identificação de endereço o número da quadra e do lote. A informação será atribuída pelo cliente e um agente de cadastro não tem como validar a informação a não ser que possua o projeto do loteamento ou com uma visita ao campo.

Isto significa uma encruzilhada para as organizações, pois construir uma base de endereços e manter a mesma atualizada demanda tempo e altos investimentos, bem como a sua não existência demanda a criação de *back offices* para tratamento de inconsistências de cadastro. A simples existência de uma plataforma de SIG não é a garantia de certificação do cadastro, pois esta apresenta dificuldades para manutenção de erros de integração entre cadastros alfanuméricos e dados gráficos quando estes não são feitos a partir de bases e cadastros certificados.

Davis (Davis, 2002) cita ainda que o intenso ritmo de mudanças característico dos ambientes

urbanos mais densamente ocupados leva a uma grande dificuldade de manutenção dos dados geográficos, sejam eles básicos (componentes do MUB) ou temáticos. Mesmo quando se tem ferramentas metodológicas para organizar a atualização. A grande variedade de classes de informações acaba por exigir a intervenção de especialistas em cada área, tornando a centralização da tarefa algo bastante difícil.

Um exemplo é a desorganização dos processos de fornecimento de um endereço para um novo domicílio, em grandes cidades. No Rio de Janeiro, por exemplo, a numeração é fornecida pela empresa de água ou energia elétrica quando se pede a instalação dos referidos pontos de fornecimento no canteiro de obra. Este número só é fornecido a prefeitura quando da entrada dos documentos para solicitação a licença de obra ou construção. Em algumas cidades é comum o cliente, quando muda de domicílio, levar o número do domicílio anterior. Nestes casos se torna rotineiro encontrar a mesma numeração repetida até quatro vezes no mesmo logradouro.

Este artigo não tem como objetivo descrever os processos de cadastramento de clientes e endereços. Estas situações foram apresentadas somente com o intuito de traçar um cenário mais amplo da totalidade dos possíveis erros e de sua propagação.

Na verdade a questão central, ou melhor, o divisor de águas deste assunto é o trabalho de campo. Uma vez que este exista, seja a ação de qualquer forma, pré-ativa, ativa ou reativa, permitirá a certificação das informações e a correção dos valores dos atributos.

3.1 – Tipos de erros

Vários são os possíveis erros em um cadastro. Por questão de organização este artigo irá se aprofundar nos erros de conteúdo da extensão das tabelas de uma das bases de dados e a sua correspondência quando as bases não se encontram integradas. Portanto dividiram-se estes erros em:

- Natureza – erros em dados alfanuméricos ou elementos gráficos.
- Integração - erros oriundos da integração dos dados alfanuméricos e dos dados gráficos.
- Inferência - de uma forma mais pontual são os erros introduzidos pelo homem, de forma manual ou automática através da utilização de sistemas de limpeza de dados, como por exemplos as ferramentas de *Data Care*, que se vale de fontes externas para promover uma padronização de dados na base.

É importante destacar que os erros de inferência são decorrentes de uma falha na manutenção das extensões dos meta-dados alfanuméricos, seja na base cadastral de endereços, seja no meta-dados das cartografias digitais utilizadas nos SIG's. Esta falha pode ser a falta do preenchimento do valor do atributo ou até mesmo a valoração errada para o atributo. Estes erros serão abordados no item 4, que trata da limpeza dos dados.

3.1.1 Erros alfanuméricos

São os que dizem respeito ao preenchimento da extensão dos meta-dados cadastrais. De certo estes tipos de dados, suas extensões e domínios deveriam estar bem definidos visto que se trata de modelos de dados implantados em aplicações do tipo OLTP, ou seja, não estariam aqui relacionados os erros provenientes de falta de padronização de campos. Porém isto não é verdade na prática, pois os campos do tipo livre permitem a mais variada sorte de combinações e associações em seu preenchimento.

Desta forma poderíamos citar como exemplos de erros alfanuméricos:

- Bairros – criação de bairros não oficiais em função de nomes populares ou sub-bairros.

- Logradouro – duplicação de logradouros no mesmo bairro, duplicação do mesmo logradouro em bairros diferentes em um mesmo município, criação de cadastros de logradouros para bairros não oficiais.
- Tipo de logradouro – logradouros que atravessam bairros com identificação de tipo diferenciada por trecho, caracterização errônea do tipo do logradouro.
- Título do Logradouro – despadronização do preenchimento do campo. Ex: Gal., General, GL.
- Nome – campo do tipo string longo de preenchimento livre.
- Número – criação de referências de endereçamento inválidas. Variadas formas de numeração.
- Complemento – despadronização do preenchimento do campo complemento. Ex: Apto., Ap., Apm.

Na prática, em função da desestruturação urbana e do crescimento desordenado, o campo complemento é o que apresenta a maior incidência de erros, pois na maioria das vezes a quantidade de campos destinados a criar a referência de endereçamento não é suficiente para realizar a tarefa. Nestes casos é comum a necessidade de se ter mais de três *strings* alfanuméricas para realizar esta tarefa.

Ex. Rua “X” nº 230, FDS 01, BL 01, AP 201.

3.1.2 Erros gráficos e de posição

De uma forma mais ampla trata da existência das camadas do modelo de dados, como entidades gráficas, e o seu real posicionamento, bem como a padronização gráfica de cada um dos elementos (simbologia, linhas, etc.).

De uma forma geral os erros constituídos neste tipo de carta referem-se à questão da existência e do posicionamento. Com relação à existência o erro está relacionado com a informação, ou seu plano, constarem do MUB; e com relação ao posicionamento está relacionado ao deslocamento máximo entre a linha de centro do elemento real e da sua contrapartida gráfica (desenho); ou seja, a feição criada exceder o intervalo de tolerância permitido.

Cabe ressaltar que os mapas digitais devem seguir as especificações técnicas que regem as questões de topologia e tolerância. Este artigo não irá tratar deste assunto e dos respectivos erros associados. O foco principal entre a entidade cadastral e a sua correspondência gráfica é a existência, devida ou indevida, da entidade.

De certo, a adoção de um SIG, vem contribuir na questão da manutenção de um cadastro de endereços único e consistido, sendo inclusive possível constituírem regras de negócios que venham a impedir a inserção e propagação de erros.

3.1.3 – Erros de Integração

A adoção de um SIG como ferramenta de manutenção de cadastros demanda a atividade de integração da base de dados do cadastro de endereços, do cadastro de clientes e o mapeamento para sua utilização. Esta manutenção se dá no sentido amplo da palavra, ou seja, editar as bases alfanuméricas e de mapeamento.

A implantação de uma solução com estas características demanda a aplicação, e provavelmente reformulação dos processos de manutenção das bases cadastrais, visto que serão adotadas novas técnicas de manutenção que utilizarão novas regras de negócio. Isto inclusive deve acontecer em estágios,

para que se permitam adotar, por parte dos artífices, as melhores práticas de trabalho, e consequentemente se estabelecerem os melhores processos operacionais.

Por exemplo, se atualmente a inserção de um logradouro em uma base alfanumérica é feita diretamente sobre o sistema legado isto constitui uma operação única. Todas as tabelas e campos previstos serão preenchidos para esta atividade e um novo registro de logradouro é criado.

Quando se adota uma base de mapeamentos, que em um primeiro momento é departamental, este processo de criação de um logradouro demandará uma edição no mapeamento. Caso este mapeamento possua características de uso para SIG, além da edição gráfica dos segmentos referentes ao novo logradouro, também demandará a atualização das tabelas com os mesmos campos do sistema legado. Isto significa uma segunda operação similar à realizada no cadastro original.

Caso a solução seja corporativa toda edição poderá ser feita diretamente a partir do SIG com o devido *up-load* no sistema transacional, ou seja, a manutenção do sistema legado se dará a partir de uma ferramenta SIG e deste ponto em diante as bases encontram-se no melhor estado de sincronismo.

Isto não é um processo simples e demanda uma série de outras decisões que envolvem entre outras o desenho da arquitetura funcional, padronização de entrada de dados, tratamento de inconsistências e a forma de replicação dos dados através da estrutura de disseminação adotada.

Neste processo de integração pode-se elencar os seguintes fatores como críticos para o sucesso:

- Geocodificação das feições do MUB – verificar a existência e a padronização dos códigos identificadores dos elementos gráficos existentes.
- Atributos de Endereçamento: checar a validade da nomenclatura de tipo, título, nome do logradouro e complementos, bem como a relação entre os códigos identificadores da base de dados alfanumérica e as tabelas das feições da base do SIG, ou seja, a correspondência entre a associação do banco de dados e o elemento gráfico.
- Geocodificação de Clientes: é a associação entre os códigos de identificação da base de clientes com os respectivos pontos representativos dos domicílios MUB. Neste caso dever-se-á tratar as inconsistências referentes aos números indicados pelos clientes e o levantamento de campo.

Com relação à associação entre a base alfanumérica e os mapas é importante ressaltar que nesta integração deverão ser tratados os erros de duplicação de entidades, seja na base alfanumérica seja no mapeamento, bem como os códigos e entidades que ficaram sem contrapartida entre as bases. Os elementos indevidos, como bairros e logradouros inexistentes deverão ser excluídos das fontes de dados.

4 – A limpeza de dados

Han (Han, 2001) salienta que as rotinas de limpeza de dados buscam basicamente preencher os valores ausentes (*missing values*), padronizar dados enquanto identificam valores fora de padrões (*outliers*) e corrigir inconsistências nos dados.

Pela natureza da rotina pode-se concluir que este tipo de procedimento pode inserir, ou alterar o valor dos dados quando são encontrados campos nulos que deveriam estar preenchidos ou quando a referência de comparação não contém o valor esperado.

Para o caso de valores ausentes, Han (Han, 2001) caracteriza este procedimento quando existirem

registros ausentes em diversas tuplas, os quais podem ser importantes para o KDD. O mesmo autor sugere alguns métodos para atribuir valores a estas tuplas:

- Ignorar o registro – não preencher.
- Imputar o valor manualmente – incrementar a base com um valor a ser definido para cada tupla nula.
- Preencher com uma constante global – atualizar todos os valores ausentes com um único valor constante.
- Utilizar um atributo médio – quando o atributo é do tipo numérico aplicar um valor médio para o referido atributo.
- Utilizar um atributo médio pertencente à mesma classe à qual a tupla pertença - quando o atributo é do tipo numérico aplicar um valor médio para o referido atributo, e este pode ser estimado em função de uma classificação.
- Utilizar o valor de maior probabilidade para preencher os valores ausentes – determinar um valor através da aplicação de uma técnica de regressão, inferência ou mesmo a partir de uma árvore de decisão.

Avaliando a figura 1, pode-se verificar que tal rotina está relacionada à etapa anterior a carga no DW; neste ponto cabem algumas considerações com relação à informação espacial ou a sua referência cadastral, quando se trata do cadastro urbano.

Como citado, o processo de KDD utiliza a informação do endereçamento como elemento básico para disposição espacial de clientes ou contribuintes, dependendo do uso. Desta forma o processo de limpeza de dados sobre o meta-dados que trata da base de endereços pode trazer problemas.

Valores nulos preenchidos por um dos métodos de inferência citados vão alterar a disposição da referência locacional alterando desta forma a distribuição dos eventos e por consequência alterando as estatísticas básicas do evento. Por exemplo, uma contagem: um conjunto de clientes que habitam no logradouro “x” qualquer que está com a referência de bairro vazia. Ao adotar-se um valor conhecido ou inferido para este campo estar-se-á possivelmente alterando a informação de clientes totais por bairros.

Cabe lembrar que o processo proposto é basicamente alfanumérico, desta forma o cadastro se apresenta ao processo como um modelo de dados, um glossário de tabelas. Na existência de um SIG de certo estes problemas irão ser reduzidos, mas não serão eliminados, por que neste caso apesar de existir uma base alfanumérica consistida com uma referência espacial, o MUB, ainda podem existir os erros citados no item 3.1.2.

Outro ponto importante é a aplicação de ferramentas de inferência básica, como por exemplo, as ferramentas de ajuste de dados, os *Data Care's*. Estas ferramentas utilizam como processo a comparação léxica entre a base do transacional e uma fonte externa, ou seja, comparando *strings* de caracteres em campos conhecidos. A primeira pergunta sobre este processo é: a fonte externa é atual ou certificada? A ausência de dados muita das vezes leva os analistas a procurarem a melhor forma possível de suprir esta ausência, o que não quer dizer que a melhor forma.

Realmente encontrando-se a base a ser comparada com um grau de desatualização muito grande o processo de comparação pode trazer algum tipo de benefício, porém é necessário após este trabalho certificar estes ajustes a partir de uma massa de testes.

O segundo ponto com relação a este procedimento é o fato que se cria uma segunda referência de endereçamentos. Por exemplo, a fonte oficial, implantada normalmente nos sistemas OLTP's da

organização constitui a fonte que alimenta os sistemas legados. Desta forma também alimenta o DW. Quando se realiza um processo de limpeza de dados com a inferência sobre os campos, está sendo gerada uma segunda referência a partir deste ponto do processo. Estas modificações valerão para os processos subsequentes ao transacional, ou seja, pode não representar a realidade.

Sabendo que o DW é a referência de armazenamento para os dados oriundos dos sistemas transacionais, uma vez que tenha sido estabelecida uma segunda referência a partir deste ponto compromete-se o resultado alcançado vis a vis a realidade. Isto se refletirá de forma mais contundente sobre os processos de alocação da base de clientes (cadastro de clientes) a base de endereços.

Um sistema de informações deve prover a aquisição, armazenamento e disponibilização dos dados, e estas são atividades perenes. Desta forma uma vez que se estabelece uma segunda verdade cria-se um desvio com relação à manutenção desta, ou seja, em um tempo futuro a fonte que foi utilizada para comparar deverá estar atualizada sobre as mesmas regras de negócios antes estabelecidas, caso contrário potencializa-se a possibilidade de criar uma terceira referência e assim por diante.

5 – Conclusões

Os processos de KDD, suas análises e ferramentas, baseiam-se principalmente em análises de natureza alfanumérica, ou seja, análises baseadas em comparações e / ou aglutinações de campos em formato de *string*. Como apresentado no trabalho estes campos podem ser fortemente influenciados pelos processos de limpeza, visto que interferem sobre o valor do atributo.

A posição espacial de um cliente pode ser dada pela sua posição geodésica ou por uma referência que permita encontrá-lo em um mapa, por exemplo, um logradouro e um número de domicílio. Quando é utilizada esta segunda forma, passa-se a trabalhar com conceitos de análises sobre banco de dados, ou seja, os resultados serão obtidos a partir de consultas aos bancos de dados que utilizaram o valor do atributo constante nas tabelas como elemento de análise.

Pode-se concluir que a inferência sobre os valores armazenados na base de endereço, quando da sua extração dos sistemas transacionais para a carga no DW, pode comprometer a qualidade do dado distribuído, pois esta se alterando a referência de localização. Esta alteração tem um impacto maior quando a organização urbana de uma determinada porção do território tem menos qualidade, como por exemplo, nas periferias.

Pode-se concluir também que a informação do endereço é de primordial importância para o processo de KDD, visto que algumas de suas ferramentas, como por exemplo, a análise de *clustering*, visa exatamente descobrir padrões de distribuição. Neste sentido o resultado deste tipo de análise fica comprometido por ser uma comparação alfanumérica não sendo permitido a aplicação de técnicas mais apuradas de decisão, como a interpolação e as análises baseadas nas distâncias entre os eventos.

Apesar de não serem incluídas claramente como ferramentas de KDD pode-se concluir que em função do tipo de análises que são realizadas nestes processos a utilização de ferramentas SIG apresenta vantagens técnicas quando comparados aos métodos ora empregados, visto que trabalham com uma informação real de posicionamento.

Outra vantagem do emprego dos SIG's no processo de KDD é o fato de que os erros do cadastro podem ser detectados e tratados ao longo do tempo fazendo com que exista uma melhora contínua da qualidade dos dados que são disponibilizados para o ambiente de análise. Esta vantagem é proveniente do

fato que a constituição do SIG demanda uma base de mapeamento atualizada, o que permitirá levantar as inconsistências cadastrais a partir das atividades de campo.

Na verdade a constituição do mapeamento e sua integração aos cadastros existentes, automaticamente forçará o ajuste das inconsistências para que possa haver integração entre o cadastro alfanumérico e o mapeamento.

Como análise final pode-se afirmar que a aplicação de SIG's como ferramentas de certificação da base de endereços melhora a qualidade das análises finais dos processos de KDD, visto que quando os dados forem armazenados no DW não mais demandaram atividades de limpeza e tratamento da referência de localização, bem como permitirá integridade de qualidade entre a base de clientes e a base de endereços.

6 – Bibliografia.

[1] **Côrtes, S.; Azevedo, H.** *Data Mining: Conceitos, Técnicas, Ferramentas e Aplicações*. CCE – Coordenação Central de Extensão – PUC RJ, 5ª Versão, Rio de Janeiro, RJ, 2003.

[2] **Han, J.; Kamber, M.** *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, Inc, 2001.

[3] **Aguiar, C. D.; Medeiros, C. B.** Construção de um Modelo básico unificado a partir de sistemas Stand-Alone. DCC - UNICAMP - In. Anais Gis Brasil 1996.

[4] **Davis Jr. C.A.; Oliveira, P. A.** *SIG Interoperável e Distribuído para Administrações Municipais de Grande Porte*. Revista Informática Pública, nº. 2, ano 4. 2002.