Descoberta de conhecimento em bases de dados públicas: uma proposta de estruturação metodológica*

Jair Sampaio Soares Junior**
Rogério Hermida Quintella***

PALAVRAS-CHAVE: descoberta de conhecimento; bases de dados; mineração de dados; sistema de apoio à decisão; gestão do conhecimento; pobreza.

KEY WORDS: knowledge discovery; databases; data mining; decision support system; knowledge management; povertry.

O mundo contemporâneo assiste ao crescimento acentuado de dois fenômenos que motivam este artigo. O primeiro deles é a difusão das tecnologias digitais e o segundo, o crescimento da parcela de sua população que vive em condições de pobreza. A humanidade gera e armazena dados e informações em uma velocidade até recentemente inimaginável. Este artigo analisa a transformação de dados públicos em conhecimento de valor social com o uso da descoberta de conhecimento em bases de dados (DCBD ou *knowledge discovery in databases* — KDD) com dois objetivos: criar uma proposta para utilização da DCBD em bases de dados públicas

* Artigo recebido em set. e aceito em out. 2005.
** Doutorando em administração na NPGA/UFBA. Endereço: NPGA/UFBA — Avenida Reitor Miguel Calmon, s/n — Vale do Canela — CEP 40110-110, Salvador, BA, Brasil. E-mail: jairsoaresjr@ yahoo.com.br.
*** Professor titular na NPGA/UFBA Endereço: NPGA/UFBA — Avenida Reitor Miguel Calmon, s/n — Vale do Canela — CEP 40110-110, Salvador, BA, Brasil. E-mail: npga@ufba.br.

e demonstrar que o uso dessa metodologia pode gerar conhecimento útil a políticas de combate à pobreza. A partir de uma reflexão teórico-metodológica foi elaborado um modelo completo de KDD, aplicado na mineração de dados domiciliares coletados pelo IBGE no censo de 2000. Neste processo, o fenômeno da pobreza foi tratado com as abordagens heurística e estatística, resultando em uma representação multidimensional baseada nas características dos domicílios e de seus moradores. Essa abordagem permitiu a criação de uma tipologia da pobreza para a cidade de Salvador, sendo os respectivos tipos georreferenciados em seguida.

Knowledge discovery in public databases: a methodological structure proposal

The contemporary world experiences a considerable growth of two phenomena that motivate this article. The first is the diffusion of digital technologies and the second is the growth of the share of its population that lives in poverty. Modern society generates and stores data and information in a very large scale. This article analyzes the transformation of public data in socialy valuable knowledge through the use of KDD (knowledge discovery in databases) with two main objectives: to propose a model for use of KDD in public databases and to demonstrate that the use of this methodology can generate useful knowledge for policies related to poverty relief. A complete model of KDD was elaborated and applied in the mining of Salvador's domiciliary data collected in the Brazilian census for the year 2000. In this process, the phenomenon of poverty was treated through a heuristical and statistical approach, resulting in a multidimensional representation based upon the characteristics of the domiciles and their inhabitants. This process allowed the creation of a tipology of poverty for the city of Salvador.

1. Introdução

No final do século passado, a tecnologia da informação (TI) na esfera pública deixou de ter um papel restrito ao suporte administrativo, passando a participar, também, em aplicações estratégicas nas tomadas de decisão, auxiliando, por exemplo, na implementação e avaliação de políticas governamentais. O processo de globalização, a internet e, no Brasil, a consolidação da democracia tornaram os cidadãos mais exigentes, ao mesmo tempo o mercado tornou-se mais competitivo e o cidadão passou a demandar mais do poder público em defesa de seus direitos. O aumento da procura por informações e a necessidade legal de maior transparência nas ações do gestor público culminaram em uma crescente disponibilização de informações por parte dos principais órgãos de governo na esfera federal, levando gradativamente as unidades da Federação a também estruturarem e disponibilizarem mais informações à sociedade.

Entre os movimentos recentes da tecnologia da informação na esfera pública, está o desenvolvimento de sistemas que permitem análises e tomada de decisões a partir de bases de dados disponibilizadas na internet.

Assim, este artigo pretende:

- propor, a partir da análise do referencial teórico que se segue, uma padronização de procedimentos que, em seu conjunto, configure um modelo atual e simples para descoberta de conhecimento em bases de dados públicas (DCBDp ou *knowledge discovery in public databases* — KDDp);
- elaborar e georreferenciar uma tipologia de pobreza para a cidade de Salvador.

2. Referencial teórico

Em geral, a gestão do conhecimento pode ser definida como o conjunto de processos para identificar o conhecimento que está presente nas pessoas e proporcionar condições adequadas para sua transferência, utilização e criação (Liebowitz e Beckman, 1998; Beckman, 1999). Já para Davenport e Prusak (1998), a gestão do conhecimento é o conjunto de atividades relacionadas com a geração, codificação e transferência do conhecimento.

A discussão sobre o conhecimento, apesar de sua aparente modernidade, é na realidade milenar. Ainda que não se possa traçar um paralelo direto, há uma aparente relação da visão do primeiro grupo de autores com a linha do racionalismo de Platão, enquanto a conceituação de Davenport e Prusak encontraria maior respaldo no empirismo de Aristóteles.

Davenport e Prusak (1998) afirmam que a gestão do conhecimento (GC) deve ter os seguintes objetivos: criar um repositório de conhecimento constituído por conhecimento externo e conhecimento interno estruturado; melhorar o acesso ao conhecimento; desenvolver um ambiente e uma cultura organizacional propícios à criação, à transferência e ao uso do conhecimento e tratar o conhecimento como um recurso mensurável.

A literatura apresenta diversas outras definições sobre gestão do conhecimento. Claramente, pode-se perceber, na atualidade, a existência de duas correntes principais: a do suporte tecnológico e a do comportamento. Na corrente tecnológica parece haver um predomínio de autores com formação na área de tecnologia da informação. Eles enfocam mais os conceitos de armazenamento, reaproveitamento e descoberta do conhecimento em detrimento de uma abordagem mais comportamental relacionada ao elemento humano adotada no segundo grupo. Assim, parece, novamente, ser possível perceber maiores relações do primeiro grupo — o da TI — com o empirismo (de Aris-

tóteles na antigüidade e Davenport e Prusak na atualidade), enquanto, por outro lado, a corrente do comportamento encontraria maior suporte no racionalismo de Platão (na antigüidade) e em autores contemporâneos (Liebowitz e Beckman, 1998; Beckman, 1999).

Nonaka e Takeuchi (1997), talvez os mais importantes autores da GC na atualidade, classificam o conhecimento humano em dois tipos: o conhecimento explícito, que pode ser articulado na linguagem formal, inclusive em afirmações gramaticais, expressões matemáticas, especificações e manuais, entre outros, e o conhecimento tácito, mais difícil de ser expresso na linguagem formal. Esta segunda corrente, identificada por Nonaka e Takeuchi, claramente melhor se coaduna com a corrente racionalista do pensamento grego e, dentro desta, na abordagem do comportamento com os trabalhos, por exemplo, de Liebowitz e Beckman.

Com base na discussão apresentada, pode-se classificar o presente artigo na linha de pensamento do empirismo de Aristóteles e da tecnologia de Davenport e Prusak. Em um esforço para tornar ainda mais clara a inserção da presente pesquisa no vasto campo da GC, considerou-se o trabalho de O'Dell e Grayson Jr. (2000). Estes autores dividem a aplicação das ferramentas de tecnologia na gestão do conhecimento em duas subclasses: transmissão e troca de conhecimento e análise de dados e suporte ao desempenho.

Considerando-se as duas subclasses de O'Dell e Grayson Jr., este artigo tem foco no segundo grupo, mais especificamente em exploração de dados, suporte à decisão e análise de dados, que é tido por esses autores "o território inexplorado da gestão do conhecimento" (O'Dell e Grayson Jr., 2000:124).

Sistemas de informação

De acordo com Laudon e Laudon (1994), o estudo de sistemas de informação (SI) constitui um campo multidisciplinar. Este novo campo lida com questões e reflexões derivadas de disciplinas como sociologia, economia e psicologia, no comportamento, e disciplinas como ciências da computação, pesquisa operacional e ciências da administração, nas abordagens técnicas.

O conceito de sistemas de informação (SI) tem evoluído substancialmente, fugindo de uma visão puramente técnica para uma visão social, mesmo que a palavra social ainda tenha um sentido vago na ciência da computação (Ivanov, 1998). É importante lembrar, também, que esta evolução e as orientações de pesquisa se diferenciam significativamente de um país para outro, e de uma escola de pensamento para outra, não havendo, portanto, um paradigma universal de pesquisa em informática social.

Múltiplas perspectivas contribuem para a formação do conceito de informática social como área de estudo dos diferentes aspectos sociais das atividades computadorizadas nas organizações. Para Friedman e Kahan Jr. (1999), as preocupações éticas e sociais devem ser partes integrantes do desenvolvimento de sistemas de computadores. Portanto, se a tecnologia da informação tem um grande potencial para alterar nossas vidas, o desenvolvimento da informática social é uma oportunidade que não podemos simplesmente ignorar (Schuler, 1994).

Sistema de apoio à decisão

O avanço tecnológico propiciou a redução dos custos e a difusão dos computadores. Conseqüentemente, houve um aumento da capacidade de coleta e armazenamento de dados não ocorrendo um aumento simultâneo e equivalente na capacidade de utilizar esses dados. Em meio a essa dinâmica, cresceu a demanda por diferentes sistemas de informação para apoiar a tomada de decisões, surgindo assim os chamados sistemas de suporte à decisão (SSDs), aqui denominados sistemas de apoio à decisão (SADs).

As definições de SSD e SAD podem ser reunidas em dois extremos conceituais: o de escopo mais amplo, onde os SADs "[...] são aqueles que contribuem de alguma forma para a tomada de decisão", e o de interpretação mais restrita, pelo qual, "SADs são sistemas baseados em computador, interativos, que auxiliam gerentes na utilização de dados através de modelos para resolver problemas não-estruturados" (Sprague e Watson, 1991:78).

A definição de SAD adotada neste artigo é: "sistemas que utilizam TI para tratar dados ou informações pouco estruturadas, de forma sistemática, visando transformá-las em conhecimento ou informações mais estruturadas destinadas a apoiar a tomada de decisões".

Dhar e Stein (1997 segundo Laudon e Laudon, 1994) reconhecem dois tipos básicos de SAD, o primeiro, chamado de SAD *guiado por modelo*, caracteriza os sistemas desenvolvidos de maneira isolada dos principais sistemas de informação da organização. Esses sistemas são baseados numa forte teoria ou modelo que se combina com uma boa interface, facilitando a execução pelo decisor por simulações e outros tipos de análises.

Já o segundo tipo de SAD, o *guiado por dados*, é mais recente e voltado para a extração de informações úteis previamente desconhecidas independentemente da existência de um modelo prévio. Neste grupo, podem ser encontradas ferramentas como Olap e *data mining.*

Características dos sistemas de apoio à decisão

Apesar de haver grande concordância entre as definições adotadas neste artigo e os conceitos mais amplamente utilizados na literatura de sistemas de informação, propõe-se aqui uma abordagem distinta daquela feita por alguns importantes autores. Por exemplo, para Damiani (1998), os sistemas de informação podem ser divididos em três categorias básicas: de apoio à gestão estratégica, de apoio à gestão tática e de apoio à decisão de nível operacional. Segundo este autor, a aplicação de SAD ocorre apenas no nível tático. Diferentemente do que preconiza Damiani e de acordo com a definição aqui adotada, entende-se que também o nível estratégico, e não apenas o tático, requer sistemas de apoio à decisão. Outra distinção entre a abordagem de Damiani e a aqui adotada é a clara dicotomia entre os três níveis de gestão assumida pelo referido autor. Tal dicotomia parece, hoje, um tanto quanto extemporânea, já que no paradigma da administração estratégica (no qual se insere o presente artigo) se pressupõe que mesmo a gestão operacional deve ser vinculada e sincrônica às grandes estratégias organizacionais.

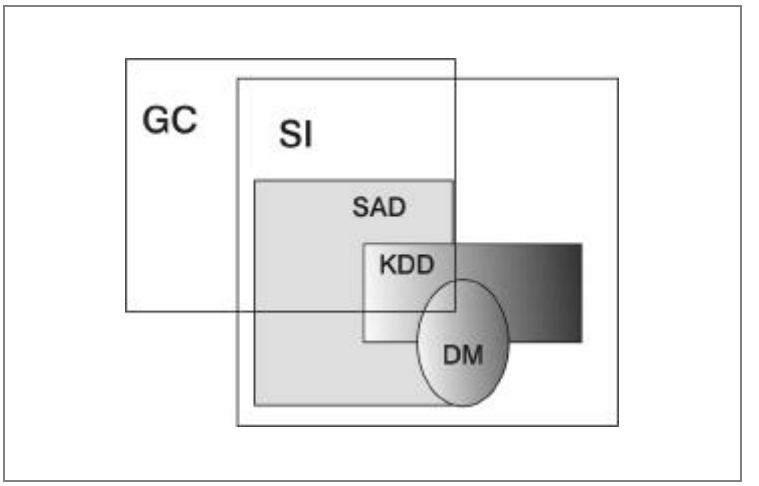
Com base nas definições acima, optou-se, neste artigo, por uma perspectiva sistêmica, a qual engloba todo o processo de descoberta de conhecimento útil em bases de dados. Visando oferecer uma melhor percepção dos principais conceitos relacionados a este artigo, procurou-se identificar a relação existente entre: gestão do conhecimento; sistema de informação; sistema de apoio à decisão; descoberta de conhecimento em base de dados e *data mining*. Essas relações são ilustradas na figura 1.

Conforme foi definido anteriormente, a corrente do suporte tecnológico na gestão do conhecimento pode utilizar os sistemas de informação que, por seu turno, possuem um tipo mais específico que é o SAD. Esse tipo de sistema incorpora ferramentas analíticas avançadas, possibilitando simulações e elaboração de cenários. Assim, os SADs envolvem, mas não limitam, o processo de KDD, metodologia que, por sua vez, utiliza o *data mining* (DM) como uma de suas ferramentas ou técnicas. Observa-se, por fim, que o *data mining* pode ser utilizado em processos outros que não o KDD, não estando portanto por ele limitado, conforme mostra a figura 1.

Figura 1

Gestão do conhecimento, sistemas de informação,
sistemas de apoio à decisão, *knowledge discovery*

in databases, *data mining* e suas interfaces



Fonte: Adaptado de Quintella e Soares Jr. (2003:10).

3. Descoberta de conhecimento em bancos de dados

O crescimento rápido do volume e da dimensão das bases de dados criou a necessidade e a oportunidade de se extrair sistematicamente o conhecimento nelas contido e de se produzir novos conhecimentos. Neste contexto, surge, no final da década de 1980, um novo ramo da computação, a descoberta de conhecimento em bases de dados (DCBD), com o objetivo principal de encontrar uma maneira estruturada de, com o uso da TI, explorar essas bases de dados e reconhecer os padrões existentes pela modelagem de fenômenos do mundo real (Fayyad et al., 1996).

O KDD engloba, portanto, as etapas que produzem conhecimentos a partir de dados relacionados e sua principal característica é a extração não-trivial de informações e conhecimentos implicitamente contidos em uma base de dados. Essas informações e conhecimentos são, usualmente, de difícil detecção por métodos tradicionais de análise, sendo também típica e potencialmente úteis na tomada de decisões (Frawley, Piatetsky-Schapiro e Matheus, 1992; Fayyad et al., 1996). Assim, enquanto os métodos tradicionais são capazes de tratar apenas as informações explícitas, o KDD é capaz de detectar informações armazenadas nas bases de dados, transformando-as em conhecimento.

O processo de KDD é iterativo e, embora apresente uma definição semelhante também ao DM, deve ser composto de uma série de etapas seqüenciais, podendo haver retorno a etapas anteriores, isto é, às descobertas realizadas (ou à falta delas). Esse

processo conduz, eventualmente, a novas hipóteses e descobertas. Neste caso, o usuário pode decidir pela retomada dos processos de DM ou uma nova seleção de atributos, por exemplo, para validar as hipóteses que tenham surgido ao longo do processo.

Funcionamento do KDD

O processo do KDD, diferentemente do *data mining*, exige três atores de distintas habilidades: o usuário, o especialista do domínio e o analista de dados. O usuário é o demandante do trabalho, que irá potencialmente desfrutar os resultados obtidos. O especialista do domínio é quem conhece o tema que será estudado, normalmente um pesquisador ou profissional com larga experiência. O analista de dados, por sua vez, é quem deve executar o processo de verificação (tese) ou anulação das hipóteses (antítese) criadas pelo especialista do domínio (eventualmente em conjunto com o usuário), que, por sua vez, irá entrar no ciclo de reformular as hipóteses (síntese) para que sejam novamente testadas, seguindo uma "trajetória em espiral rumo à finalização do processo". Observa-se que na abordagem de Inmon, Terdeman e Imhoff (2001) não há distinção formal entre usuário e especialista do domínio.

Etapas da descoberta de conhecimento em bancos de dados

O KDD é composto por um conjunto de etapas que, em geral, podem ser reunidas em três fases: preparação, análise e interpretação (Adriaans e Zantige, 1996; Brachman e Anand, 1996; Fayyad et al., 1996; Han e Kamber, 2000). Todas essas fases são críticas, sendo usualmente a fase de análise a mais complexa. Ela compreende, entre outras, a etapa de mineração de dados, que tem como objetivo encontrar padrões nos dados armazenados. Esta etapa é freqüentemente confundida na literatura com o próprio processo de KDD (Han e Kamber, 2000).

O primeiro registro descritivo dos processos de KDD data de 1996 no artigo intitulado *"The KDD process for extracting useful knowledge from volumes of data"*, de autoria dos pesquisadores Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth do Massachusetts Institute of Technology (MIT). Ele demonstra a preocupação dos autores em sistematizar as etapas do processo KDD, já que, segundo eles: "A maioria dos trabalhos anteriores sobre o tema dava ênfase à etapa de *data mining*. No entanto, os outros passos são igualmente, se não mais, importantes para o sucesso da aplicação de KDD na prática".

Em um outro importante trabalho sobre o tema, Han e Kamber (2000) alertam para a importância de um *data warehouse* previamente concebido. Esses autores apresentam o processo KDD dividindo-o em sete etapas: limpeza dos dados;

integração; seleção dos dados; transformação; *data mining*; avaliação de padrões e disseminação do conhecimento.

Adriaans e Zantige (1996) se diferenciam dos demais autores apresentados por evidenciarem a necessidade de um dinamismo para o processo, pois, segundo eles, em qualquer etapa os dados podem ser incluídos, alterados ou descartados. Por outro lado, somente depois de selecionados os dados é que os autores propõem o início da etapa de codificação, com o objetivo de formatá-los e recodificá-los de forma a atender às exigências dos algoritmos da etapa seguinte, o *data mining*. Para eles é nesta etapa, com uso intensivo de recursos computacionais, que efetivamente se extrai o conhecimento. Estes autores afirmam que 80% do conhecimento são extraídos com uma "análise menos trivial" por consultas *ad hoc* com o uso de ferramentas SQL, só então se devendo utilizar técnicas mais avançadas.

Entre as abordagens revisadas na literatura, Amaral (2001) apresenta uma das mais simplificadas. O autor procura descrever e agrupar todo processo em duas fases principais. A primeira delas envolveria a preparação dos dados e a segunda, a mineração propriamente dita. Durante este processo, cada resultado encontrado seria registrado em relatórios de descobertas e, com o auxílio de técnicas de visualização, os analistas de mineração procurariam interpretar as informações para, só então, obter o conhecimento.

Vale ainda ressaltar a contribuição de Reinartz (1999). O autor, em seu trabalho, evidencia a importância de documentar a experiência adquirida durante todo o processo.

Na literatura brasileira da área, o que se observa é uma quase total ausência de trabalhos com abordagem de KDD. Praticamente, toda a pesquisa nacional tem sido desenvolvida com enfoque em mineração de dados e *business inteligence*. Uma exceção é o trabalho em que Quintella e Soares Jr. (2003:89) descrevem o KDD de forma muito simplificada como "o processo não-trivial para geração de conhecimento a partir da busca sistemática de padrões em grandes volumes de dados". Tal definição confunde-se com o entendimento geral do que é o *data mining*, assunto tratado a seguir.

Mineração de dados — data mining

Como já mencionado, o *data mining* é tratado como uma das etapas da descoberta de conhecimento em bases de dados. Reconhece-se, no entanto, que nem todo processo de DM é conduzido em um contexto de KDD.

Segundo Cabena e colaboradores (1998:36), *data mining* é a "técnica de extrair informação, previamente desconhecida e de máxima abrangência a partir de bases de dados, para usá-la na tomada de decisão". Han e Kamber (2000:8), por sua

vez, conceituam a técnica de forma mais detalhada e coincidentemente mais alinhada com os objetivos deste artigo. Para eles, *data minning* é "uma etapa na descoberta do conhecimento em bancos de dados que consiste no processo de analisar grandes volumes de dados sob diferentes perspectivas, a fim de descobrir informações úteis que normalmente não estão sendo visíveis". Por outro lado, de acordo com Harrison (1998), o *data mining* contempla a exploração e a análise, por meios analíticos ou semi-analíticos, de grandes quantidades de dados para descobrir modelos e regras significativas, conceito semelhante ao utilizado por Frawley, Piatetsky-Schapiro e Matheus (1992:214).

Uma vez apresentados diferentes definições e conceitos, discute-se, a seguir, as técnicas de operacionalização do DM e suas inter-relações.

Técnicas e funções do data mining

As técnicas empregadas em DM podem ser divididas em dois grandes grupos: heurísticas e matemáticas. Entre as heurísticas, as redes neuronais artificiais são as que mais se destacam, seguidas da inteligência artificial simbolista. Dentro do grupo da matemática, por sua vez, destacam-se a análise estatística e a modelagem matemática.

Os algoritmos de DM mais empregados são comumente divididos em cinco funções: classificação; regressão; associação e modelos de dependência e análise de seqüência; *clustering;* e sumarização. Estas funções são descritas resumidamente a seguir.

CLASSIFICAÇÃO

Para Carvalho (2001) a classificação é uma das funções mais utilizadas no DM, simplesmente porque é uma das tarefas cognitivas humanas mais utilizadas na busca da compreensão do ambiente em que vivemos. Ela pressupõe características que definem grupos específicos e associa ou classifica um item em uma ou várias classes predefinidas (Fayyad et al., 1996).

Os algoritmos clássicos empregados na função classificação baseiam-se em árvores de decisão, regras de decisão e análise discriminante, recomendada para identificar as variáveis (explicativas) que melhor discriminam grupos previamente identificados (variáveis explicadas). A maioria desses algoritmos utiliza a função discriminante de Fischer para dois ou mais grupos.

REGRESSÃO

A função regressão é similar à classificação, diferencia-se desta por objetivar a predição de um valor real em vez de um atributo nominal ou uma categoria.

Com a popularização do computador, os cientistas sociais passaram a utilizar técnicas de regressão até então impossíveis de serem operacionalizadas (Inmon, Terdeman e Imhoff, 2001). Atualmente, as ferramentas de análise de regressão são encontradas nos diversos níveis de plataformas de computação, até mesmo no popular MS-Excel. Existem, no entanto, outros modelos de regressão mais complexos, envolvendo maior número de variáveis explicativas e relacionamentos não-lineares, e entre eles destacam-se a regressão múltipla[1] (RLM), probito e a regressão não-linear.

ASSOCIAÇÃO

Identifica relações significativas existentes entre os eventos ocorridos em determinada ocasião (relações entre campos de um banco de dados) baseada em modelos de dependência. Esses modelos procuram descrever dependências significantes entre variáveis (Agrawal, 1995), podendo ser divididos em dois níveis: o estrutural e o quantitativo. Nos modelos de dependência estruturais, são especificadas as variáveis localmente dependentes umas das outras, enquanto nos modelos de dependência quantitativos são utilizadas escalas numéricas para determinar as forças das dependências entre as variáveis.

Cabena e colaboradores (1998) descrevem a função associação como o processo de interconexão de objetos, na tentativa de expor características e tendências. Os principais métodos são "regras de associação" e "característica seqüencial".[2]

---

[1] A regressão múltipla é o método de análise mais apropriado quando o problema de pesquisa envolve mais de duas variáveis. Na análise de regressão clássica, há uma única variável dependente e múltiplas variáveis de predição (independentes). Quando se procura mensurar a probabilidade de ocorrência dos resultados entre uma variável resposta (explicada) do tipo dicotômica e as variáveis explicativas são categóricas ou contínuas, é utilizada a regressão logística ou modelo logístico. No jargão estatístico, os modelos de predição com classificação e com regressão são chamados, respectivamente, árvore de classificação e árvore de regressão. Para as árvores de regressão devem ser considerados os aspectos relacionados ao ajuste do modelo e sua verificação, bem como a seleção de variáveis explicativas que farão parte do modelo.

[2] Uma regra de associação possui como grande vantagem sua simplicidade. Diferentemente da técnica "característica seqüencial", exige um grande número de registros para assegurar a representatividade dos resultados. Da mesma forma, procura determinar a freqüência de combinação de cada transação que pode ser produzida nas seqüências de registros. Por fim, a análise de seqüência procura identificar desvios e tendências no tempo. Tem comportamento semelhante à associação, diferenciando-se apenas pelo fato de que a relação existe durante um dado período de tempo.

CLUSTERING OU AGRUPAMENTO

Diferentemente da classificação, em que os grupos são predefinidos, os *clusters* são definidos por agrupamentos dos dados baseados em medidas de semelhança ou modelos de densidade de probabilidade. Os grupos são sugeridos pelos dados, e não predefinidos. A fase de *clustering* ou agrupamento é também chamada de análise de classificação, taxonomia numérica ou análise Q (Malhotra, 2001).

A função *clustering* freqüentemente está presente também nas primeiras fases da mineração de dados, com o intuito de reunir os registros em grupos com características em comum para serem utilizados nas fases seguintes. Procura identificar, baseada em modelos probabilísticos ou em medidas de similaridade, grupos (*clusters*) que compartilham de uma característica específica.

O objetivo da função agrupamento é classificar, com base em um conjunto de variáveis considerado, os indivíduos pertencentes a uma população em subconjuntos (*clusters*) relativamente homogêneos. Os principais algoritmos utilizados nessa função já são antigos na estatística, mas só foram disseminados após a difusão dos computadores. Entre os vários algoritmos usados na função *clustering*, merecem destaque aqueles baseados na teoria de conjuntos nebulosos, particularmente apropriados para este fim: o *fuzzy c-means*, o *extended fuzzy c-means* e o algoritmo de agrupamento participativo (Silva, 2003).

SUMARIZAÇÃO

Engloba a organização e o resumo dos dados. É utilizada em uma fase preliminar aos demais modelos ou funções. Visa, principalmente, orientar e motivar análises posteriores mais complexas. Pode ser relacionada à estatística, mais especificamente, à análise exploratória de dados ou estatística descritiva.

Principalmente nos processos de DM, a sumarização utiliza as funções complementares de caracterização e visualização para observar a presença de alguma característica estrutural nos dados. A visualização é um poderoso recurso de análise de dados, sendo muitas vezes suficiente para obter as respostas necessárias. Já a caracterização permite a generalização de qualidades relevantes dos dados através de análises quantitativas que propiciam descrições compactas.

Como um exercício da aplicação do KDD, elegeu-se neste artigo a temática referente à mensuração e distribuição da pobreza na cidade de Salvador, utilizando-se os dados do Censo Demográfico do ano 2000. A possibilidade de conhecer o comportamento deste fenômeno e visualizar sua distribuição espacial com o uso de ferramentas de geoprocessamento torna este estudo bastante desafiador.

4. Pobreza e desigualdade social: conceitos e mensuração

Nesta parte do artigo, buscou-se fazer um breve levantamento acerca de algumas questões que envolvem os conceitos e as formas de mensuração da pobreza, sem pretensão de abarcar, muito menos esgotar a discussão. O propósito desta seção é apenas o de subsidiar o entendimento do tema utilizado neste artigo, como exemplo de aplicação dos processos e sistemas de KDD em bases de dados públicas.

A pobreza no Brasil

O elevado grau de pobreza da população brasileira remonta aos primórdios de sua formação histórica, tendo se mantido presente ao longo do tempo, resistindo ao crescimento da economia e à aparente ampliação das políticas sociais (Schwartzman, 1996). Por outro lado, apesar de não ser novo, o tema vem ganhando importância e visibilidade nos últimos anos. Esta afirmação pode ser ilustrada pelo exemplo de políticas públicas como o Comunidade Solidária e, mais recentemente, o Programa Fome Zero.

Devido à complexidade do problema, a tentativa de mensuração ou de apreensão de situações de pobreza não é uma tarefa fácil. Diversas questões conceituais e metodológicas se interpõem no percurso daqueles que se propõem a estudar tal questão (Lima, 2004).

O termo pobreza encontra a sua origem no adjetivo *pauper-eris*. Sua interpretação tem sofrido variações ao longo dos anos. Atualmente, a compreensão mais comum sobre o conceito de pobreza é associada à falta de renda e ao estado de privação e incapacidade de mobilizar esforços para satisfazer às necessidades básicas do cidadão (Sen, 1992).

Ainda no século XIX autores ingleses (Booth, 1889, 1892; e Rowntree, 1901, segundo Ciaris, 2003) estabeleceram valores mínimos para a questão alimentar humana, em uma abordagem biofisiológica do fenômeno por eles atribuído ao caráter desigual da propriedade dos meios de produção e distribuição de riquezas.

Recentemente esta visão já não é tão amplamente aceita. Por exemplo, na abordagem de Max-Neef, Elizalde e Hopenhayn (1996) pobres são aqueles que não têm atendidas suas necessidades "existenciais" nas esferas do ser, ter, fazer e interagir, além de suas necessidades "axiológicas" de subsistência, liberdade, identidade, participação, ócio, proteção e afeto.

Em função da complexidade da temática, é consenso que seja difícil mensurar a pobreza. Neste artigo foi adotado um conceito de caráter operacional para a medição do fenômeno, definido como "a privação do indivíduo ao acesso ao bem-estar". De forma sincrônica a essa definição, o fenômeno será analisado neste artigo

a partir de uma *proxy* da "renda" e de indicadores socioeconômicos selecionados, reunindo assim elementos de diferentes correntes de pensamento e aliando-se ao enfoque usado por Sen (1992).

Entre outros objetivos, este artigo pretende, através do KDD, mensurar a pobreza a partir de bases de dados públicas. Para isso, será utilizado um conjunto de indicadores diretos de pobreza e uma série de fatores que, presumivelmente, têm um impacto (mesmo que indireto) sobre a situação de privação. Ambos serão descritos e discutidos na seção relativa às escolhas metodológicas a seguir.

5. Procedimentos metodológicos

O objeto do estudo aqui descrito é a descoberta de conhecimento em bases de dados, mais especificamente em bases de dados públicas (DCBDp), tendo como recorte sua aplicação ao estudo da pobreza na cidade de Salvador. Este recorte foi escolhido por ser foco de atenção cada vez maior por parte dos governos, organizações não-governamentais nacionais e internacionais e, naturalmente, institutos de pesquisa e estatística. Já o recorte geográfico pode ser justificado por tratar-se da terceira maior região metropolitana do país, simultânea e paradoxalmente uma das mais pobres. Observa-se ainda que não há literatura suficiente (nem em qualidade nem em quantidade) tratando de aplicações de KDD na área pública no Brasil. Nos poucos trabalhos existentes, observa-se a falta de uma estruturação específica de fases do processo KDD para bases de dados públicas. São apresentadas, a seguir, as opções feitas neste artigo em termos de estrutura, métodos e técnicas de pesquisa apropriados ao contexto do trabalho.

Etapas do KDD

A bibliografia descreve diversas abordagens para o KDD, algumas delas com um encadeamento linear e sucessivo das fases, procedimento este que nesta pesquisa não foi possível, já que as fases aconteceram muitas vezes de maneira simultânea ou, em alguns casos, fora da ordem proposta pelos principais autores. Dessa forma, para esse estudo, optou-se por um modelo híbrido derivado das semelhanças e diferenças observadas nas propostas encontradas na literatura e na experiência dos autores deste trabalho.

Aqui são descritas as duas fases principais, "prospecção" e "mineração" de dados, empregadas durante realização da pesquisa que originou o presente artigo (figura 2).

Conforme pode ser visto na figura, as etapas identificadas para a fase de "prospecção" foram respectivamente: identificação de "objetivos"; "levantamento" (identificação e classificação das fontes existentes e definição do "modelo" de análise); "reunião"; "seleção" e "criação" das bases de dados; "consistência" (limpeza ou eliminação de ruído e enriquecimento) das bases de dados e "compatibilização" das bases de dados.

A fase de "mineração" compreende as etapas: "transformação" dos dados; "função"; "técnicas e algoritmos" e "avaliação" dos resultados.

Comungando com o pensamento de Reinartz (1999) foi feita a documentação de todo o processo, porém, com o objetivo de tornar a leitura mais agradável, optou-se por apresentar neste artigo apenas uma síntese dos principais procedimentos adotados em cada etapa.

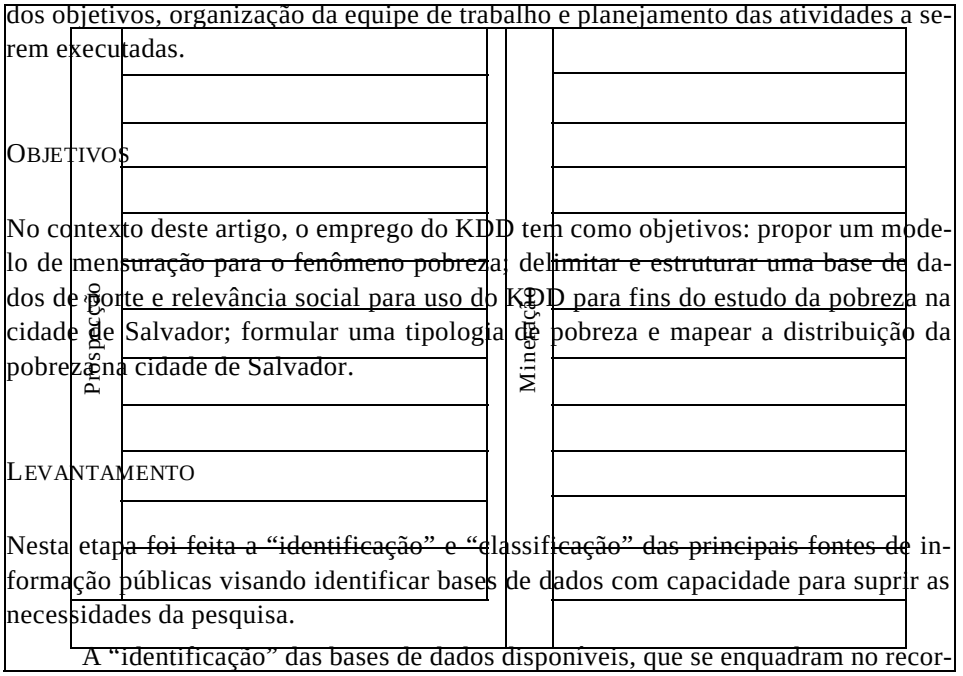
Figura 2

Fases e etapas em um processo de descoberta de conhecimento em bases de dados públicas (DCBDp)

- Objetivo
- Levantamento
  - Identificação
  - Classificação
  - Modelo de análise
- Reunião
- Seleção
- Criação
- Consistência
  - Limpeza
  - Enriquecimento
- Compatibilização
- Transformação
- Função
  - Sumarização
    - Caracterização
    - Vizualização
  - Classificação
  - Associação
  - Regressão
  - *Clustering*
- Técnicas e algoritmos
  - Heurísticas
  - Matemáticas
- Avaliação dos resultados

Prospecção

Em um processo de descoberta de conhecimento em bases de dados, a fase de prospecção inicialmente destina-se à delimitação das perguntas de pesquisa, definição dos objetivos, organização da equipe de trabalho e planejamento das atividades a serem executadas.



OBJETIVOS

No contexto deste artigo, o emprego do KDD tem como objetivos: propor um modelo de mensuração para o fenômeno pobreza; delimitar e estruturar uma base de dados de porte e relevância social para uso do KDD para fins do estudo da pobreza na cidade de Salvador; formular uma tipologia de pobreza e mapear a distribuição da pobreza na cidade de Salvador.

LEVANTAMENTO

Nesta etapa foi feita a "identificação" e "classificação" das principais fontes de informação públicas visando identificar bases de dados com capacidade para suprir as necessidades da pesquisa.

A "identificação" das bases de dados disponíveis, que se enquadram no recorte proposto neste artigo, não foi tarefa das mais difíceis, já que no Brasil apenas o Instituto Brasileiro de Geografia e Estatística (IBGE) dispõe de dados com as características e recortes desejados. Assim, para consecução dos objetivos desta pesquisa foi utilizado o arquivo Agregado de Setores Censitários 2000 (ASC2000), disponibilizado pelo IBGE no site <www.ibge.gov.br>.

No arquivo ASC2000, os dados estão agrupados por unidades da Federação totalizando 215.811 setores censitários para todo território nacional, 15.342 setores para a Bahia e 2.523 setores para a cidade de Salvador. A base de dados analisada é, portanto, relevante, pois, além de ser oficial e pública, abrange informações de todas as pessoas residentes e seus domicílios na cidade de Salvador no ano de realização do último censo.

Para finalizar a etapa de "levantamento" é preciso definir o "modelo" de análise e seus respectivos indicadores. A partir dessa estruturação, buscou-se medir, com um único indicador, um fenômeno de caráter multidimensional — a pobreza, elegendo-se para isso dimensões focadas nas características dos *domicílios*, de seus *responsáveis* e de seus *residentes*. As razões para esta opção serão descritas a seguir.

Para elaborar a tipologia da pobreza para os setores censitários e atender ao modelo de análise, foram selecionados 12 indicadores de privação relativos às características básicas dos domicílios e de seus moradores.

Em sintonia com o que é preconizado na literatura (Sen, 1992; Rocha, 2000 e 2001; Jarman 1983 segundo Lacerda, Calvo e Freitas, 2002; Lopes, 2003; Townsend, 1993), os indicadores de pobreza e desigualdade social foram selecionados a partir das piores condições identificadas nas variáveis existentes no rol disponibilizado pela base ASC2000.

A escolha desse conjunto de indicadores considerou alguns critérios pragmáticos defendidos por Tironi e colaboradores (1991), Quintella e Soares Jr. (2003), Jannuzzi (2001) e Trzesniak (1998). Entre os principais critérios observados destacam-se: relevância, gradação de intensidade, univocidade, padronização, rastreabilidade, estabilidade, representatividade e simplicidade.

Para operacionalização do modelo de análise proposto foram construídos indicadores correspondentes às dimensões de análise detalhadas a seguir.

Domicílio. A dimensão *domicílio* é composta pelos indicadores *abastecimento de água* (v_1), *esgotamento sanitário* (v_2), *destino do lixo* (v_3) e *moradia* (v_4). A escolha dos três primeiros indicadores é justificada pelas práticas da Organização Internacional do Trabalho (OIT) e da Organização das Nações Unidas (ONU) que consideram, entre outros, água, esgotamento sanitário e coleta de lixo como necessidades mínimas de uma família (Lopes, 2003). Já o quarto indicador (moradia) visa identificar os setores censitários com maior número de pessoas por domicílio, o que, em princípio, sugere um maior compartilhamento dos recursos (Merrick, 2002).

Pessoa responsável. Para representar esta dimensão foram selecionados os dados referentes aos moradores em domicílio particular permanente, em função da relação existente entre cada pessoa e o responsável pelo domicílio. A dimensão é composta pelos indicadores *instrução* (v_5 e v_6) e *renda* do responsável (v_7; v_8; v_9 e v_{10}). O indicador *instrução* oferece a possibilidade de identificar maiores concentrações de famílias cerceadas do acesso à educação. Segundo Lopes (2003), entre vários outros autores, a educação é um bem imprescindível para que os indivíduos possam levar vidas saudáveis e ter chances de inserção na sociedade. A *renda* do responsável, por sua vez, é, entre os indicadores que integram o modelo de mensuração da pobreza aqui proposto, o mais universalmente aceito, sendo selecionado com o objetivo de identificar a concentração da população carente, pois, segundo Rocha (2003), em sociedades modernas urbanizadas, boa parte do bem-estar está associada à renda de que as pessoas dispõem para ter acesso a bens e serviços adquiridos no mercado.

Diversos estudos, a exemplo de Schwartzman (1996), Torres e colaboradores (2003) e outros, sinalizam para a correlação entre a pobreza e famílias chefiadas por

mulheres. A opção de distinguir os *responsáveis do sexo feminino* (v_9; v_{10}) também pode ser justificada pela observação de que, "famílias chefiadas por mulheres com baixa escolaridade apresentam altas correlações com renda familiar baixa e presença de apenas um provedor adulto" (Torres et al., 2003:24).

Pessoas residentes. Mingione (1999), fundamentado em inúmeras pesquisas, relaciona o aumento da freqüência de crianças nas famílias à situação de pobreza. Em sintonia com esta constatação a terceira e última dimensão do modelo proposto é composta pelos indicadores: alta incidência de crianças com idade até seis anos no domicílio (v_{11}). Por outro lado, como já mencionado, é notório que a ausência da educação possui forte associação com a pobreza. Assim sendo, o outro indicador que compõe esta dimensão é a elevada proporção de pessoas residentes *não-alfabetizadas* com mais de 10 anos (v_{12}).

Uma vez estabelecidos os objetivos, feita a "identificação" e "classificação" das fontes, definido o "modelo" de análise a ser utilizado com suas respectivas dimensões, selecionados, avaliados e justificados os indicadores, deu-se início à criação da base de dados efetiva — "jazida de dados" [3] — que serviu ao processo KDD.

A partir deste ponto, os dados das fontes selecionadas foram trabalhados com o objetivo de estruturar a "jazida de dados" para atender à fase da mineração de dados.

Reunião, seleção e criação

O emprego do KDD pressupõe que serão trabalhadas bases de dados já existentes (dados secundários), freqüentemente essas bases são provenientes da agregação de outras bases de dados.[4]

Antes da etapa de *reunião*, normalmente os registros e as variáveis de interesse para o estudo estão dispersos em vários arquivos e em diferentes formatos, no caso desta pesquisa a base de dados ASC2000 utilizada estava estruturada em matrizes sob a forma de planilhas agrupadas por unidades da Federação e subdivididas em quatro pastas: domicílio; pessoas — características gerais; pessoas — instrução e responsável pelo domicílio.

---

[3] A denominação alternativa para "base de dados" adotada neste texto (jazida de dados) origina-se da mesma metáfora tradicionalmente adotada pelos usuários da mineração de dados, ou seja, o forte paralelismo existente entre as atividades de quem busca conhecimento em bases de dados e daqueles que buscam por minérios em bases territoriais.

[4] Eventualmente podem ser, também, usados dados primários agregando-os às bases preexistentes para o emprego do KDD.

Para efeito de processamento do KDD, na etapa de *seleção* foram apurados os dados referentes apenas ao município de Salvador. Em seguida, foram excluídos 21 setores censitários[5] considerados áreas não-urbanas.

Visando uma melhor aproximação do fenômeno, optou-se por trabalhar apenas com os domicílios particulares permanentes construídos para servir exclusivamente à habitação. Foram excluídos da base de dados os domicílios particulares improvisados[6] e coletivos.[7] A população da pesquisa, portanto, foi constituída de todos os setores comuns ou não-especiais e seus respectivos domicílios e moradores residentes na área urbanizada da cidade de Salvador em 1º de agosto de 2000.

Por fim, o emprego da etapa de *criação* possibilitou a elaboração de um arquivo em formato compatível com o conjunto de softwares empregados. Este procedimento de manter apenas um arquivo, em um único formato, com todos os dados que foram trabalhados, favoreceu significativamente as etapas seguintes, tanto em relação à performance quanto à praticidade das operações subseqüentes.

A partir do *subset* de dados oriundos das fases de *reunião*, *seleção* e *criação*, deu-se início à fase de *consistência* com dados de todos os domicílios de Salvador, totalizando 2.502 setores censitários.

A base de dados resultante foi composta por 103 variáveis, das 527 disponíveis nos quatro arquivos originais.

CONSISTÊNCIA

A principal vantagem de se trabalhar com bases de dados provenientes de estatísticas oficiais decorre do fato de que estas, usualmente, passaram previamente por um processo de consistência.

Como na presente pesquisa utilizou-se de dados procedentes do IBGE, as subetapas de *limpeza* ou *eliminação de ruído* já haviam sido executadas, tornando-se necessária apenas uma rápida verificação para tratar os registros incompletos. As-

[5] Áreas não-urbanizadas de cidade; áreas urbanas isoladas; aglomerados rurais de extensão urbana; aglomerados rurais isolados — tais como zonas rurais existentes em algumas ilhas pertencentes ao município.

[6] Para o IBGE o domicílio particular improvisado foi aquele localizado em unidade não-residencial que não tinha dependências destinadas exclusivamente à moradia, mas que, na data de referência, estava ocupado por morador. São enquadrados nesta definição as lojas, fábricas, os prédios em construção, vagões de trem, carroças, tendas, barracas, grutas etc.

[7] O domicílio coletivo é caracterizado quando a relação entre as pessoas que nele habitavam é restrita a normas de subordinação administrativa. Ficam incluídos nesta definição os hotéis, pensões, presídios, cadeias, penitenciárias, quartéis, postos militares, asilos, orfanatos, conventos, hospitais e clínicas (com internação), alojamento de trabalhadores, motéis, campings etc.

sim como, em função da inexistência de outra base de dados com a abrangência e nível de detalhe (granularidade) dos dados utilizados, também não foi possível realizar a etapa de *enriquecimento*.

COMPATIBILIZAÇÃO

A fase de compatibilização envolve a unificação das diferentes bases de dados originais já consistidas, resultando na "jazida de dados". Nas abordagens tradicionais de KDD, a fase de compatibilização poderia também resultar no *data warehouse.*

Após a etapa de *compatibilização*, a jazida de dados passa a ter o formato requerido para as transformações sintáticas e semânticas que compõem o início da mineração de dados.

Mineração

Após o desenlace do processo de *prospecção*, dá-se início à mineração de dados. Observa-se aqui que este segundo processo é mais complexo que o primeiro, embora, paradoxalmente, seja o que exige menor tempo para sua execução. No KDD, assim como na pesquisa mineral, freqüentemente dedica-se mais tempo à delimitação da jazida que à sua explotação.

No decorrer da mineração de dados realizada durante esta pesquisa, foram executadas as tarefas de "transformação" dos dados; escolha da "função" de mineração; "técnica e algoritmo" de busca e "avaliação" dos resultados.

TRANSFORMAÇÃO

Durante esta etapa a "jazida de dados" sofre uma transformação sintática e semântica. A transformação sintática é aquela que não altera o significado dos dados, visa apenas atender os requisitos das ferramentas de mineração utilizadas nas etapas subseqüentes. Já a transformação semântica busca atender, com o cálculo de indicadores, o modelo de análise previamente definido.

FUNÇÕES E ALGORITMOS

Após todo o trabalho de *prospecção* e de posse da "jazida de dados" transformada, dá-se início à escolha da função ou conjunto de funções, no caso desta pesquisa, *sumarização, associação, regressão* e *clustering.*

A seleção da *função* determina a maneira como é feita a busca por reconhecimento de padrões e relacionamentos complexos, o sucesso desta seleção, para Diniz e Louzada Neto (2000:28), "está diretamente ligado à experiência e intuição do analista".

Neste artigo, o conjunto de *funções* selecionadas para esta etapa foi:

- análise preliminar dos dados pela *sumarização*;
- cálculo da matriz de correlações para os indicadores (variáveis) selecionados pela função *associação* e redução de dimensionalidade com a análise de componentes principais;
- emprego da função *regressão* para obtenção do índice de pobreza para cada setor censitário estudado;
- aplicação da função *cluster* para posterior emprego na construção da tipologia proposta para o fenômeno da pobreza em Salvador;
- *sumarização* e suas funções complementares: *caracterização* e *visualização* dos *clusters* encontrados na etapa anterior.

Definido o conjunto de funções e seu respectivo encadeamento, partiu-se então para a busca do grupo de *técnicas e algoritmos* mais apropriados para cada *função.* Nesta pesquisa optou-se por trabalhar com algoritmos derivados da análise estatística.

Descreve-se a seguir cada uma dessas etapas.

SUMARIZAÇÃO

Foi aplicada em dois momentos. No primeiro, buscou-se um estudo preliminar dos indicadores selecionados e armazenados na "jazida de dados", bem como a orientação quanto à escolha das técnicas para as funções de mineração subseqüentes. No segundo momento, após as funções *associação*, *regressão* e *clustering*, foi possível elaborar uma síntese dos tipos de pobreza descobertos durante o processo. Nos dois momentos foi aplicada a função complementar *caracterização*, optando-se por estatística descritiva por mera questão de disponibilidade de software. Cabe ressaltar que também poderiam ser utilizadas as técnicas de *SQL* tradicional ou *Olap*, entre outras.

A função complementar *visualização,* por sua vez, teve maior participação no segundo momento. Este importante recurso de análise contribuiu de forma significativa para a interpretação dos resultados. Além dos tradicionais gráficos de análise, foram utilizados recursos de geoprocessamento, principalmente na elaboração de cartogramas da tipologia obtida. Foram construídos gráficos, tabelas e cartogramas dos resultados obtidos durante e após a conclusão da etapa de *cluster*.

ASSOCIAÇÃO

Para atender a esta função, foi escolhido o método de redução de dimensionalidades análise de componentes principais (ACP) para reduzir o número de variáveis. Com a ACP foi possível identificar um subconjunto de 12 indicadores correlacionados com pobreza nas 527 variáveis dos 2.502 setores censitários. Dessa forma, contribuindo para confirmação das dimensões selecionadas — domicílio, família e pessoa — conforme o modelo de análise proposto na fase de levantamento do processo de prospecção.

REGRESSÃO

A partir dos 12 indicadores validados durante a etapa de *associação* foi utilizada a função *regressão* para obter os escores fatoriais de cada setor censitário, com o objetivo de construir o *índice de pobreza*. A construção do índice com a técnica estatística "regressão" possibilitou caracterizar e hierarquizar os setores censitários segundo as dimensões de análise da pobreza estudadas.

AGRUPAMENTO

Após a modelagem do fenômeno da pobreza obtida nas fases anteriores e de acordo com o índice de pobreza construído, foi aplicada a função *clustering* com o objetivo de particionar os setores censitários e reuni-los em grupos homogêneos de pobreza e assim permitir formular uma tipologia do fenômeno.

Conforme orienta Hair Jr. e colaboradores (1992:269), a aplicação da função *clustering* foi dividida em três diferentes estágios: particionamento, interpretação e validação. Este procedimento auxiliou na identificação dos setores censitários pertencentes a cada uma das classes homogêneas e mutuamente exclusivas de pobreza, bem como a descrever as características de cada uma delas.

6. Avaliação dos resultados

Esta seção tem o objetivo de apresentar os principais resultados da aplicação do KDD nas bases de dados públicas trabalhadas, visando à identificação e mensuração da pobreza na cidade de Salvador. Assim como na etapa de "avaliação" de resultados do processo do KDD em sua fase de mineração, serão aqui apresentadas as "funções" utilizadas, bem como as respectivas verificações de adequação das "técnicas e algoritmos" empregados e os principais resultados encontrados (conhecimento descoberto).

Os resultados da fase de mineração devem ser avaliados sob quatro aspectos: adequação do modelo de análise, conveniência das funções, adequação das técnicas e algoritmos escolhidos para processá-las e principalmente os achados e descobertas de conhecimento resultantes da fase de mineração e, conseqüentemente, de todo o processo de KDD.

Deve-se observar que enquanto para a fase de prospecção o produto final é a "jazida de dados", na fase de mineração os resultados esperados podem ser representados, de maneira simplificada, como: resultado = $f(a, b, c, d)$.

Sumarização da base de dados

Descreve-se brevemente nesta subseção apenas a sumarização de cunho exploratório da "jazida de dados transformada". A maior parte da função sumarização realizada, no entanto, deu-se na descrição dos *clusters* que compuseram a tipologia desenvolvida, não sendo aqui apresentada por não fazer parte dos objetivos deste artigo.

Foram estudados 2.439.255 habitantes distribuídos em 651.051 domicílios particulares permanentes das zonas urbanas da cidade de Salvador. Segundo os dados investigados, no ano 2000 existiam na capital baiana 51.030 domicílios em condições de privação no abastecimento de água; 107.949 domicílios em condição de privação de esgotamento sanitário; 42.871 apresentando privação de coleta de lixo e 51.425 domicílios em condições precárias de moradia.

Ao analisar a dimensão família, constata-se que 214.971 chefes de família, em agosto de 2000, possuíam menos de quatro anos de estudo, sendo que, entre estes, 91.227 são do sexo feminino. No que tange à renda a situação é mais alarmante, 338.841 chefes de família declararam-se sem rendimentos ou com rendimentos iguais ou inferiores a dois salários mínimos, destes 153.387 são mulheres. Por fim, 293.707 habitantes estudados são crianças com idade até seis anos e 124.517 são pessoas residentes não-alfabetizadas com mais de 10 anos.

A partir de uma primeira sumarização dos dados é possível obter um panorama mais geral dos resultados e também verificar a ocorrência de violação das premissas

que muitas técnicas exigem, como por exemplo a normalidade, homocedasticidade e linearidade, e, dessa maneira, considerar a possibilidade de aplicar ou não certos algoritmos.

Resultado da associação

Com o objetivo de testar a conveniência da técnica fatorial, os dados foram submetidos ao teste de esfericidade de Bartlett, sendo, em seguida, calculada a medida de Kaiser-Meyer-Olkin (KMO) para todos os 2.502 setores censitários da capital baiana. Os resultados encontrados demonstraram que a análise fatorial é apropriada e que pode, portanto, ser utilizada na mineração desses dados. O resultado da medida KMO foi 0,848, o que pode ser considerado muito bom, já o teste de esfericidade de Bartlett, com significância 0 para um qui-quadrado 43.932,969 com 66 graus de liberdade foi aceito. Assim sendo, a etapa de análise fatorial contribuiu para a seleção dos indicadores mais relevantes para o estudo e posterior validação do modelo de análise a ser empregado. Resumidamente, a verificação de adequação do modelo de análise proposto consistiu em dois passos:

- seleção dos indicadores significantes — após a extração pelo método da análise das componentes principais, todos os indicadores apresentaram comunalidade maior que 50%, ratificando a presença de todos os indicadores no modelo de mensuração da pobreza;
- determinação do número de dimensões de análise selecionadas — utilizando o critério adotado por vários autores, entre eles Johnson e Wichern (1998) e Hair Jr. e colaboradores (1992), foram aceitas apenas duas dimensões para compor o modelo, visto que a primeira e segunda dimensões possuem autovalores, 7,954 e 1,497 respectivamente, ambos superiores à unidade preconizada pelos autores.

Diante dos resultados encontrados, conclui-se que a construção abstrata, feita a partir das definições e convenções terminológicas sobre o conceito da pobreza discutido ao longo deste artigo, pode representar a realidade através de apenas duas dimensões (no caso, domicílio e família) em vez das três (domicílio, responsável e residentes) inicialmente propostas.

A análise dos resultados para os 12 indicadores (definidos na etapa de levantamento) permite observar, após a rotação pelo método Varimax, que as duas dimensões de análise selecionadas para o modelo permitem explicar 78,76% da variância total dos indicadores.

Confirmado o modelo de análise para mensuração da pobreza em Salvador, contemplando duas dimensões (domicílio e família), foi calculado o índice de pobreza através da função regressão, conforme apresentado na próxima subseção.

Resultado da regressão

A construção do índice de pobreza (IP) foi motivada pela necessidade de classificar os setores censitários segundo a condição de pobreza encontrada. Dessa forma, foi possível avaliar a condição do setor censitário como um todo e assim identificar aqueles setores com maior probabilidade de se encontrar moradores e domicílios em condições de privação. O índice de pobreza foi construído a partir dos valores dos escores fatoriais para cada setor censitário. Concluída essa função e de posse da hierarquização dos setores censitários, obtida a partir dos índices de pobreza, foi dado prosseguimento à análise de agrupamento com a função *cluster*.

Resultado do agrupamento

Com o objetivo de evidenciar as associações entre as variáveis, permitindo o agrupamento segundo suas similitudes, optou-se pela utilização função *cluster* da mineração de dados. Mais uma vez recorreu-se à técnica estatística, mais especificamente à análise de agrupamentos. Foi selecionado o algoritmo hierárquico para executar a função *cluster* e o número de classes foi definido a partir da análise do dendograma obtido. Esta análise ofereceu a possibilidade de subdividir os 2.502 setores censitários em nove ou cinco tipos (grupos). Diante das dificuldades de se especificar as características distintivas entre nove grupos, optou-se por trabalhar com a segunda opção, acreditando-se que esta representaria de maneira satisfatória a distribuição do fenômeno estudado.

Tipologia da pobreza

Para construir uma tipologia de pobreza para os setores censitários da cidade de Salvador foi necessário classificá-los segundo suas características, descritas no modelo de análise e sintetizadas através do índice de pobreza (IP) calculado. A elaboração da tipologia aqui proposta foi feita utilizando uma fusão de dois métodos: o conceitual-analítico (heurístico) e a taxonomia numérica (matemático). O que norteou toda construção da tipologia aqui proposta foi a finalidade à qual a pesquisa se prestava a atender: “mapear a distribuição da pobreza”.

Os setores censitários foram classificados como de “pobreza muito alta” se seu índice de pobreza (IP) for superior a 0,700 — nesta condição encontram-se 111 setores censitários. Já os 517 setores com IP até 0,180 ficam situados no tipo de “pobreza muito baixa”. O tipo mais numeroso, de “pobreza moderada”, reúne os setores com índice entre 0,410 e 0,494. Os demais tipos “pobreza alta” e “pobreza baixa” possuem IP entre 0,580 e 0,628, e IP entre 0,180 e 0,311 respectivamente.

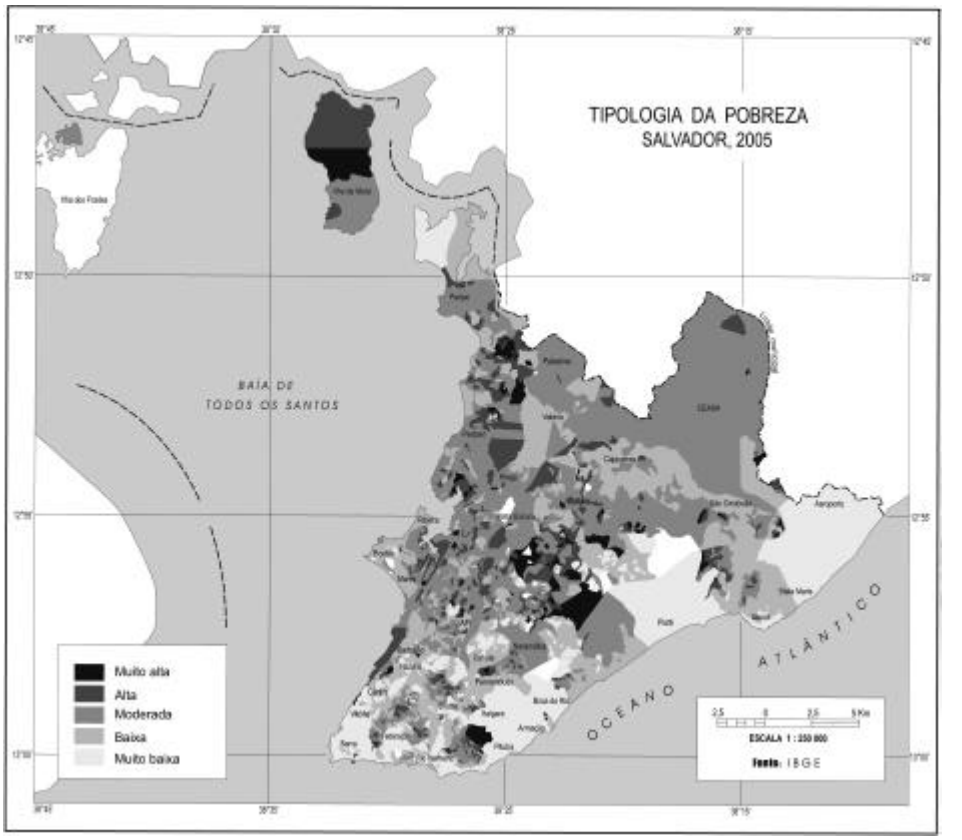
O resultado final dessas escolhas e procedimentos (mapeamento) é descrito a seguir.

7. Considerações finais

Foi possível observar que a aplicação do KDD em bases de dados públicas permitiu identificar, mesmo com as limitações impostas pelas bases de dados utilizadas, quais indicadores, em termos de saneamento básico, educação, condições de moradia, renda etc., estão associados aos altos níveis de pobreza em Salvador, resultando na elaboração de um índice de pobreza que reflete as múltiplas dimensões que envolvem o fenômeno. A visualização integrada desse resultado está sintetizada na figura 3.

Figura 3

Distribuição da tipologia da pobreza por setor censitário — Salvador, 2005



Neste artigo diversas contribuições foram apresentadas ao estudo do KDD ou mais especificamente do KDDp, entre as principais destacam-se: elaboração de uma proposta de estruturação e sistematização de etapas para o processo KDD em bases de dados públicas; a modelagem do fenômeno social da pobreza, oferecendo ao gestor público a possibilidade de ajustar a política de acordo com as características de cada grupo; o mapeamento do fenômeno (que oferece uma análise ampla e sistêmica da pobreza e poderá ser útil no desenvolvimento de ações antipobreza); outra importante contribuição que esta pesquisa traz está relacionada à demonstração da importância do KDD na construção de índices baseados em dados socioeconômicos pela elaboração de modelos multidimensionais de análise e dos métodos utilizados para aglutinação dos indicadores; o estudo demonstrou a possibilidade de utilização

das informações de bases de dados públicas, especialmente do IBGE na identificação de grupos homogêneos de pobreza na capital baiana (a utilização do KDD e a facilidade e rapidez do acesso a dados secundários, potencializam o seu uso como instrumento de planejamento).

Os resultados deste estudo demonstram para os analistas de dados e especialistas do domínio que a utilização dos algoritmos disponíveis nos softwares de DM ou estatística exige, além de uma postura responsável, o conhecimento aprofundado de cada etapa do processo, bem como do domínio que está sendo estudado. Espera-se que este artigo contribua para uma reflexão acerca da forma com que essas bases de dados públicas vêm sendo utilizadas.

Referências bibliográficas

ADRIAANS, P.; ZANTIGE, D. *Data mining.* Harlow: Addison-Wesley, 1996.

AGRAWAL, R. et al. Fast discovery of association rules. In: FAYYAD, U. M. et al. (Eds.). *Advances in knowledge discovery and data mining*. Menlo Park: AAAI/MIT Press, 1995.

AMARAL, Fernanda Cristina. *Data mining:* técnicas e aplicações para o marketing direto. São Paulo: Berkeley Brasil, 2001.

ANDERBERG, Michael R. *Cluster analysis for applications*. New York: Academic Press, 1973.

BECKMAN, T. The current state of knowledge management. In: LIEBOWITZ, J. (Ed.). *Knowledge management handbook*. New York: CRC Press, 1999.

BRACHMAN, Ronald J.; ANAND, Tej. The process of knowledge discovery in databases. In: *Advances in knowledge discovery and data mining.* Menlo Park: AAAI Press, 1996.

BUSSAB, W. de O.; MIAZAKI, E. S.; ANDRADE, D. F. Introdução à análise de agrupamentos. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 9., 1990. *Anais...* São Paulo. Associação Brasileira de Estatística.

CABENA P. et al. *Discovering data mining*: from concept to implementation. Englewood Cliffs: Prentice Hall, 1998.

CARVALHO, Luís Alfredo Vidal de. *A mineração de dados no marketing*. São Paulo: Érica, 2001.

CIARIS (CENTRO DE APRENDIZAGEM E DE RECURSOS PARA A INCLUSÃO SOCIAL). *Uma questão terminológica?* 2003. Disponível em: <http://ciaris.ilo.org/portugue/frame/r1-2.htm>. Acesso em: 20 jun. 2004.

DAMIANI, W. B. Estudo do uso de sistemas de apoio ao executivo (EIS — Executive Information Systems). In: ENCONTRO ANUAL DA ANPAD, 22., Foz do Iguaçu, 1998. *Anais...* Foz do Iguaçu: Enanpad, 1998.

DAVENPORT, T. H.; PRUSAK, L. *Conhecimento empresarial*: como as organizações gerenciam o seu capital intelectual. Rio de Janeiro. Campus, 1998.

DINIZ, Carlos Alberto R.; LOUZADA NETO, Francisco. *Data mining:* uma introdução. São Paulo: ABE, 2000.

FAYYAD, Usama M. et al. *Advances in knowledge discovery and data mining*. Menlo Park: AAAI Press, 1996.

———; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, New York: ACM Press, v. 39, n. 11, p. 27-34, Nov. 1996.

FRAWLEY, W.; PIATETSKY-SCHAPIRO, G.; MATHEUS, C. Knowledge discovery in databases: an overview. *AI Magazine*, p. 213-228, Fall 1992.

FRIEDMAN, Batya; KAHAN Jr.; Peter H. Educating computer scientists: inking the social and the technical. *Communications of the ACM*, v. 37, n. 1, p. 65-70, Jan. 1999.

HAIR JR., Joseph F. et al. *Multivariate data analysis.* 5. ed. Upper Saddle River (NJ): Prentice Hall, 1992.

HAN, J.; KAMBER, M. *Data mining*: concepts ans techiniques. New York: Morgan Kaufmann, 2000.

HARRISON, Thomas H. *Intranet data warehouse*. São Paulo: Berkeley, 1998.

INMON, W. H.; TERDEMAN, R. H.; IMHOFF, Claudia. *Data warehousing*: como transformar informações em oportunidades de negócios. São Paulo: Berkeley, 2001.

IVANOV, K. *Strategies and design for information technology*: Eastern or neo-romantic wholes, and the return to Western systems. Aix-en-Provence: University of Aix-Marseille III, 1998.

JANNUZZI, P. M. Repensando a prática de uso de indicadores sociais na formulação e avaliação de políticas públicas municipais. In: ENCONTRO NACIONAL DA ANPAD, 25., Campinas, 2001. *Anais...* Campinas: Anpad, 2001.

JOHNSON, Richard A.; WICHERN, Dean W. *Applied multivariate statistical analysis.* 4. ed. Saddle River: Prentice Hall, 1998.

LACERDA, Josimari Telino de; CALVO, Maria Cristina Marino; FREITAS, Sérgio Fernando Torres de. Intra-urban differentials in Florianópolis, Santa Catarina State, Brazil, and their potential use in health services planning. *Cad. Saúde Pública*, v. 18, n. 5, p. 1331-1338, Sept./Oct. 2002. Disponível em: <www.scielosp.org>. Acesso em: 10 dez. 2004.

LAUDON, K. C.; LAUDON, J. P. *Management information systems*: organization and technology. 3. ed. New York: McMillan, 1994.

LIEBOWITZ, J.; BECKMAN, T. *Knowledge organizations*: what every manager should know. Boca Raton: CRC Press, 1998.

LIMA, Ana Luiza M. de Codes. Mensuração da pobreza: uma reflexão sobre a necessidade de articulação de diferentes indicadores. *Caderno CRH*, Salvador, n. 1, 2004.

LOPES, Marra Helger. *Análise de pobreza com indicadores multidimensionais*: uma aplicação para Brasil e Minas Gerais. 2003. Dissertação (Mestrado em Economia) — Faculdade de Economia, Universidade de Minas Gerais, Belo Horizonte.

MALHOTRA, Naresh K. *Pesquisa de marketing*: uma orientação aplicada. 3. ed. Porto Alegre: Bookman, 2001.

MAX-NEEF, M.; ELIZALDE, A.; HOPENHAYN, M. *Desarrollo a escala humana una opción para el futuro*. Cepaur, Fundación Dag Hammarskjõld. Medellín, Colombia: Proyecto 20 Editores, 1996.

MERRICK, B. G. *The ethics of hiring in the new workplace*: men and women managers face changing stereotypes discover correlative patterns for success. Indiana: Competitiveness Review, 2002.

MINGIONE, Enzo. Urban poverty in the advanced industrial world: concepts, analysis and debates. In: ———. *Urban poverty and the underclass.* New York: Blackwell, 1999.

NONAKA, I.; TAKEUCHI, H. *Criação de conhecimento na empresa*: como as empresas japonesas geram a dinâmica da inovação. São Paulo: Campus, 1997.

O'DELL, Carla; GRAYSON Jr., C. Jackson; ESSAIDES, Nilly. *Ah. Se soubéssemos antes o que sabemos agora*: as melhores práticas gerenciais ao alcance de todos. São Paulo: Futura, 2000.

QUINTELLA, Rogério Hermida; SOARES JUNIOR, Jair Sampaio. Sistemas de apoio à decisão e descoberta de conhecimento em bases de dados: uma aplicação potencial em políticas públicas. *Organizações e Sociedade*, Salvador, v. 28, p. 83-98, 2003.

REINARTZ, Thomas. *Focusing solution for data mining*: analytical studies and experimental results in real-world domains. New York: Springer-Verlag, 1999.

ROCHA, S. Estimação de linhas de indigência e de pobreza: opções metodológicas no Brasil. In: HENRIQUES, R. O. (Ed.). *Desigualdade e pobreza no Brasil*. Rio de Janeiro: Ipea, 2000. p. 109-127.

———. Medindo a pobreza no Brasil: evolução metodológica e requisitos de informação básica. In: LISBOA, M. B.; MENEZES-FILHO, N. A. (Orgs.). *Microeconomia e sociedade no Brasil*. Rio de Janeiro: Contra Capa, 2001. p. 51-78.

———. *Pobreza no Brasil*: afinal, de que se trata? Rio de Janeiro: FGV, 2003.

SCHULER, D. Social computing. *Communications of the ACM*, New York: ACM Press. v. 37, n. 1, p. 28-29, Jan. 1994.

SCHWARTZMAN, Simon. *As diversas faces da pobreza no Brasil*. 1996. Disponível em: <www.schwartzman.org.br/simon/pobreza.htm>. Acesso em: 10 out. 2004.

SEN, A. *Inequality reexamined.* New York: Russell Sage, 1992.

SILVA, L. *Aprendizagem participativa em agrupamento nebuloso de dados*. 2003. Dissertação (Mestrado em Engenharia) — Faculdade de Engenharia Elétrica e de Computação, Unicamp, Campinas.

SOARES JUNIOR, Jair Sampaio; QUINTELLA, Rogério Hermida. Indicadores sociais de baixo custo e sua utilidade na gestão da interface entre os governos estadual e municipal. *Organizações e Sociedade*, Salvador, v. 25, p. 45-60, 2002.

TORRES, Haroldo da Gama; MARQUES, Eduardo; FERREIRA; Maria Paula; BITAR, Sandra. Pobreza e espaço: padrões de segregação em São Paulo. *Estudos Avançados*, São Paulo, v. 17, n. 47, p. 97-128, 2003.

SPRAGUE, R. H.; WATSON, H. J. *Sistema de apoio à decisão*. Rio de Janeiro: Campus, 1991.

TIRONI, L. F. et al. *Critérios para geração de indicadores de qualidade e produtividade no serviço público*. Brasília: Ipea/MEFP, 1991.

TOWNSEND, P. Conceptualising poverty. In: ———. *The international analysis of poverty*. London: Harvester Wheatsheaf, 1993.

TRZESNIAK, P. Indicadores quantitativos: reflexões que antecedem seu estabelecimento. *Revista de Ciência da Informação*, Brasília, v. 27, n. 2, p. 159-164, maio/ago. 1998.