

Conceitos e Aplicações de *Data Mining*

Data Mining Concepts and Applications

HELOISA HELENA SFERRA

Universidade Metodista de Piracicaba (Piracicaba, Brasil)
hhsferra@uol.com.br

ÂNGELA M. C. JORGE CORRÊA

Universidade Metodista de Piracicaba (Piracicaba, Brasil)
ajcorrea@unimep

RESUMO Atualmente, muito se fala em *Data Mining*, encontrando-se na literatura significativa variedade de estudos sobre o tema. Este artigo tem como objetivo introduzir conceitos básicos dessa tecnologia a interessados que ainda estão iniciando o estudo de *Data Mining*. Nesse contexto, o presente texto pretende apresentar alguns desses conceitos sobre as técnicas que envolvem a descoberta de conhecimento em grandes conjuntos de dados, além de registrar algumas características de um software específico para mineração de dados, o *Clementine*, da SPSS, bem como algumas aplicações realizadas nessa ferramenta.

Palavras-chave MINERAÇÃO DE DADOS – DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS – ANÁLISE EXPLORATÓRIA DE DADOS – MODELOS ESTATÍSTICOS DE RELACIONAMENTO ENTRE VARIÁVEIS – *CLEMENTINE*/SPSS.

ABSTRACT Much is said about Data Mining nowadays and there is a significant variety of studies on the subject. This paper's aim is to introduce some of the technology's basic concepts to those who are beginning their studies on Data Mining. In such context, the present article presents some of the concepts related to the techniques involved in knowledge discovery within large databases. Moreover, it presents some features of a specific software for Data Mining: Clementine, from SPSS. The paper also indicates some applications for this tool's use.

Keywords DATA MINING – KNOWLEDGE DISCOVERY IN DATABASES – EXPLORATORY ANALYSIS – STATISTICAL MODELS OF RELATIONSHIP BETWEEN VARIABLES – *CLEMENTINE*/SPSS.

INTRODUÇÃO

A rápida evolução dos recursos computacionais ocorrida nos últimos anos permitiu que, simultaneamente, fossem gerados grandes volumes de dados. Estima-se que a quantidade de informação no mundo dobra a cada 20 meses e que o tamanho e a quantidade dos bancos de dados crescem com velocidade ainda maior (Dilly, 1999). O explosivo crescimento do volume de dados tem gerado uma urgente necessidade de novas técnicas e ferramentas capazes de transformar, de forma inteligente e automática, terabytes de dados em informações significativas e em conhecimento. Essas informações, de grande valia para o planejamento, gestão e tomadas de decisão, estão, na verdade, implícitas e/ou escondidas sob uma montanha de dados, e não podem ser descobertas ou, no mínimo, facilmente identificadas utilizando-se sistemas convencionais de gerenciamento de banco de dados. Em resposta a essa necessidade, surgiu o *Data Mining* (DM), também chamado de Mineração de Dados.

Data Mining é uma tecnologia que emergiu da intersecção de três áreas: estatística clássica, inteligência artificial e aprendizado de máquina, sendo a primeira a mais antiga delas. Observa-se que o *Data Mining* é parte de um processo maior conhecido como KDD (*Knowledge Discovery in Databases*) – em português, Descoberta de Conhecimento em Bases de Dados –, que, segundo Addrians & Zantinge (1996), permite a extração não trivial de conhecimento previamente desconhecido e potencialmente útil de um banco de dados. Esse conceito é enfatizado por Fayyad *et al.* (1996b), ao afirmar que é “o processo não trivial de identificação de padrões válidos, desconhecidos, potencialmente úteis e, no final das contas, compreensíveis em dados”.

Nesse contexto, o presente artigo tem como finalidade apresentar conceitos sobre as principais técnicas que envolvem a descoberta de conhecimento em grandes conjuntos de dados e relatar algumas características de um software específico para mineração de dados, o *Clementine*, da SPSS, bem como aplicações realizadas nesta ferramenta. Assim, são mostrados o processo de descoberta de conhecimento (KDD) e o *Data Mining* (DM), como parte desse processo, bem como suas técnicas e as metodologias estatísticas que as fundamentam. Em seguida, são discutidas as características de uma ferramenta de *Data Mining*, o *Clementine*, da SPSS, com a qual se desenvolve a aplicação relatada neste texto.

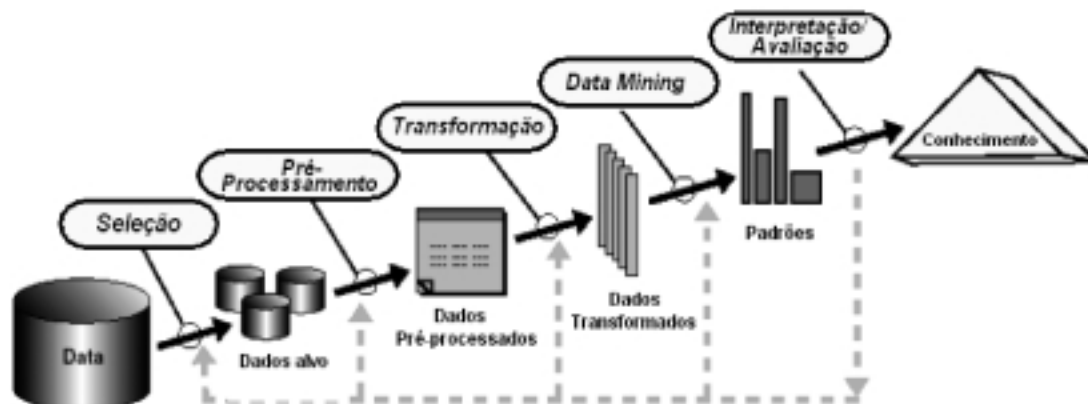
DESCOBERTA DE CONHECIMENTO (KDD) E *DATA MINING* (DM)

Considere-se uma hierarquia de complexidade: se algum significado especial é atribuído a um dado, ele se transforma em uma informação (ou fato). De acordo com Sade (1996), se uma norma (ou regra) é elaborada, a interpretação do confronto entre o fato e a regra constitui um conhecimento.

O processo KDD é constituído de várias etapas, como ilustrado na figura 1, que são executadas de forma interativa e iterativa. De acordo com Brachman & Anand (1996), as etapas são interativas porque envolvem a cooperação da pessoa responsável pela análise de dados, cujo conhecimento sobre o domínio orientará a execução do processo. Por sua vez, a iteração deve-se ao fato de que, com frequência, esse processo não é executado de forma sequencial, mas envolve repetidas seleções de parâmetros e conjunto de dados, aplicações das técnicas de *Data Mining* e posterior análise dos resultados obtidos, a fim de refinar os conhecimentos extraídos.

Dentre as várias etapas do processo KDD, a principal, que forma o núcleo do processo e que, muitas vezes, confunde-se com ele, chama-se *Data Mining*.

Fig. 1. Visão geral das etapas que constituem o processo KDD (Fayyad *et al.*, 1996b).



Esse processo tem início com o entendimento do domínio da aplicação e dos objetivos a serem atingidos. Em seguida, é realizado um agrupamento organizado da massa de dados alvo da descoberta. Como em toda análise quantitativa, a qualidade dos dados é essencial para a obtenção de resultados confiáveis. Portanto, dados limpos e compreensíveis são requisitos básicos para o sucesso do *Data Mining*, como afirmam Diniz & Louzada-Neto (2000). A limpeza dos dados (identificada na literatura como *Data Cleaning*) é realizada por meio de um pré-processamento, visando assegurar a qualidade dos dados selecionados. Destaca-se que, segundo Mannila (1996), essa etapa pode tomar até 80% do tempo necessário para todo o processo, devido às dificuldades de integração de bases de dados heterogêneas.

Os dados pré-processados devem passar por outra transformação, que os armazena adequadamente, visando facilitar o uso das técnicas de *Data Mining*. Nessa fase, o uso de *Data Warehouses* expande-se consideravelmente, já que, nessas estruturas, as informações estão alocadas da maneira mais eficiente. Addrians & Zantinge (1996) definem *Data Warehouse* como um depósito central de dados, extraído de dados operacionais, em que a informação é orientada a assuntos, não volátil e de natureza histórica. Devido a essas características, *Data Warehouses* tendem a se tornar grandes repositórios de dados extremamente organizados, facilitando a aplicação do *Data Mining*.

Prosseguindo no processo KDD, chega-se especificamente à fase de *Data Mining*. O objetivo principal desse passo é a aplicação de técnicas de mineração nos dados pré-processados, o que envolve ajuste de modelos e/ou determinação de características nos dados. Em outras palavras, exige o uso de métodos inteligentes para a extração de padrões ou conhecimentos dos dados.

É importante destacar que cada técnica de *Data Mining* utilizada para conduzir as operações de Mineração de Dados adapta-se melhor a alguns problemas do que a outros, o que impossibilita a existência de um método de *Data Mining* universalmente melhor. Para cada problema particular, tem-se uma técnica particular. Portanto, o sucesso de uma tarefa de *Data Mining* está diretamente ligado à experiência e à intuição do analista.

A etapa final do processo de mineração consiste no pós-processamento, que engloba a interpretação dos padrões descobertos e a possibilidade de retorno a qualquer um dos passos anteriores. Assim, a informação extraída é analisada (ou interpretada) em relação ao objetivo proposto, sendo identificadas e apresentadas as melhores informações. Dessa forma, o propósito do resultado não consiste somente em visualizar, gráfica ou logicamente, o rendimento do *Data Mining*, mas, também, em filtrar a informação que será apresentada, eliminando possíveis ruídos (ou seja, padrões redundantes ou irrelevantes) que podem surgir no processo. Apresenta-se, a seguir, uma breve caracterização do processo de *Data Mining*.

DATA MINING

Data Mining, ou Mineração de Dados, pode ser entendido como o processo de extração de informações, sem conhecimento prévio, de um grande banco de dados e seu uso para tomada de decisões. É uma metodologia aplicada em diversas áreas que usam o conhecimento, como empresas, indústrias e instituições de pesquisa. *Data Mining* define o processo automatizado de captura e análise de grandes conjuntos de dados para extrair um significado, sendo usado tanto para descrever características do passado como para prever tendências para o futuro.

Para encontrar respostas ou extrair conhecimento interessante, existem diversos métodos de *Data Mining* disponíveis na literatura. Mas, para que a descoberta de conhecimentos seja relevante, é importante estabelecer metas bem definidas. Essas metas são alcançadas por meio dos seguintes métodos de *Data Mining*: Classificação, Modelos de Relacionamento entre Variáveis, Análise de Agrupamento, Sumarização, Modelo de Dependência, Regras de Associação e Análise de Séries Temporais, conforme citação e definição feita por Fayyad *et al.* (1996a). É importante ressaltar que a maioria desses métodos é baseada em técnicas das áreas de aprendizado de máquina, reconhecimento de padrões e estatística. Essas técnicas vão desde as tradicionais da estatística multivariada, como análise de agrupamentos e regressões, até modelos mais atuais de aprendizagem, como redes neurais, lógica difusa e algoritmos genéticos.

Os métodos tradicionais de *Data Mining* são:

- **Classificação:** associa ou classifica um item a uma ou várias classes categóricas pré-definidas. Uma técnica estatística apropriada para classificação é a análise discriminante. Os objetivos dessa técnica envolvem a descrição gráfica ou algébrica das características diferenciais das observações de várias populações, além da classificação das observações em uma ou mais classes predeterminadas. A ideia é derivar uma regra que possa ser usada para classificar, de forma otimizada, uma nova observação a uma classe já rotulada. Segundo Mattar (1998), a análise discriminante permite que dois ou mais grupos possam ser comparados, com o objetivo de determinar se diferem uns dos outros e, também, a natureza da diferença, de forma que, com base em um conjunto de variáveis independentes, seja possível classificar indivíduos ou objetos em duas ou mais categorias mutuamente exclusivas.
- **Modelos de Relacionamento entre Variáveis:** associa um item a uma ou mais variáveis de predição de valores reais, consideradas variáveis independentes ou exploratórias. Técnicas estatísticas como regressão linear simples, múltipla e modelos lineares por transformação são utilizadas para verificar o relacionamento funcional que, eventualmente, possa existir entre duas variáveis quantitativas, ou seja, constatar se há uma relação funcional entre X e Y. Observa-se, conforme Gujarati (2000), que o método dos mínimos quadrados ordinários, atribuído a Carl Friedrich Gauss, tem propriedades estatísticas relevantes e apropriadas, que tornaram tal procedimento um dos mais poderosos e populares métodos de análise de regressão.
- **Análise de Agrupamento (Cluster):** associa um item a uma ou várias classes categóricas (ou *clusters*), em que as classes são determinadas pelos dados, diversamente da classificação em que as classes são pré-definidas. Os *clusters* são definidos por meio do agrupamento de dados baseados em medidas de similaridade ou modelos probabilísticos. A análise de *cluster* (ou agrupamento) é uma técnica que visa detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso de sua existência, determinar quais são eles. Nesse tipo de análise, segundo Pereira (1999), o procedimento inicia com o cálculo das distâncias entre os objetos estudados dentro do espaço multiplano constituído por eixos de todas as medidas realizadas (variáveis), sendo, a seguir, os objetos agrupados conforme a proximidade entre eles. Na sequência, efetuam-se os agrupamentos por proximidade geométrica, o que permite o reconhecimento dos passos de agrupamento para a correta identificação de grupos dentro do universo dos objetos estudados.
- **Sumarização:** determina uma descrição compacta para um dado subconjunto. As medidas de posição e variabilidade são exemplos simples de sumarização. Funções mais sofisticadas envolvem técnicas de visualização e a determinação de relações funcionais entre variáveis. As funções de sumarização são fre-

qüentemente usadas na análise exploratória de dados com geração automatizada de relatórios, sendo responsáveis pela descrição compacta de um conjunto de dados. A sumarização é utilizada, principalmente, no pré-processamento dos dados, quando valores inválidos são determinados por meio do cálculo de medidas estatísticas – como mínimo, máximo, média, moda, mediana e desvio padrão amostral –, no caso de variáveis quantitativas, e, no caso de variáveis categóricas, por meio da distribuição de frequência dos valores. Técnicas de sumarização mais sofisticadas são chamadas de visualização, que são de extrema importância e imprescindíveis para se obter um entendimento, muitas vezes intuitivo, do conjunto de dados. Exemplos de técnicas de visualização de dados incluem diagramas baseados em proporções, diagramas de dispersão, histogramas e *box plots*, entre outros. Autores como Levine *et al.* (2000) e Martins (2001), entre outros, abordam com grande detalhamento esses procedimentos metodológicos.

- **Modelo de Dependência:** descreve dependências significativas entre variáveis. Modelos de dependência existem em dois níveis: estruturado e quantitativo. O nível estruturado especifica, geralmente em forma de gráfico, quais variáveis são localmente dependentes. O nível quantitativo especifica o grau de dependência, usando alguma escala numérica. Segundo Padovani (2000), análises de dependência são aquelas que têm por objetivo o estudo da dependência de uma ou mais variáveis em relação a outras, sendo procedimentos metodológicos para tanto a análise discriminante, a de medidas repetidas, a de correlação canônica, a de regressão multivariada e a de variância multivariada.
- **Regras de Associação:** determinam relações entre campos de um banco de dados. A idéia é a derivação de correlações multivariadas que permitam subsidiar as tomadas de decisão. A busca de associação entre variáveis é, freqüentemente, um dos propósitos das pesquisas empíricas. A possível existência de relação entre variáveis orienta análises, conclusões e evidencição de achados da investigação. Uma regra de associação é definida como *se X então Y*, ou $X \Rightarrow Y$, onde X e Y são conjuntos de itens e $X \cap Y = \emptyset$. Diz-se que X é o antecedente da regra, enquanto Y é o seu conseqüente. Medidas estatísticas como correlação e testes de hipóteses apropriados revelam a freqüência de uma regra no universo dos dados minerados. Vários métodos para medir associação são discutidos por Mattar (1998), de natureza paramétrica e não-paramétrica, considerando a escala de mensuração das variáveis.
- **Análise de Séries Temporais:** determina características sequenciais, como dados com dependência no tempo. Seu objetivo é modelar o estado do processo extraindo e registrando desvios e tendências no tempo. Correlações entre dois instantes de tempo, ou seja, as observações de interesse, são obtidas em instantes sucessivos de tempo – por exemplo, a cada hora, durante 24 horas – ou são registradas por algum equipamento de forma contínua, como um traçado eletrocardiográfico. As séries são compostas por quatro padrões: tendência, variações cíclicas, variações sazonais e variações irregulares. Há vários modelos estatísticos que podem ser aplicados a essas situações, desde os de regressão linear (simples e múltiplos), os lineares por transformação e regressões assintóticas, além de modelos com defasagem, como os autoregressivos (AR) e outros deles derivados. Uma interessante noção introdutória ao estudo de séries temporais é desenvolvida por Morettin & Toloi (1987).

Diante da descrição sumária de metodologias estatísticas aplicáveis ao procedimento de Mineração de Dados, registra-se que, embora Hand (1998) afirme que o termo *Data Mining* possa trazer uma conotação simplista para os estatísticos, Fayyad *et al.* (1996a) mostraram a relevância da estatística para o processo de extração de conhecimentos, ao afirmar que essa ciência provê uma linguagem e uma estrutura para quantificar a incerteza resultante quando se tenta deduzir padrões de uma amostra a partir de uma população.

De acordo com Hand (1998), a estatística preocupa-se com a análise primária dos dados, no sentido de que eles são coletados por uma razão particular ou por um conjunto de questões particulares *a priori*. *Data Mining*, por outro lado, preocupa-se também com a análise secundária dos dados, num sentido mais amplo e mais indutivo do que uma abordagem hipotético-dedutiva, freqüentemente considerada como o paradigma para o progresso da ciência moderna. Assim, *Data Mining* pode ser visto como o descendente direto da estatística, já que são técnicas metodológicas complementares.

CLEMENTINE: UMA FERRAMENTA DE DATA MINING

Nos itens anteriores deste artigo, foram apresentados conceitos e técnicas para descoberta de conhecimentos em banco de dados. Esta sessão tem como objetivo discutir a utilização de uma ferramenta de *Data Mining* e analisar sua aplicação em uma base de dados de natureza econômica. A ferramenta em questão é o *Clementine*, da empresa SPSS Inc.

Todos os passos do processo de descoberta de conhecimento podem ser realizados pelo *Clementine*. No entanto, segundo o manual do usuário (*Clementine Users Guide*, 2001), a metodologia indicada para ser usada em conjunto com a ferramenta é o modelo CRISP-DM (*Cross-Industry Standard Process for Data Mining*), que foi desenvolvido a partir da experiência de três empresas pioneiras no setor: a DaimlerChrysler, que aplica análises de *Data Mining* em seus negócios desde 1996; a NCR, que provê soluções de *Data Warehouse*; e a SPSS, que disponibiliza soluções baseadas no processo de mineração de dados desde 1990. Essa metodologia é composta por seis fases, como ilustrado na figura 2.

Fig. 2. O modelo CRISP-Data Mining (CRISP-DM, 2001).



Como pode ser observada na figura 2, a sequência das fases desse processo não é rígida. Voltar e ir avante entre as diferentes fases é sempre necessário. Dessa forma, uma fase depende do resultado de outra, ou da tarefa particular de uma fase que precisa ser executada na próxima etapa. O círculo externo simboliza a natureza cíclica do processo de *Data Mining*. As fases desse processo são:

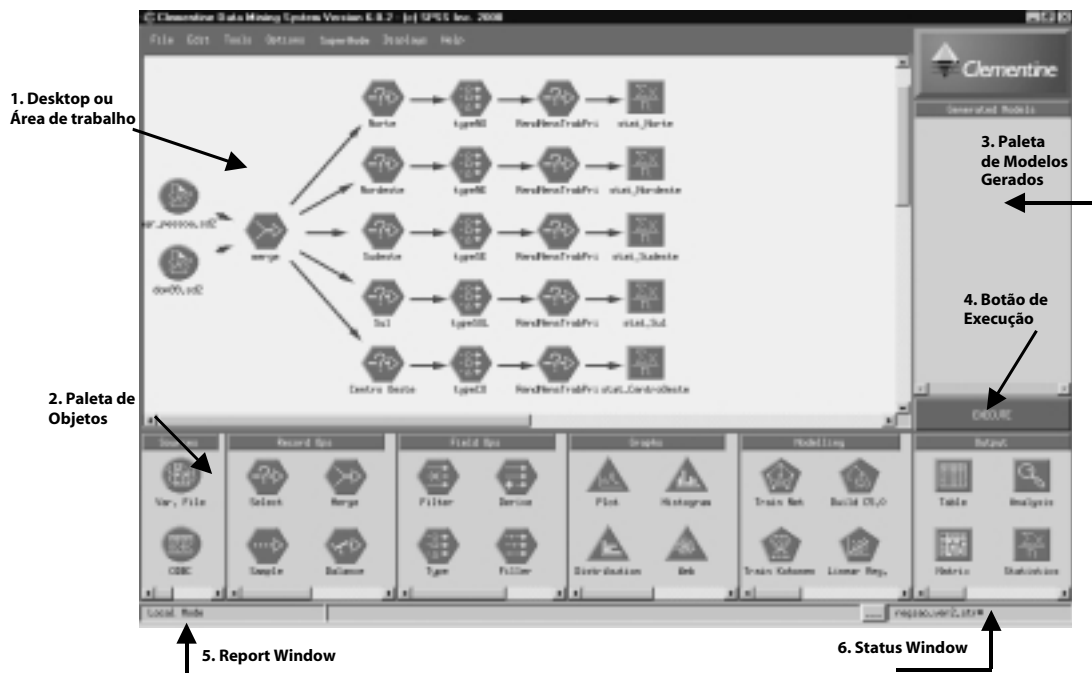
- *Entendimento do Negócio (Business Understanding)*: visa o entendimento dos objetivos e requisitos do projeto, do ponto de vista do negócio. Baseado no conhecimento adquirido, o problema de mineração de dados é definido e um plano preliminar é projetado para alcançar os objetivos.
- *Entendimento dos Dados (Data Understanding)*: inicia com uma coleção de dados e prossegue com atividades que visam buscar familiaridade, identificar problemas de qualidade, descobrir os primeiros discernimentos nos dados ou detectar subconjuntos interessantes para formar hipóteses da informação escondida.

- *Preparação dos Dados (Data Preparation)*: cobre todas as atividades de construção do *dataset* final. As tarefas de preparação de dados são, provavelmente, desempenhadas várias vezes e sem qualquer ordem prescrita. Essas tarefas incluem a seleção de tabelas, registros e atributos, bem como a transformação e limpeza dos dados para as ferramentas de modelagem.
- *Modelagem (Modelling)*: várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são ajustados para valores ótimos. Geralmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas delas têm requisitos específicos na formação de dados. Portanto, retornar à fase de preparação de dados é frequentemente necessário.
- *Avaliação (Evaluation)*: o modelo (ou modelos) construído na fase anterior é avaliado e os passos executados na sua construção são revistos, para se certificar que o modelo representa os objetivos do negócio. Seu principal objetivo é determinar se existe alguma questão de negócio importante que não foi suficientemente considerada. Nesta fase, uma decisão sobre o uso dos resultados de mineração de dados deverá ser obtida.
- *Utilização ou Aplicação (Deployment)*: após a construção e avaliação do modelo (ou modelos), ele pode ser utilizado de duas formas: em uma, o analista pode recomendar ações a serem tomadas baseando-se, simplesmente, na visão do modelo e de seus resultados; na outra, o modelo pode ser aplicado a diferentes conjuntos de dados.

A Interface do Clementine

O *Clementine* possui uma interface de programação visual que facilita a construção de modelos de *Data Mining* para o processo de descoberta de conhecimento. A ferramenta oferece ricas facilidades para a exploração e manipulação de dados, além de várias técnicas de modelagem e recursos gráficos, para a visualização de dados. As operações são representadas em uma área de trabalho por nós (*nodes*) que, conectados, formam o fluxo de dados, chamado de *streams*, conforme ilustra a figura 3.

Fig. 3. A Interface de Programação Visual do *Clementine*.



A área de trabalho, ou *desktop*, também chamada de *stream pane*, é a área de construção e manipulação dos *streams* e dados. Em outras palavras, é a área de construção do modelo de *Data Mining*.

Os nós apresentam-se agrupados de acordo com seu tipo de funcionalidade na paleta de objetos localizada na parte inferior da área de trabalho do *Clementine*, que pode ser acessado e do qual é possível importar dados, por meio das funcionalidades dos nós do grupo *Source*; manipular registros e campos, através do grupo *Record Ops*; visualizar os dados a partir de gráficos contidos nos diversos nós do grupo *Graphs*; construir modelos por meio de uma variedade de técnicas de modelagem disponíveis no grupo *Modelling*; e avaliar os resultados com os recursos do grupo *Output*.

A paleta de modelos gerados, localizada à direita do leitor na área de trabalho, contém os resultados de um modelo construído depois de executado. Para executar um modelo, basta clicar no botão de execução, abaixo da paleta de modelos gerados. Ao se clicar no botão, todos os *streams* válidos são executados. O *Report Window*, localizado abaixo da paleta de objetos, provê um *feedback* do progresso de várias operações, tal como quando os dados estão sendo lidos. O *Status Window*, também abaixo da paleta de objetos, provê informação sobre o que a aplicação está realizando no momento, bem como mensagens de pedido de retorno do usuário.

Aplicação

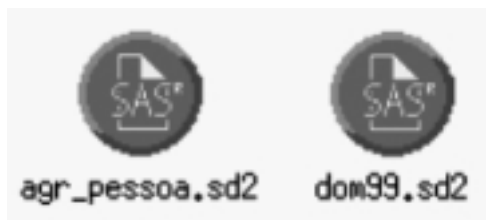
Visamos aqui mostrar algumas aplicações realizadas com a ferramenta em estudo. É importante destacar que o objetivo é mostrar como o *Data Mining* pode ser aplicado por meio dessa ferramenta e, não, fazer *marketing* do *Clementine*.

Os dados utilizados para essa pesquisa foram fornecidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE), em CD (microdados), e referem-se à Pesquisa Nacional por Amostra de Domicílios (PNAD) do ano de 1999. É importante lembrar que essa massa de dados tem sua origem temporal no sistema de pesquisas domiciliares, que foi implantado progressivamente no Brasil a partir de 1967, com a criação da Pesquisa Nacional por Amostra de Domicílios, e tem por finalidade a produção de informações básicas para o estudo do desenvolvimento socioeconômico do País (PNAD, 1999). Os dados encontram-se organizados e disponíveis em CD-ROM, divididos em dois arquivos: um referente a pessoas e o outro, a domicílios.

O arquivo de pessoas contém, em síntese, informações sobre a identificação dos moradores, suas características gerais, educação, trabalho e rendimento, entre outras. O arquivo de domicílios possui, em resumo, informações sobre características da unidade domiciliar. Os dados do presente estudo referem-se a pessoas ocupadas em atividades agrícolas. As variáveis selecionadas para essa aplicação são: UF – Unidade da Federação; V0302-Sexo; V8005-Idade; V0404-Cor ou Raça; V0601-Sabe ler e escrever; V4703-Anos de Estudo; e V4614-Rendimento Mensal Domiciliar.

O primeiro passo realizado para essa aplicação foi carregar os arquivos de Pessoas e Domicílios. Para isso, utilizou-se o nó SAS¹ da paleta *Source* (fig. 4). Em seguida, foi preciso ligar os arquivos, por meio do nó *Merge*, já que são usadas variáveis dos dois arquivos. Em seguida, os dados foram separados em regiões, para se ter clareza dos resultados em relação à Unidade da Federação a que pertencem. Para isso, foi utilizado o nó *Select*, conforme ilustra a figura 5.

Fig. 4. Nó SAS.



¹ O SAS (*Statistical Analysis System*) é um pacote para análises estatísticas compatível para aplicações no *Clementine* – SPSS.

Fig. 5. Nó *Select* para divisão em regiões.

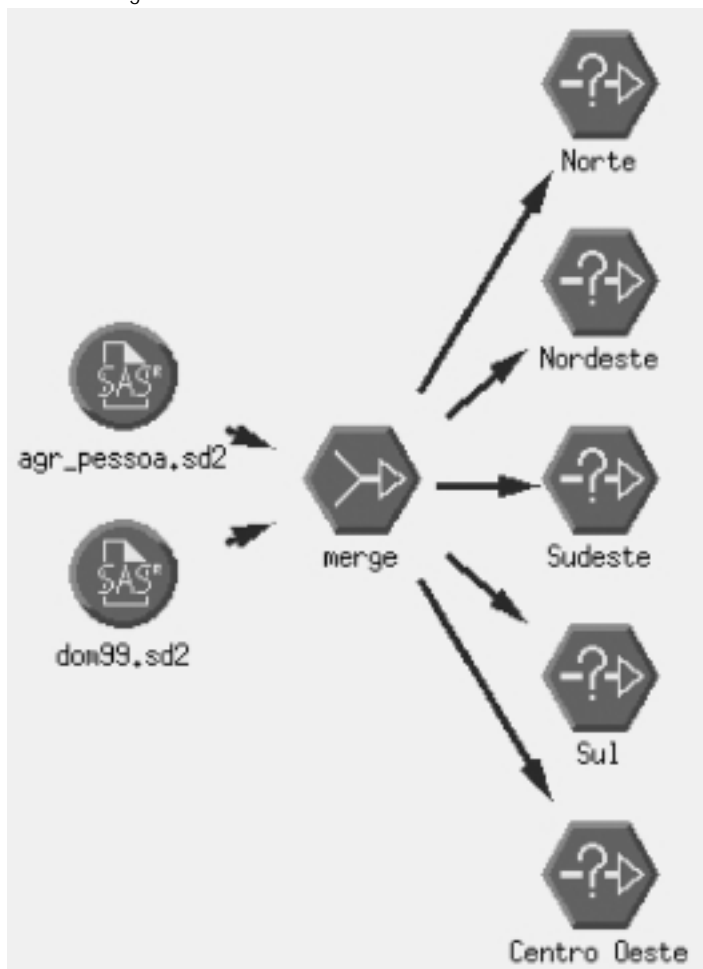
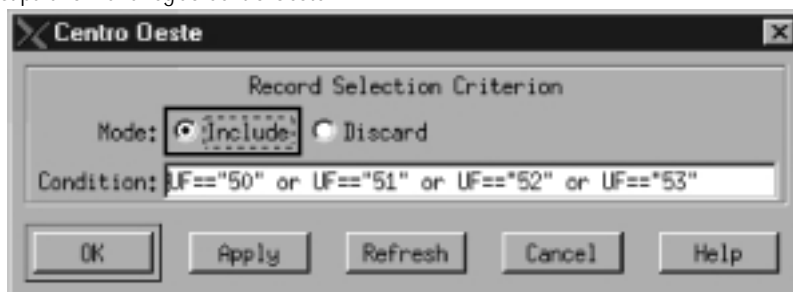


Fig. 6. Nó *Select* para formar a região Centro-Oeste.



Para cada nó *Select* utilizado, foi selecionado o modo *Include* para incluir o resultado da condição (*Condition*) que seleciona a Unidade da Federação (UF) correspondente a cada região, caso ela seja verdadeira. A figura 6 mostra as condições usadas para formar a região Centro-Oeste, constituída pelas seguintes UFs: 50. Mato Grosso do Sul; 51. Mato Grosso; 52. Goiás; e 53. Distrito Federal.

Ao iniciar uma aplicação com o *Clementine*, não é preciso, necessariamente, saber o que se está procurando. É possível explorar os dados investigando diferentes relacionamentos até encontrar informações úteis. Desse modo, para melhor entendimento das informações, foi realizada uma análise exploratória de dados usando o nó *Statistics* da paleta *Output* (veja quadro correspondente).

Análise exploratória de dados, apenas com nó *Statistics* da paleta *Output*: resultados.

NORTE

STATISTICS FOR FIELD: RENDMENSALDOM

OCCURRENCES	=	1486
MINIMUM	=	0.0000
MAXIMUM	=	9612.0
RANGE	=	9612.0
MEAN	=	522.59
STANDARD DEVIATION	=	807.40
STANDARD ERROR OF THE MEAN	=	20.945
VARIANCE	=	651887.5
MEDIAN	=	331.00
SUM	=	776570.0

NORDESTE

STATISTICS FOR FIELD: RENDMENSALDOM

OCCURRENCES	=	14745
MINIMUM	=	0.0000
MAXIMUM	=	9936.0
RANGE	=	9936.0
MEAN	=	311.06
STANDARD DEVIATION	=	353.92
STANDARD ERROR OF THE MEAN	=	2.9146
VARIANCE	=	125259.0
MEDIAN	=	236.00
SUM	=	4586630.0

SUL

STATISTICS FOR FIELD: RENDMENSALDOM

OCCURRENCES	=	5776
MINIMUM	=	0.0000
MAXIMUM	=	17300.0
RANGE	=	17300.0
MEAN	=	633.63
STANDARD DEVIATION	=	913.36
STANDARD ERROR OF THE MEAN	=	12.018
VARIANCE	=	834232.6
MEDIAN	=	408.00
SUM	=	3659861.0

SUDESTE

STATISTICS FOR FIELD: RENDMENSALDOM

OCCURRENCES	=	6927
MINIMUM	=	0.0000
MAXIMUM	=	18000.0
RANGE	=	18000.0
MEAN	=	583.94
STANDARD DEVIATION	=	807.79
STANDARD ERROR OF THE MEAN	=	9.7057
VARIANCE	=	652529.9
MEDIAN	=	380.00
SUM	=	4044930.0

CENTRO-OESTE
 STATISTICS FOR FIELD: RENDMENSALDOM

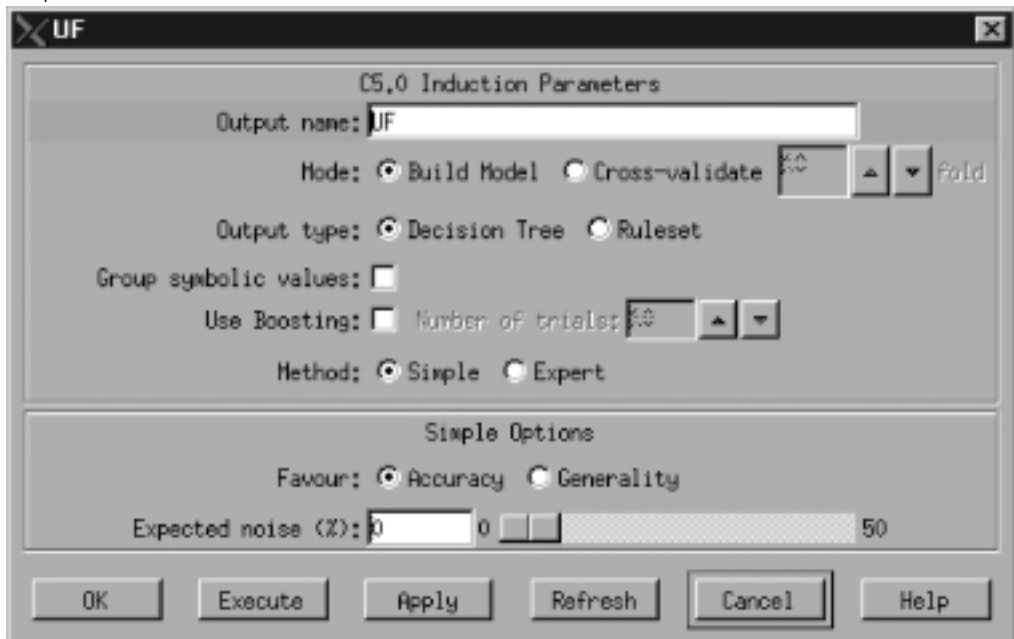
OCCURRENCES	=	3521
MINIMUM	=	0.0000
MAXIMUM	=	36700.0
RANGE	=	36700.0
MEAN	=	624.40
STANDARD DEVIATION	=	1242.1
STANDARD ERROR OF THE MEAN	=	20.932
VARIANCE	=	1542758.2
MEDIAN	=	350.00
SUM	=	2198522.0

Várias medidas estatísticas – como contagem, média, mínimo, máximo, amplitude, desvio padrão, variância, soma e erro padrão da média – foram obtidas a partir do nó *Statistics* para a variável Renda Mensal Domiciliar associada às pessoas ocupadas na agricultura em 1999, de forma a subsidiar uma análise exploratória do comportamento dessa variável nas grandes regiões geográficas do país.

A figura 7 apresenta a utilização do nó *Build C5.0* para a Região Centro-Oeste. Esse nó utiliza o algoritmo C5.0² para construir uma árvore de decisão ou um conjunto de regras (*ruleset*). Todas as opções selecionadas para aplicação desse nó são ilustradas pela referida figura.

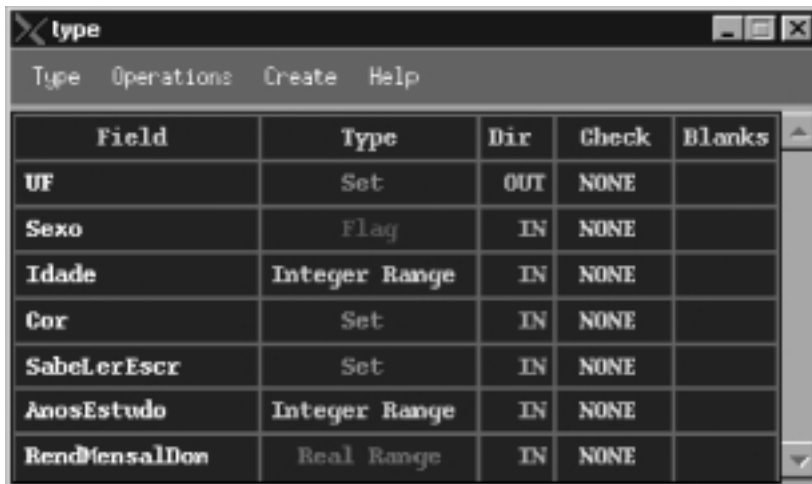
Para gerar uma árvore de decisão através do nó *Build C5.0*, são necessários uma ou mais variáveis de entrada (*In*) e apenas um campo de saída (*Out*). Para selecionar essas variáveis foi usado o nó *Type*, conforme ilustra a figura 8. O resultado obtido é apresentado na figura 9.

Fig. 7. Opções do nó *Build C5.0*.



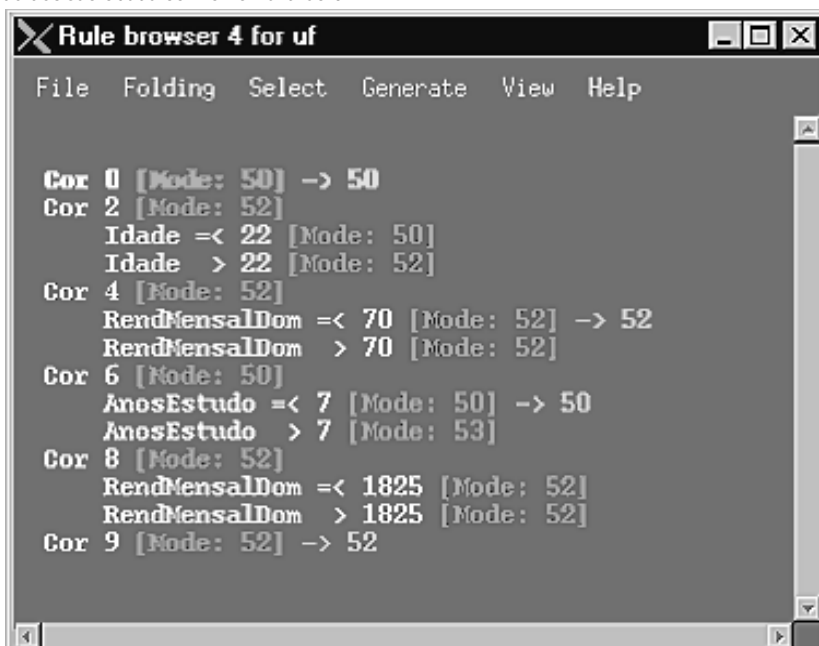
² O algoritmo de indução de regras capaz de produzir árvores de decisão compactas – *rulesets*, conjunto de regras. A versão anterior desse algoritmo foi chamada de C4.5 (*Clementine Users Guide*).

Fig. 8. Nó *Type*.



Field	Type	Dir	Check	Blanks
UF	Set	OUT	NONE	
Sexo	Flag	IN	NONE	
Idade	Integer Range	IN	NONE	
Cor	Set	IN	NONE	
SabeLetEscr	Set	IN	NONE	
AnosEstudo	Integer Range	IN	NONE	
RendMensalDom	Real Range	IN	NONE	

Fig. 9. Árvore de decisão obtida com o nó *Build C5.0*.



```

Cor 0 [Mode: 50] -> 50
Cor 2 [Mode: 52]
  Idade =< 22 [Mode: 50]
  Idade > 22 [Mode: 52]
Cor 4 [Mode: 52]
  RendMensalDom =< 70 [Mode: 52] -> 52
  RendMensalDom > 70 [Mode: 52]
Cor 6 [Mode: 50]
  AnosEstudo =< 7 [Mode: 50] -> 50
  AnosEstudo > 7 [Mode: 53]
Cor 8 [Mode: 52]
  RendMensalDom =< 1825 [Mode: 52]
  RendMensalDom > 1825 [Mode: 52]
Cor 9 [Mode: 52] -> 52
  
```

A figura 9 apresenta informações sobre o número de observações usadas para gerar as ramificações da árvore de decisão, bem como os níveis de certeza. Analisando os resultados obtidos pela árvore, é possível visualizar a classificação da variável de saída (ou seja, variável de predição) Unidade da Federação (UF) na região Centro-Oeste. Para melhor entendimento dos resultados obtidos, apresenta-se, a seguir, uma explicação do funcionamento das árvores de decisão. É importante ressaltar que esse exemplo é apenas ilustrativo e seu objetivo é somente explicar um resultado obtido com a ferramenta de estudo.

Entendendo as Árvores de Decisão

Ao navegar pelos nós de uma árvore de decisão, é possível verificar uma lista de condições que definem a divisão dos dados que foram descobertos pelo algoritmo no *Clementine*.

As árvores de decisão funcionam/trabalham recursivamente, dividindo os dados com base nos valores dos campos de entrada. Os dados que foram divididos são denominados ramo, ou galho. O galho inicial (também denominado raiz) engloba todos os registros. A raiz é dividida em subconjuntos, ou galhos filhos, baseados no valor de um particular campo de entrada. Cada galho filho pode ser dividido, mais de uma vez, em subgalhos, que podem ser divididos novamente, e assim por diante. No nível mais baixo da árvore, encontram-se os galhos que não podem mais ser divididos, conhecidos como galhos terminais, ou folha.

O navegador da árvore de decisão mostra os valores de entrada que definem cada divisão, ou galho, e um resumo do campo (ou variável) de saída para os registros da divisão. Para divisões baseadas em campos numéricos, o galho é mostrado por uma linha, na forma:

nome_do_campo relação valor [resumo]

em que a relação é uma relação numérica. Por exemplo, um galho definido por valores maiores que 22 para a variável Idade aparecerá como:

idade = < 22 [resumo]

Para divisões baseadas em campos simbólicos, o galho é mostrado da seguinte forma:

nome_do_campo valor [resumo]

ou

nome_do_campo [valores] [resumo]

em que os valores são os da variável que define o galho. Por exemplo, um galho que inclui registros onde o valor da variável Cor ou Raça³ pode ser 0 (índigena), 2 (branca) ou 4 (preta), deve ser representado como:

cor 2 [resumo]

ou

cor ['0','2','4'] [resumo]

Para galhos terminais, uma predição é também dada adicionando-se uma seta e o valor que foi previsto para o final da condição da regra. Por exemplo, uma folha definida por *Anos de Estudo* = < 7, que prediz um valor 7 para o campo de saída, a árvore mostrará:

AnosEstudo = < 7 [mode: 50] -> 50

O resumo para o galho é definido diferentemente para campos de saída simbólica e numérica. Para árvores com campos de saída numérica, o resumo é o valor médio para o galho, e o efeito do galho, definido como a diferença entre a média deste e a média para seus pais. Para árvores com campos de saída simbólica, o resumo é a moda, ou seja, o valor mais freqüente para os registros no galho.

Após entendimento do funcionamento da árvore de decisão gerada por meio do *Clementine*, pode-se chegar às seguintes conclusões ou regras:

- 1) se a variável Cor ou Raça for igual a indígena (0), a variável UF é Mato Grosso do Sul;
- 2) se a variável Cor ou Raça for branca (2), a variável UF é Goiás; porém, se a idade for menor ou igual a 22 anos, a variável UF é Mato Grosso do Sul; e se a idade for maior que 22 anos, a variável UF é Goiás;
- 3) se a variável Cor ou Raça for Preta (4), a variável UF é Mato Grosso do Sul;
- 4) se a variável Cor ou Raça for amarela (6), a variável UF é Mato Grosso do Sul; porém, se a variável Anos de Estudo for menor ou igual a sete anos, a UF é Mato Grosso do Sul; e se a variável Anos de Estudo for maior que sete anos, a variável UF é Distrito Federal;
- 5) se a variável Cor ou Raça for Parda (8), a variável UF é Goiás; e,
- 6) se a variável Cor ou Raça for ignorada, a UF também é Goiás.

³ A definição da variável como *Cor ou Raça* segue a classificação padrão da PNAD-IBGE, assim como suas categorias, definidas como branca, preta, amarela, parda e indígena (conforme dicionário de dados, PNAD 1998, cd-ROM).

Em síntese, de acordo com as regras mencionadas e com a figura 9, a UF da Região Centro-Oeste com maior diversidade de pessoas de Cor ou Raça diferente é Goiás (UF igual a 52).

CONSIDERAÇÕES FINAIS

A idéia central deste artigo foi a de apresentar, de forma sucinta, os principais conceitos e técnicas envolvidos na nova área interdisciplinar *Data Mining*. Além desses conceitos e técnicas, foram apresentadas as características de um software específico para mineração de dados, o *Clementine*, da SPSS.

Destaca-se *Data Mining* como parte de um processo maior, denominado KDD, e que se refere ao meio pelo qual padrões são extraídos e enumerados a partir dos dados, ou seja, ao uso de métodos inteligentes para se extrair novos conhecimentos. Entendendo-se por métodos inteligentes a aplicação de alguma técnica específica de *Data Mining*, neste artigo foi destacada a utilização de classificação por meio de árvores de decisão, com o apoio do software *Clementine*. Constata-se que todos os passos do processo de descoberta de conhecimento podem ser realizados pelo *Clementine*. Uma grande vantagem dessa ferramenta é sua interface de programação visual, o que favorece a construção de modelos de *Data Mining* para o processo de descoberta de conhecimento e ainda oferece ricas facilidades para exploração e manipulação de dados, além de várias técnicas de modelagem e recursos gráficos para visualização de dados.

A utilização de um software de *Data Mining* pode trazer descobertas inovadoras para estudiosos da área econômica. Porém, deve ficar claro que nenhuma ferramenta de *Data Mining* trabalha por si só e elimina a necessidade de conhecimento e entendimento do negócio e a compreensão dos dados a serem minerados, nem mesmo substitui analistas e pesquisadores da área (ou gestores de negócios). Mas deve ficar claro que o uso da ferramenta proporciona aos usuários meios para encontrar tesouros de informações que permitam detectar tendências e características disfarçadas, confirmar a necessidade de estudos de novas relações, não necessariamente previstas pela teoria econômica ou que sejam indicativas de temas a serem pesquisados, ou, ainda, reagir rapidamente a um evento que ainda pode estar por vir.

REFERÊNCIAS BIBLIOGRÁFICAS

- ADDRIANS, P. & ZANTINGE, D. *Data Mining*. Inglaterra: Addison-Wesley, 1996.
- BRACHNAD, R.J. & ANAND, T. The process of knowledge discovery in databases. In: FAYYAD, U.M. et al. *Advances in Knowledge Discovery in Data Mining*. Menlo Park: AAAI Press, 1996.
- CLEMENTINE® 6.0 User's Guide, Copyright © 2001 by SPSS Inc. Printed in the United States of America, 2001.
- CRISP-DM: *Cross Industry Standard Process Model for Data Mining*. . Printed in the United States of America, 2001.
- DILLY, R. *Data Mining: an introduction*. Belfast: Parallel Computer Centre, Queens University, 1999.
- DINIZ, C.A. & LOUZADA-NETO, F. *Data Mining: uma introdução*. São Carlos: Associação Brasileira de Estatística, 2000.
- FAYYAD, U.M. et al. The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: _____. *Advances in Knowledge Discovery in Data Mining*. Menlo Park: AAAI Press, 1996a.
- FAYYAD, U.M. et al. *Advances in Knowledge Discovery and Data Mining*. California: AAAI Press, 1996b.
- GUJARATI, D.N. *Econometria Básica*. Trad. Ernesto Yoshita. São Paulo: Makron Books, 2000.
- HAND, D.J. Data Mining: statistics and more? *The American Statistician*, England, 52 (2): 112-118, mai./98.
- IBGE. *Pesquisa Nacional por Amstras de Domicílios 1999*. Rio de Janeiro: IBGE, 1999. CD-rom.
- LEVINE, D.M. et al. *Estatística: teoria e aplicações*. Trad. Teresa C.P. de Souza. Rio de Janeiro: LTC Editora, 2000.
- MANNILA, H. Data mining: machine learning, statistics and databases. *International Conference on Statistics and Scientific Database Management*, Estocolmo, 8, 1996.
- MARTINS, G.A. *Estatística Geral e Aplicada*. São Paulo: Atlas, 2001.
- MATTAR, F.N. *Pesquisa de Marketing*. São Paulo: Atlas, 1998.
- MORETTIN, P.A. & TOLOI, C.M. *Séries Temporais*. 2.^a ed. São Paulo: Atual, 1987.
- PADOVANI, C.R. *Estatística na Metodologia da Investigação Científica*. Botucatu: UNESP, 1995.
- PEREIRA, J.C.R. *Análise de Dados Qualitativos*. São Paulo: Edusp/Fapesp, 1999.
- SADE, A.S. & SOUZA, J.M. *Prospecção de Conhecimento em Bases de Dados Ambientais*. Rio de Janeiro: UFRJ, 1996.

Dados dos autores

HELOISA HELENA SFERRA
Analista de Sistemas. Mestranda em Ciência de
Computação pela UNIMEP

ÂNGELA M. C. JORGE CORRÊA
Professora doutora do Grupo de Área em
Métodos Quantitativos (FCMNTI/UNIMEP) do
Mestrado em Administração e convidada do
Mestrado em Ciência da Computação/UNIMEP

Recebimento do artigo: 19/mar./03
Consultoria: 24/mar./03 a 18/dez./03
Aprovado: 18/dez./03

