UNIVERSIDADE FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM MÉTODOS NUMÉRICOS EM ENGENHARIA
Área de Concentração: Programação Matemática

MARCO AURÉLIO SILVA NETO

MINERAÇÃO VISUAL DE DADOS: EXTRAÇÃO DO CONHECIMENTO A PARTIR DAS TÉCNICAS DE VISUALIZAÇÃO DA INFORMAÇÃO E MINERAÇÃO DE DADOS

Experimentos: ITAIPU e SIMEPAR

DISSERTAÇÃO DE MESTRADO

Curitiba
Março/2008

MARCO AURÉLIO SILVA NETO

MINERAÇÃO VISUAL DE DADOS: EXTRAÇÃO DO CONHECIMENTO A PARTIR DAS TÉCNICAS DE VISUALIZAÇÃO DA INFORMAÇÃO E MINERAÇÃO DE DADOS

Experimentos: ITAIPU e SIMEPAR

Dissertação apresentada ao Programa de Pós-Graduação em Métodos Numéricos em Engenharia (PPGMNE) da Universidade Federal do Paraná (UFPR), como parte dos requisitos para obtenção do título de Mestre em Ciências na área de concentração Programação Matemática.

Orientador: Prof. Dr. Sérgio Scheer

Curitiba
Março/2008

TERMO DE APROVAÇÃO

MARCO AURÉLIO SILVA NETO

MINERAÇÃO VISUAL DE DADOS: EXTRAÇÃO DO CONHECIMENTO A PARTIR DAS TÉCNICAS DE VISUALIZAÇÃO DA INFORMAÇÃO E MINERAÇÃO DE DADOS

Estudos de Casos: ITAIPU e SIMEPAR

Dissertação aprovada como requisito parcial para obtenção do título de Mestre em Ciências, na área de concentração Programação Matemática, do Programa de Pós-Graduação em Métodos Numéricos em Engenharia (PPGMNE) da Universidade Federal do Paraná (UFPR), pela comissão formada pelos professores:

Aprovada em 06 de março de 2008.

Comissão examinadora:

Prof. Dr. Sérgio Scheer
Universidade Federal do Paraná - UFPR
Orientador

Profª. Drª. Maria Teresinha Arns Steiner
Universidade Federal do Paraná - UFPR

Prof. Dr. Hélio Pedrini
Universidade Federal do Paraná – UFPR

Profª. Drª. Cinthia Obladen de Almendra Freitas
Pontifícia Universidade Católica do Paraná - PUCPR

Este trabalho é dedicado a todas as pessoas que me acompanharam ao longo destes anos e que comigo compartilharam as dúvidas, as certezas, as alegrias e as frustrações que nos cercaram ao longo deste período. Em especial dedico a meus Pais Suely e Carlos e minha esposa Margarete que sempre me encorajaram a ir adiante, mesmo nos momentos de dificuldade.

Agradecimentos

A obtenção do título de Mestre em Ciências é antes de tudo uma realização pessoal. Ao longo destes dois anos muitas dificuldades foram superadas e novas experiências foram adquiridas. Hoje, tenho certeza que este título só foi possível graças à ajuda daqueles que fizeram com que eu acreditasse que poderia chegar até aqui.

Meu primeiro agradecimento é para Margarete, minha esposa, que pacientemente soube aceitar os meus picos de mau-humor e às longas horas de estudos madrugadas adentro. Obrigado pelo apoio, incentivo e confiança.

Agradeço a meus pais, Suely Maria Silva e Carlos Eduardo Neto, e a meus irmãos Priscilla e Carlos, que me ajudaram nos momentos que mais precisei e que nunca deixaram de acreditar em meu potencial. Obrigado pelo companheirismo nesta jornada.

Ao professor Sérgio Scheer, meu orientador, por me introduzir na área de Visualização e me guiar em todos os passos. Minha eterna gratidão.

Aos professores do PPGMNE, especialmente à Maria Teresinha Arns Steiner, Celso Carnieri, Arinei Carlos Lindback da Silva, Neida Maria Patias Volpi, Sérgio Scheer, Liliana Madalena Gramani Cumin, Anselmo Chaves Neto e Klaus de Geus, por contribuírem com minha formação.

À professora Andrea Sell Dyminski por sua dedicação e colaboração junto ao projeto ITAIPU e pela ajuda fornecida na interpretação dos resultados.

Aos meus amigos, que em momentos distintos da vida, me fizeram feliz, e agradecido a Deus por tê-los. Obrigado por aceitarem as minhas ausências em muitas festas realizadas por ter que ficar estudando.

Aos colegas de pós, em especial, Chico (Francisco), Vanessa, Ricardo, Wirllen, Bernadete, Ana Beatriz, Carmem, Marcelo, Cristiane (UTFPR), Michelly (UTFPR) e Vanessa (UTFPR), pelo companheirismo e apoio nas horas difíceis (e nas fáceis também...) e pelos momentos de descontração.

Aos colegas Pablo, Neile, Rosangela, Mayko, membros do projeto ITAIPU. Obrigado por ajudar!

Ao Tiago e funcionários do SIMEPAR pela ajuda durante a pesquisa e desenvolvimento deste trabalho.

Aos funcionários do CESEC, em particular à secretária Maristela, que sempre nos divertia com seu alto índice de humor e que com muita disposição fazia cafezinhos a todos.

À UFPR pela oportunidade.

Ao SIMEPAR pelo apoio financeiro e fornecimento dos dados.

À ITAIPU pelo fornecimento dos dados.

A todos que de alguma forma contribuíram para a realização deste trabalho, levo a mais completa admiração.

Marco Aurélio Silva Neto

"O cansaço físico, mesmo que suportado forçosamente,
não prejudica o corpo, enquanto o conhecimento imposto
à força não pode permanecer na alma por muito tempo".

Platão

RESUMO

Extrair rapidamente informações de grandes conjuntos de dados é, hoje, uma demanda crescente devido ao aumento da capacidade de geração de dados por sensores e outras fontes. A alta dimensionalidade e a grande quantidade de registros contidos nas bases de dados atuais são problemas não triviais na busca e extração de “conhecimento”. O *Processo KDD* (*Knowledge Discovery in Database*), termo criado em 1989, refere-se ao processo de seleção, pré-processamento e transformação de dados, necessário para avaliação e interpretação de resultados pelo uso de técnicas de Mineração de Dados (MD) que, por sua vez, possibilita a extração de padrões “escondidos” nos dados. Por ser uma área científica multidisciplinar, a MD exige o conhecimento em várias outras áreas, incluindo a Visualização e a Estatística. Assim, a Mineração Visual de Dados (MVD) é uma abordagem para integrar a Mineração de Dados com a Visualização. Refere-se à exploração visual de dados fazendo uso de recursos visuais e Computação Gráfica Interativa. Neste trabalho, é apresentado um estudo no qual foram utilizados algoritmos de MVD para analise de dados em dois experimentos. A ITAIPU, maior hidrelétrica em operação do mundo, atualmente possui mais de 2200 instrumentos de auscultação instalados, produzindo dados que vêm sendo armazenados há mais de 30 anos. Neste experimento, a MVD foi utilizada para analisar relações nos instrumentos instalados na estrutura da barragem, permitindo, por exemplo, detectar indesejáveis falhas nas leituras, e conseqüentemente, na sua segurança. No segundo experimento a MVD foi utilizada na filtragem de dados que não representam chuvas em imagens do radar meteorológico do SIMEPAR. Para tanto, um maior número de informações é extraído mais facilmente quando diferentes técnicas de Visualização da Informação (baseadas em Projeções Geométricas, Iconográficas e Orientadas a Pixels) são aplicadas aos dados. Esta análise visual dos dados mostrou-se eficiente por agilizar a detecção de padrões e anomalias nos dados, mostrando-se uma valiosa ferramenta de apoio à tomada de decisões.

Palavras-chave: Processo KDD, Mineração de Dados, Visualização da Informação, Mineração Visual de Dados, Monitoramento de Barragens, Radar Meteorológico.

ABSTRACT

Extracting information quickly from large data sets is now an increasing demand due to increased capacity to generate data for sensors and other sources. The high dimensionality and the large number of records contained in databases are non trivial current problems in the search and extraction of "knowledge". The KDD (Knowledge Discovery in Database), term created in 1989, refers to the process of selection, pre-processing and processing of data, necessary for evaluation and interpretation of results by using Data Mining (DM) techniques. These techniques enable the extraction of "hidden" patterns in data. As a multidisciplinary scientific area, the DM requires knowledge in several other areas, including Visualization and Statistics. Thus, the Visual Data Mining (VDM) is an approach to integrate the Data Mining with Visualization, and refers to the visual exploration of data by making use of interactive computer graphics. This work presents a study where VDM algorithms are used to analyze two experiments data. The active largest hydroelectric power plant in the world is ITAIPU in the frontier of Brazil and Paraquay. Its dam currently has more than 2200 monitoring instruments installed. They are continuously producing data that have been stored for more than 30 years. It served as the first experiment and VDM was used to examine relationships in the installed instruments, allowing, for example, to detect undesirable weaknesses in reading, and consequently, the dam structure safety. In the second experiment, VDM was used in the filtering of data that do not represent rain on the weather radar images of SIMEPAR, the meteorological system of the Parana State in the South of Brazil. Thus, a greater number of information is extracted more easily when different techniques of Information Visualization are applied to filtering radar data. This data visual analysis proved to be efficient by speeding up the pattern and anomally detection in the data. Moreover, they proved to be a valuable tool to support decision-making.

Keywords: KDD Process, Data Mining, Information Visualization, Visual Data Mining, Dam Monitoring, Weather Radar.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE FIGURAS

LISTA DE FIGURAS

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AA	Análise de Agrupamento
AM	Análise Multivariada
ANA	Agência Nacional de Águas
CAPPI	*Constant Plan Position Indicator*
CG	Computação Gráfica
COPEL	Companhia Paranaense de Energia
DM	*Data Mining*
IAPAR	Instituto Agronômico do Paraná
IHC	Iteração Humano-Computador
INPE	Instituto Nacional de Pesquisas Espaciais
KDD	*Knowledge Discovery Databases*
KM	Quilômetro
KWh	*Quilowatts-Hora*
LEMA	Laboratório de Estudos em Monitoramento e Modelagem Ambiental
MD	Mineração de Dados
MVD	Mineração Visual de Dados
MW	*Megawatts*
ONS	Operador Nacional do Sistema
PI	Processamento de Imagem
PPI	*Plan Position Indicator*
PS	Processamento de Sinal
RHI	*Range Height Indicator*
RIDAT	Rede Integrada de Detecção de Descargas Atmosféricas no Brasil
RN	Rede Neural

LISTA DE ABREVIATURAS

SV	Sistema de Visualização
V	Variável Velocidade Radial do Radar Meteorológico
VC	Visualização Científica
VI	Visualização da Informação
VRML	*Virtual Reality Modeling Language*
W	Variável Largura Espectral do Radar Meteorológico
Z	Variável Refletividade do Radar Meteorológico

1 INTRODUÇÃO

1.1 Considerações Iniciais

Os bancos de dados geralmente, possuem grandes conjuntos de dados numéricos ou categóricos (como datas e horas) definidos em domínios multidimensionais e cuja análise e interpretação não é trivial.

O termo *KDD* (*Knowledge Discovery in Databases*), que é responsável por extrair informações importantes desta base de dados, foi criado em 1989 e é constituído por uma série de etapas necessárias para que isso ocorra.

Após entender o domínio da aplicação e definir os objetivos a serem atingidos, iniciam-se as etapas do chamado *Processo KDD* com a *Seleção* dos dados de interesse do banco.

Dados estranhos ou inconsistentes geralmente podem ser pré-processados e estratégias podem ser tomadas para contornar estes problemas. A ausência de dados, por exemplo, pode ser facilmente resolvida pela exclusão dos registros que apresentam algum dado não preenchido ou então pela interpolação dos valores, com preenchimento dos campos incompletos.

Em alguns casos, poderá ser necessário fazer uma transformação nos dados, transformando, por exemplo, dados categóricos (como datas e horas) em valores numéricos.

Estas etapas, embora seja uma preparação nos dados que serão apresentadas à etapa de *Mineração de Dados*, são de extrema importância para todo o *Processo KDD*. Uma má seleção dos dados, ou uma exclusão de registros

importantes podem ser cruciais na etapa de interpretação dos resultados. Segundo Silver (1996), estas etapas podem ocupar mais da metade do tempo necessário de todo o processo.

Com base nas especificações do que se está querendo buscar nos dados, diversas técnicas podem ser usadas para se extrair o conhecimento desejado. Todo o processo pode ser realimentado e novas seleções ou novos métodos podem levar a diferentes soluções.

Devido à multidisciplinaridade do *Processo KDD*, a escolha dos métodos a serem usados não é uma tarefa fácil e exige o conhecimento em diferentes áreas (GIMENES, 2000), incluindo a *Visualização*.

A *Visualização,* processo para transformar informação em uma forma visual (GERSHON, 1994), tem sido usada para analisar e mostrar grandes volumes de dados multidimensionais. A *Visualização* permite, diferentemente dos métodos estatísticos, visualizar os resultados sem necessariamente saber que tipo de fenômeno deve ser analisado.

Diferentes áreas da *Visualização* são facilmente confundidas entre si. A *Visualização Científica*, por exemplo, se preocupa em estudar dados de natureza física, que possuem características espaciais. Já a *Visualização da Informação* se preocupa com a análise de dados abstratos, sem referências espaciais (RHYNE, 2003).

As técnicas de *Visualização* podem ser usadas para auxiliar, ou serem auxiliadas pelas técnicas de *Mineração de Dados*. A tentativa de integrar estas áreas ficou conhecida na literatura por *Mineração Visual de Dados* (WONG, 1999).

Ankerst (2000) inseriu a *Mineração Visual de Dados* no *Processo KDD*, sendo responsável pela comunicação entre o computador e o usuário, através da visualização.

1.2 Objetivos

O objetivo principal deste trabalho é buscar a experimentação de mecanismos e técnicas de integração da *Mineração de Dados* (MD) com a *Visualização de Informações* (VI), resultando em experimentos realizados em casos reais sobre a *Mineração Visual de Dados* como apoio visual na interpretação de grandes volumes de dados.

Assim, em uma primeira experimentação em caso real foi possível realizar uma análise visual geral das interações existentes entre as leituras da instrumentação de uma barragem de concreto. Para tanto, utilizou-se técnicas de *Mineração Visual de Dados* para extrair as informações existentes e "escondidas" no interior dos dados.

Num segundo experimento foi utilizado o algoritmo *Back-Propagation* para treinar uma *Rede Neural*, baseado nas informações da vizinhança dos *pixels* das imagens do radar meteorológico do SIMEPAR para classificá-los como sendo ruídos, ecos de terreno ou chuva.

Como objetivos secundários, as contribuições a serem citadas são:

- O levantamento das técnicas de *Visualização da Informação* e a sua reclassificação baseada nas literaturas existentes, mostrando suas vantagens e desvantagens;
- O referencial bibliográfico e forma de integrar as áreas de *Visualização* e *Mineração de dados*;
- A integração de duas áreas distintas de pesquisas, a *Visualização* e a *Mineração de Dados*, utilizando técnicas típicas de ambas para análise dos resultados;
- O desenvolvimento de uma classe em Java para ser integrado ao *RadVis* (*software*, disponível no SIMEPAR, para visualizar imagens de radares meteorológicos) na implementação da extração dos ruídos nas imagens de radar.

1.3 Organização do Trabalho

Este trabalho está dividido em sete capítulos:

Este capítulo é constituído por uma breve introdução dos conteúdos abordados neste trabalho, onde são mostrados os objetivos e os resultados que se deseja obter a partir da análise dos dados com técnicas de *Mineração Visual de Dados*.

O capítulo 2 é responsável por conceituar o termo *Visualização* e suas áreas de pesquisa, *Visualização Científica, Visualização da Informação* e *Mineração Visual de Dados*. Ainda neste capítulo, são estudadas algumas técnicas de interação, responsáveis por facilitar na interpretação dos resultados.

No capítulo 3 é visto como se realiza a extração de informações úteis em banco de dados. Este processo que passa por diversas etapas, incluindo a *Mineração de Dados* e a *Visualização,* é conhecido por *Processo KDD*.

Alguns métodos de *Mineração de Dados* são vistos no capítulo 4, com o objetivo de encontrar relações entre variáveis e análise de grupos (*clusters*), além de serem estudados os conceitos de *Redes Neurais.*

No capítulo 5, apresenta-se uma revisão bibliográfica sobre *Visualização da Informação.* Aqui diversas técnicas são estudadas e uma caracterização baseada nas características dos dados é realizada.

As técnicas de *Visualização da Informação* junto às técnicas de *Mineração de Dados* são usadas nos casos descritos no capítulo 6. A união delas ficou conhecida na literatura por *Mineração Visual de Dados,* usada em duas situações. O primeiro caso, com a análise de instrumentos geotécnico-estrutural da Barragem de ITAIPU, verificando as relações existentes entre eles. E em uma segunda situação de estudo, na predição de chuvas nas imagens do radar meteorológico do SIMEPAR.

No capítulo 7 são apresentadas as conclusões deste trabalho e sugestões para trabalhos futuros.

2 VISUALIZAÇÃO

2.1 Considerações Iniciais

Este capítulo é fundamental para se familiarizar com os conceitos que envolvem a *Visualização* permitindo diferenciar *Visualização Científica* (VC) de *Visualização da Informação* (VI).

A visualização de dados multivariáveis ou multidimensionais, sendo um subcampo da VC (WONG, 1997), é um conceito importante estudado por diversos cientistas. Diversas técnicas para visualizar dados multidimensionais estão sendo usadas para apoiar a tomada de decisões. Estas técnicas serão vistas separadamente no capítulo 5.

Neste capítulo será visto um modelo de visualização baseado em fluxo de dados, onde a imagem é formada a partir de uma série de transformações nos dados. Além disso, foi feito um levantamento de sistemas de visualização sob um ponto de vista de uma nova classificação de sistemas que pode ser usada para gerar as visualizações.

2.2 Definições e Conceitos Iniciais

Segundo McCormick, Defanti e Brown (1987) a *Visualização* é um método de computação que transforma o simbólico no geométrico, permitindo que

pesquisadores observem os resultados de seus experimentos e simulações computacionais. A visualização provê um método para ver o invisível.

Já Foley e Ribarsky (1994) consideram que uma definição útil de *Visualização*, poderia ser a ligação (ou mapeamento) de dados para uma representação que pode ser percebida. As ligações poderiam ser visuais, audíveis ou táteis, quem sabe, até uma combinação destas. Portanto, atualmente também se considera *Visualização* a representação de dados por meio de estímulos a outros sentidos como o tato e a audição.

Conforme Gershon (1994), *Visualização* é mais do que um método computacional. *Visualização* é um processo para transformar informação em uma forma visual, permitindo aos usuários observar a informação. A exibição visual resultante permite ao cientista ou engenheiro perceber características escondidas nos dados, porém necessárias para análises exploratórias visuais dos mesmos.

Essas definições deixam claro o principal objetivo da Visualização que é prover um maior compreendimento de um determinado processo, conjunto de dados ou informações.

Conceitualmente entende-se por *Visualização* a transformação de informações, das mais variadas naturezas, em representações gráficas com o objetivo de tornar essas informações mais inteligíveis para a mente humana. Para isso conta-se com o principal sentido humano, a visão.

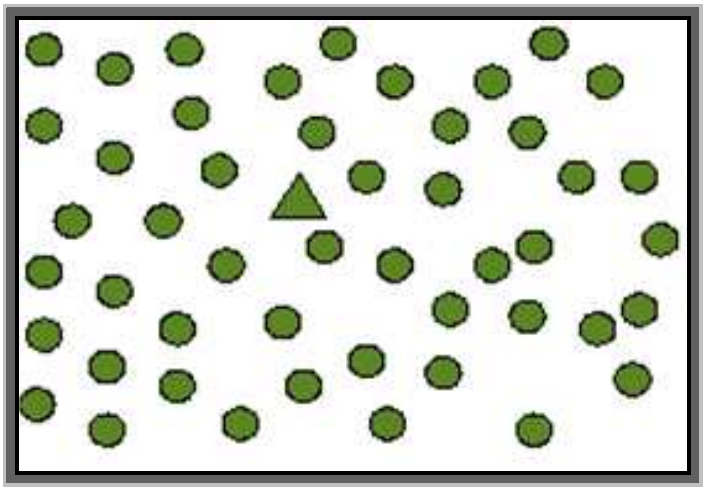


Figura 1 - Reconhecimento de padrões usando a visão. Onde está o Triângulo?

2 VISUALIZAÇÃO

A visão é o sentido humano que possui maior capacidade de captação de informação por unidade de tempo. Além de ser rápido e paralelo, a visão é treinada para reconhecer padrões. Tente-se, por exemplo, encontrar o triângulo em meio aos círculos da figura 1. Note que esta tarefa é realizada rapidamente, comprovando a afirmação inicial. Se o triângulo estivesse pintado de uma cor diferente a dos círculos, este processo de reconhecimento seria ainda mais fácil e rápido.

Para se atingir o objetivo da *Visualização*, esta é apoiada por diversas áreas científicas, como *Computação Gráfica* (CG), *Interação Humano-Computador* (IHC), *Processamento de Imagem* (PI), *Processamento de Sinal* (PS). Dessa forma a *Visualização* por meio do computador, apoiada em técnicas destas diversas áreas científicas, segundo Brodlie (1992), tem proporcionado inúmeros benefícios para as mais diversas áreas. Benefícios esses como o aumento de produtividade e maior rapidez e eficiência na tomada de decisões.

Dentre as inúmeras áreas que têm se beneficiado com produtos da *Visualização*, merecem destaque o campo de dinâmica dos fluídos, por ser o precursor do uso da *Visualização* na área científica devido aos resultados gerados por simulação numérica (LUO, STOKES e BARTON, 1996). A figura 2 mostra o fluxo de ar ao passar por um veículo em movimento, as cores representam a intensidade da pressão no encontro do ar com a superfície do automóvel.

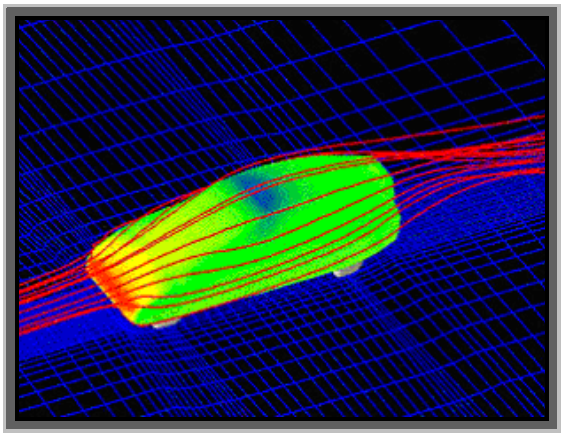


Figura 2 - Fluxo de ar em torno do automóvel (LUO, X. -L. et al., 2007)

2 VISUALIZAÇÃO

Profissionais da medicina também constituem exemplos típicos de usuários de *Visualização*, conseqüência direta do aumento da capacidade dos equipamentos de medição (tomografia computadorizada, ressonância magnética, etc.). Na figura 3 é mostrada a representação visual da anatomia humana através de renderização[1] de volumes, técnica bastante conhecida em VC.

Figura 3 - Representação visual da anatomia humana (HÖRNE, K. H. *et al*, 2007)

Outras aplicações científicas e profissionais para a *Visualização* são, por exemplo, nas áreas da Biologia Molecular, Meteorologia, Ciências Ambientais, Microscopia, Odontologia, Física Nuclear, Engenharias, Geologia e Geografia (MINGHIN; LEVKOWITZ, 2006).

Por meio de *Visualização* podem ser utilizadas técnicas interativas que permitam rápida e facilmente alterar o tipo de informação analisada. As técnicas de interação serão vistas com mais detalhes na seção 2.3. Também é útil para a percepção de características que se aplicam a pequenos subconjuntos dos dados

[1] **Renderização**: O termo "renderizar" (do inglês *to render*) vem sendo usado na computação gráfica, significando converter uma série de símbolos gráficos num arquivo visual, ou seja, "fixar" as imagens num vídeo, convertendo-as de um tipo de arquivo para outro, ou ainda "traduzir" de uma linguagem para outra.(RENDERIZAÇÃO, 2008).

e que poderiam passar despercebida se fossem utilizados somente meios estatísticos, pois estes consideram basicamente características genéricas.

Para os métodos visuais não é necessário saber que tipo de fenômeno deve ser analisado ou que questões específicas devem ser feitas, tal como acontece com métodos estatísticos, pois, em termos humanos, tais características se tornam explícitas quando os dados são representados graficamente.

Assim, técnicas e ferramentas de *Visualização* têm sido usadas para analisar e mostrar grandes volumes de dados multidimensionais, freqüentemente variantes no tempo.

2.3 Interação e Navegação

As técnicas de *Visualização* de grandes volumes de informação muito provavelmente não permitem apresentar toda informação numa única vista e com um grau de detalhes desejado. Constantemente será necessário analisar regiões dos dados mais de perto ou ver as informações em ângulos diferentes.

As técnicas de interação e navegação possibilitam uma maior varredura no espaço virtual permitindo a interação direta do usuário com as informações apresentadas. Segundo Furnas e Jul (1997), a navegação é definida como "*o processo pelo qual as pessoas determinam onde estão, onde está tudo o resto, e como chegar a objetos ou lugares específicos*".

Os principais métodos de interação com o mundo virtual são estudados pela *Computação Gráfica* (CG), dos quais a translação, a rotação, o escalamento, o espelhamento e o cisalhamento permitem ao usuário explorar todo o "mundo" virtual. Estas técnicas de CG são baseadas em transformações geométricas aplicadas aos pontos dos objetos (COHEN; MANSSOUR, 2006).

A *translação*, matematicamente, consiste em adicionar constantes de deslocamento em todos os vértices fazendo com que a imagem se posicione em outro lugar do espaço.

A *rotação* é responsável pelo giro de objetos da cena, permitindo o usuário analisar de diversos ângulos as informações mostradas. Matematicamente, a *rotação* é uma composição de cálculos de seno e cosseno do ângulo de rotação em todos os vértices da cena em torno de um dos seus eixos (*x, y* ou *z*).

A transformação geométrica de *escala* é usada para aumentar ou diminuir objetos da cena. Esta técnica consiste em multiplicar um valor de escala em todos os pontos dos objetos.

Espelhamento consiste em girar os objetos em torno de uma linha de referência (caso bidimensional) ou ao redor de plano (tridimensional) de modo que as coordenadas dos pontos dos objetos na posição inicial e os da rotação mantenham a mesma distancia em relação a estes.

Já na operação de *cisalhamento* há uma variação no valor da coordenada *x* em função do valor da de *y*. Um exemplo clássico pode ser visto na figura 4, onde é aplicada esta transformação para gerar a italização de caracteres.

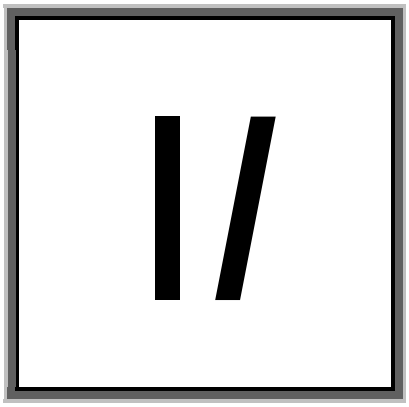


Figura 4 - Técnica de cisalhamento aplicada à italização de caracteres

Espelhamento e *cisalhamento* são transformações dificilmente usadas em técnicas de *Visualização da Informação*, porém de bastante importância na *Computação Gráfica*.

Outras técnicas bastante usadas para interagir com a cena é a aplicação de *zoom* e *pan*. Estas técnicas permitem aproximar ou afastar objetos da cena

(*zoom*) ou então deslocar os objetos da cena (*pan*) de tal forma que o usuário possa visualizar diferentes partes do universo.

No *zoom*, por exemplo, ao contrário da transformação geométrica de *escala,* a transformação não é aplicada nos pontos da cena e sim num incremento ou decremento do ângulo de visão α. Este ângulo funciona como o ângulo de abertura da lente de uma máquina fotográfica. A figura 5 mostra o efeito do *zoom* quando este ângulo é alterado.

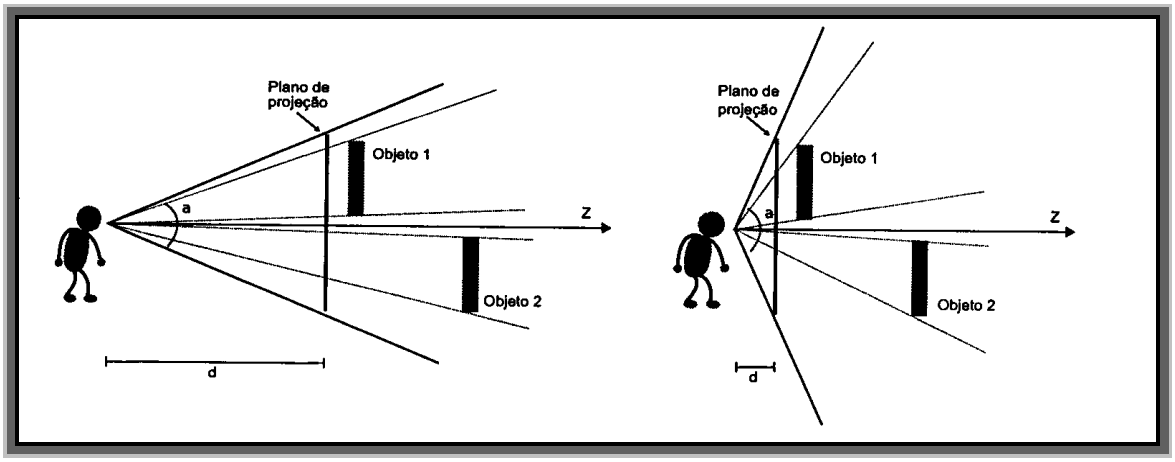


Figura 5 - Exemplo do efeito do zoom quando o ângulo α é alterado (COHEN; MANSSOUR, 2006)

Outros artefatos podem ser úteis na análise dos dados e no reconhecimento de padrões. O uso de cores, iluminação, transparência e formas diferenciadas permitem uma identificação mais rápida das informações mostradas. Assim como nas demais, um mapeamento por cores, segundo Branco (2003), aumenta o grau de percepção o que facilita a distinção na visualização (*Just Noticiable Differences*).

Estas técnicas também podem ser usadas para tratar sobreposição de dados. Em se tratando de pontos que são mapeados num mesmo lugar do espaço, um recurso denominado *jittering*, onde uma perturbação aleatória é aplicada nos pontos, resolve facilmente este problema. Mas quando a limitação está na área de visualização, a sobreposição pode ser resolvida usando

transparência/brilho nos valores com um nível proporcional ao nível de sobreposição (ARTERO, 2005).

2.4 Dados Complexos e Multidimensionais

A visualização tradicional em duas dimensões e a plotagem de linhas estão entre as técnicas mais comuns utilizadas para dados de baixa dimensionalidade. Porém, para dados com dimensões superiores, estas técnicas tradicionais não se aplicam com alto grau de confiabilidade, pois há muita perda de informações. Neste contexto, diferentes áreas da *Visualização* têm feito seus estudos para aprimorar as técnicas na interpretação de dados multidimensionais.

Segundo Wong (1997), a *Visualização de Dados* multivariáveis e multidimensionais é um subcampo importante da *Visualização Científica*. Foi estudado separadamente por estatísticos e psicólogos desde antes da computação ter sido transformada em uma disciplina.

Visualização Científica, *Visualização da Informação*, *Visualização de Dados*, *Mineração Visual de Dados*, *Discovery Visualization* e *Análise Visual* são alguns dos termos comumentes usados para designar diferentes áreas da *Visualização*. Estas áreas de estudos são facilmente confundidas entre si, pois não possuem diferenças claras e nem uma metodologia universal que permita a divisão nestas áreas.

Por exemplo, enquanto que a *Visualização Científica* está preocupada em estudar dados de natureza física, que geralmente possuem uma característica especial, facilitando seu mapeamento em representações tridimensionais (3D), a *Visualização da Informação* se preocupa na análise de dados abstratos que não possuem referências espaciais e cuja complexidade é aumentada devido às grandes quantidades de informações. Segundo Rhyne (2003), as fronteiras destas áreas não são nítidas e nem está claro que haja vantagens nesta separação.

2 VISUALIZAÇÃO

Optou-se por usar, neste trabalho, o termo *Visualização de Dados*, ou simplesmente *Visualização* à área de estudo mais genérica que engloba a *Visualização da Informação* e a *Visualização Científica*.

A tentativa de integrar os termos *Visualização da Informação* com *Mineração de Dados* (*Data Mining*), principal etapa do processo de descoberta de informações em banco de dados (*KDD – Knowledge Discovery Databases*) que será estudada com mais detalhes no capítulo 3, deu origem ao termo *Mineração Visual de Dados* (*Visual Data Mining*) (GANESH *et al*, 1996; KEIM; KRIEGEL, 1996; WONG, 1999; BRANCO, 2003). Outros termos como, *Discovery Visualization* (RIBARSKY *et al*, 1999) e *Análise Visual* (ROHRER; SIBERT; EBERT, 1999) também são usados com este propósito.

A figura 6 mostra a diferença visual destas áreas de estudos. À esquerda, na figura 6(a), uma imagem típica de *Visualização Científica* onde estão representados dados meteorológicos de pressão atmosférica. Nesta representação técnicas como mapeamento por cores, planos de cortes, isocontornos, texturas, visualização de volumes e visualização de geometrias de terreno podem ser observados. À direita, na figura 6(b), uma representação visual de dados usando técnicas de *Visualização de Informações*. As técnicas *Glyphs* e *Coordenadas Paralelas* estão nesta representação.

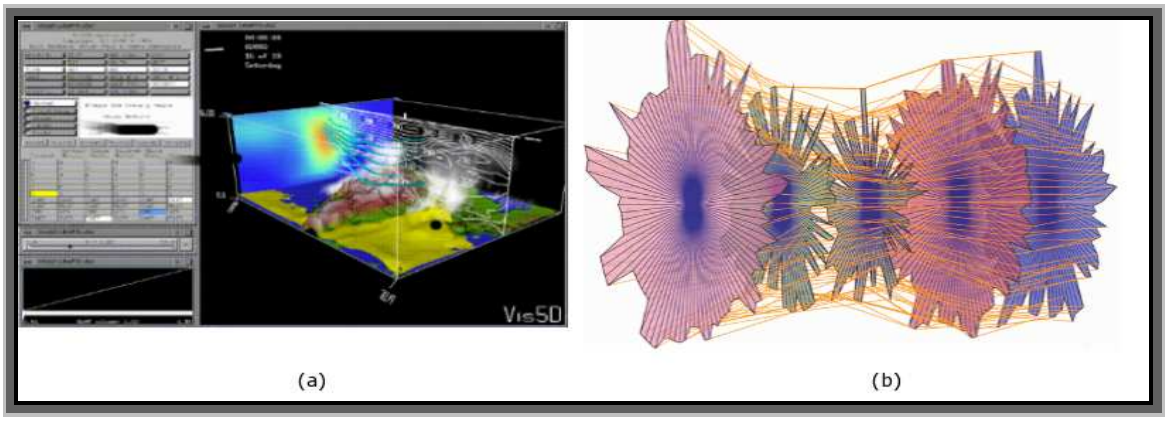


Figura 6 - Diferenças visuais entre as áreas (a) Visualização Científica (JOHNSON; EDWARDS, 2007) e (b) Visualização da Informação (FARNEA; CARPENDALE; ISENBERG, 2005; FARNEA, 2006)

2.5 Sistemas de Visualização e suas Exigências

Um *Sistema de Visualização* (SV) pode ser entendido como um processo que deve ser iniciado pela análise dos dados, verificando que tipos de informações poderão ser extraídas e quais técnicas poderão ser usadas. Então, este processo abrange desde a análise dos dados, passando pelo desenvolvimento e até a execução de um sistema usado para analisar visualmente as imagens.

Como um programa, o *Sistema de Visualização* executa certas transformações nestes dados e os exibe através de uma representação visual. O desenvolvedor da visualização deve considerar as exigências básicas para assegurar que as técnicas de transformação e de visualização a serem aplicadas aos dados sejam as apropriadas no sentido de que transmitam um alto grau de percepção. Este passo é importante nas exigências do processo, porque, utilizando-se técnicas de visualização impróprias para analisar os dados podem-se tirar conclusões errôneas.

O processo de visualização de dados passa necessariamente por três passos fundamentais, que são: a aquisição dos dados, a transformação em uma forma apropriada para representação e a “renderização” (*rendering*) ou representação na tela do monitor ou em outro *display* (ou superfície de visualização). As técnicas de visualização envolvem, portanto, algoritmos de processamento de dados que extraem os dados de interesse da amostra e os convertem em uma forma adequada para representação (BURIOL, 2006; SILVA NETO; BURIOL; SCHEER, 2007; BURIOL *et al*, 2007)

A Figura 7 ilustra o modelo de processo de visualização conhecido por *Fluxo de Dados* (*dataflow*). Este modelo, que influenciou o desenvolvimento de diversos *Sistemas de Visualização* (WALTON, 1993), é orientado aos dados, isto é, nele os dados são transformados por meio de passos lógicos até a representação final.

Figura 7 - Modelo de fluxo de dados para obtenção da imagem

A primeira transformação é descrita como filtro. Utilizando os dados brutos reais ou de uma simulação e convertendo-os em um formato que esteja pronto para as operações de visualização subseqüentes.

O segundo passo, chamado mapeamento da visualização mapeia os dados para um objeto de visualização. Estes objetos têm atributos como geometria, cor, tempo, transparência, luminosidade e reflexão, que serão usados para retratar o significado dos dados ao espectador.

O terceiro passo faz a renderização e aplica operações como translação, rotação e escurecimento, sombreando e dando aspectos mais realísticos (ou ressaltando características) aos objetos de visualização criados no passo anterior.

Sem perda da generalidade, pode-se utilizar o modelo de Card, Mackinlay e Shneiderman (1999), onde dados brutos (coletados ou gerados por algum processo) são transformados em tabelas. Aqui cabe ressaltar que o uso de tabelas é uma simplificação desnecessária, pois os dados podem ser representados em outros tipos quaisquer de estruturas de dados, dependendo da aplicação.

Um SV deve ser desenvolvido com a finalidade de transformar dados em imagens. Além disso, deve-se considerar também qual o público que este sistema irá atingir. Segundo Carvalho (2001), pode-se dividir este público em três tipos de visualização: visualização individual, visualização em grupo e visualização para o público em geral.

- Visualização individual: visualização preparada para a percepção por somente uma pessoa, não precisando de informações extras, como legenda, cores, pois o usuário já está familiarizado com os dados;

- Visualização em grupo: ao contrário da visualização individual, aqui a visualização é voltada para um grupo de pessoas partindo do princípio que estes já estejam familiarizados com os conceitos científicos e com os dados que estão sendo trabalhados. Informações adicionais, como as cores e legendas devem ser adotados;
- Visualização para o público em geral: neste caso, um protótipo deverá ser realizado para verificar a correta e adequada utilização dos dados, fazendo uso de diversas técnicas e explorando as visualizações extras como barra de cores, legendas. Devem-se transmitir conceitos de *Visualização Científica* para o usuário não científico, pois estes não têm experiências com técnicas de *Visualização* e não entendem a complexidade dos dados científicos.

Os avanços ocorridos nas tecnologias de obtenção de dados e informações por simulações computacionais e medidores (sensores), fizeram com que surgisse uma demanda de ferramentas gráficas e de auxílio computacional como apoio ao processo de interpretação dessas informações. Estes avanços, segundo Santos (2007), permitiram que os SV evoluíssem de acordo com quatro tipos diferentes que ainda se utilizam: *Bibliotecas Gráficas*, *Sistemas Interativos*, *Sistemas de Programação Visual* e *Visualização Através da Internet*.

Neste trabalho, elaborou-se uma nova classificação para *Sistemas de Visualização*. Além dos propostos por Santos (2007), um novo tipo foi adicionado, *Sistemas Interpretativos*. E o *Sistema de Programação Visual* foi subdividido em dois tipos, *Linguagens de Programação Visual* e *Ambientes de Programação Visual*. Alguns conceitos foram modificados a fim de reclassificar os novos tipos de *Sistema de Visualização*.

Assim, esta nova classificação para SV resultou em seis tipos:

- *Linguagens de Programação Visual*
- *Ambientes de Programação Visual*
- *Bibliotecas Gráficas*
- *Sistemas Interpretativos*
- *Sistemas Interativos*

- *Sistemas de Visualização Através da Internet*

Os *Sistemas de Programação Visual*, que surgiram em torno de 1980, são sistemas que fornecem módulos que implementam passos simples do *Pipeline de Visualização*[2]*.* Os módulos são ligados por programação visual sem necessidade de programação. Ou então, através da programação, novos módulos podem ser adicionados.

Os *Sistemas de Programação Visual* podem ser divididos em dois tipos: *Ambientes de Programação Visual* e *Linguagens de Programação Visual*. Ambientes, como o *Microsoft Visual Studio*, que possuem recursos que englobam uma ou mais linguagens de programação e permitam desenvolver interfaces gráficas (Visual C#, Visual J#, Visual Basic, Qt, Fltk, NetBeans, Tcl/TK), são facilmente confundidos com as Linguagens de Programação Visual.

Aqueles são *Ambientes de Programação Visual*, que utilizam ferramentas que facilitam na organização textual do código fonte para uma determinada linguagem de programação (Java, C, C++, C#, Visual Basic), ou seja, permitem a construção de um código fonte em forma de texto (especificação de diálogos e interfaces de usuário, baseadas em janelas, por exemplo) e que corresponde às informações gráficas adicionadas (como os botões, formulários, áreas de texto). Já as *Linguagens de Programação Visual*, são linguagens de programação que permitem aos usuários desenvolvedores manipular elementos graficamente e não só por especificação textual. A maioria das *Linguagens de Programação Visual* são baseadas na idéia de “caixas e setas”, ou seja, caixas ou círculos e outros elementos gráficos úteis na construção de diagramas, tratados como objetos da tela, ligados por setas, linhas ou arcos (JOHNSTON; HANNA; MILLAR, 2007). A tabela 1 exemplifica algumas das *Linguagens de Programação Visual* existentes. Na primeira coluna é apresentado o nome da linguagem, na segunda, uma breve descrição e onde esta linguagem pode ser encontrada.

[2] **Pipeline de Visualização:** seqüência de passos que devem ser realizados para a visualização do manipulador na tela. A cada alteração no manipulador, o *pipeline* deve ser novamente aplicado de forma a manter a consistência da exibição.

Tabela 1 - Linguagens de Programação Visual

Linguagem de Programação Visual	Informações
Iris Explorer	Ires Explorer: Desenvolvido inicialmente pela Silicon Graphics, é uma poderosa ferramenta para desenvolvimento de aplicações gráficas. http://www.nag.co.uk/welcome_iec.asp
AVS	AVS – *Advanced Visual System: Software* usado para criar aplicações de visualização de dados multidimensionais. www.avs.com
VisiQuest	VisiQuest: Sucessor do Khorus. Permite processamento de imagens, e análise de dados através de soluções visuais. www.accusoft.com/imaging/visiquest
OpenDx	OpenDx: Desenvolvido com base no IBM Visualization Data Explorer. OpenDX é uma ferramenta que possui diversas funcionalidades e possui vários pacotes para a visualização de informações científicas, de engenharia e de análise de dados. Seu sofisticado modelo de dados fornece aos usuários uma grande flexibilidade na criação de visualizações. www.opendx.org
LabVIEW	LabVIEW - L*aboratory Virtual Instrument Engineering Workbench*: É uma linguagem de programação gráfica pertencente à *National Instruments.* A primeira versão surgiu em 1986 para o *Macintosh*, e atualmente existem ambientes de desenvolvimento integrado também para Windows, Linux e Solaris. O principal campo de aplicação é na técnica de medição e na automatização. http://www.ni.com/labview/whatis/
MeVisLab	MeVisLab: Linguagem de programação visual destinado a criação de métodos científicos e *software* para medicina assistida e radiologia, em especial, incluindo diagnóstico auxiliado pelo computador, planejamento terapêutico. http://www.mevislab.de/
AgentSheets	AgentSheets: Usado principalmente na educação para ensinar os alunos sobre a programação e multimídia através de jogos e ciência computacional. http://www.agentsheets.com/
Outros	Alice, Amiga Vision, Analytica, Automator, Baltie, CanDO, CODE, DRAKON, Flow, G, Hollywood Designer, jMax, Ladder logic, Lava, Limnor, Max/MSP, Mindscript, OpenMusic, Pipeline Pilot, Prograph, Pure Data, Quartz, Composer, SCADE, Scala Multimedia Authoring, Simulink, Built on Squeak, Etoys scripting, Scratch, Stagecast Creator, Subtext, SynthMaker, Tersus, ThingLab, ToonTalk, Turtle Art, VEE, VisSim, virtools, WireFusion.

2 VISUALIZAÇÃO

A partir de 1960, as *Bibliotecas Gráficas* já permitiam desenhar contornos e outros tipos de gráficos simples. Estas são utilizadas via códigos em várias linguagens e exigem programação e são bastante flexíveis quanto à necessidade do usuário. Atualmente, as *Bibliotecas Gráficas* permitem desde o tratamento simples nos dados, como a criação de polígonos e linhas, até um tratamento mais avançado, como desenvolvimento de aplicações gráficas que abusam de recursos gráficos como desenvolvimento de desenhos animados e cenas de efeitos especiais. Diversas técnicas estão sendo estudadas e desenvolvidas por pesquisadores a fim de aprimorar e aperfeiçoar estes conceitos.

A tabela 2 ilustra alguns exemplos de bibliotecas usadas no desenvolvimento de animações das mais diferentes áreas, incluindo a indústria de jogos, científica e de animação.

Tabela 2 - Bibliotecas Gráficas

Biblioteca gráfica	Informações
NAG	NAG - *Numerical Algorithms Group*: É uma biblioteca open source composta de um vasto conjunto de rotinas em FORTRAN para solução de problemas numéricos e estatísticos. www.nag.co.uk
VTK	VTK – *Visualization ToolKit*: Biblioteca Gráfica para desenvolvimento de aplicações baseado em computação gráfica, processamento de imagem e visualização. Rotinas desenvolvidas em C++. Possui interface para Tcl/TK, Java, Python. Desenvolvido pela KitWare. Inc. www.vtk.org
ITK	ITK – *Insight Toolkit*: Biblioteca open source usada para desenvolvimento de aplicações médicas. Implementada em C++ e possui interface para Tcl/TK, Java, Python. Desenvolvida pela KitWare. Inc. www.itk.org
IVTK	IVTK – InfoVis ToolKit: Pacote gráfico interativo escritos em Java. Inclui uma série de componentes para análise visual de dados, dentre as quais, *Árvore de Decisões* e *Coordenadas Paralelas.* http://ivtk.sourceforge.net/

Biblioteca gráfica	Informações
VIS5D	VIS5D – *Visualization in Five Dimension*: Biblioteca científica usada para visualização volumétrica baseada em OpenGl. É um sistema interativo para a visualização de grandes em dimensões como os produzidos pelos modelos numéricos meteorológicas. http://www.ssec.wisc.edu/~billh/vis5d.html
VISAD	VISAD - *Visualization for Algorithm Development*: É uma biblioteca desenvolvida originalmente em Java usada na visualização e análise de dados numéricos. http://www.ssec.wisc.edu/~billh/visad.html
OpenGL	OpenGl – *Open Graphics Library*: API gráfica multiplataforma e multilinguagem usada na construção de aplicações 3D ou 2D. Possui mais 250 funções diferentes capazes de construir cenas tridimensionais complexas. Bastante usada na indústria de jogos. Compete diretamente com o DirectX (no *microsoft windows*). www.opengl.org
OpenMap	OpenMap – *Open System Mapping Technology*: Biblioteca em Java de desenvolvimento de aplicações e Applets aplicados em informações geográficas. http://openmap.bbn.com/

Neste trabalho, adotou-se o nome de *Sistemas Interpretativos* aos sistemas que utilizam menus ou uma linguagem de comandos. Em geral, estes sistemas permitem a visualização dos dados através da interpretação de *scripts* oriundas de arquivos textos ou de linhas de comandos não havendo necessidade de escrever programas.

Este mesmo conceito foi enunciado por Santos (2007), porém para *Sistemas Interativos*. Aqui não se considerou adequado usar esta nomenclatura pelo fato deste tipo de sistema interpretar linhas de comando além de não lidar diretamente com a interação das visualizações. Os *Sistemas Interpretativos* são mais simples de serem usados do que as *Bibliotecas Gráficas*, porém são menos flexíveis e exigem um conhecimento dos comandos disponíveis. A tabela 3 ilustra alguns exemplos destes sistemas.

Tabela 3 - Sistemas Interpretativos

Sistemas interpretativos	Informações
GnuPlot	GnuPlot: É um aplicativo de domínio público, destinado à construção de gráficos e superfícies. É uma poderosa ferramenta. Uma característica importante deste aplicativo é o fato de se ter arquivos binários para diferentes sistemas operacionais, possibilitando que um arquivo *script* seja executado em diferentes plataformas. www.gnuplot.info
MatLab	MATLAB - *MATrix LABoratory:* É um *software* interativo de alta performance voltado para o cálculo numérico. O MATLAB integra análise numérica, cálculo com matrizes, processamento de sinais e construção de gráficos 3D e 2D. Permite tratamento de imagens e uso de técnicas de visualização científica e computação gráfica. http://www.mathworks.com/
IDL	IDL: *Software* ideal para análise de dados, visualização, e desenvolvimento de aplicação multi-plataforma. http://www.ittvis.com/idl/
Maple	Maple: É um sistema de álgebra computacional comercial de uso genérico. Constitui um ambiente informático para a computação de expressões algébricas, simbólicas, permitindo o desenho de gráficos a duas ou a três dimensões. O seu desenvolvimento começou em 1981 pelo Grupo de Computação Simbólica na Universidade de Waterloo em Waterloo, no Canadá, província de Ontário. http://www.scientific.de/maple.html
GraDS	GraDS: É uma ferramenta de *desktop* interativa que está atualmente em uso global à analise e exibição de ciências da Terra. Trabalha com dados de modelos de 4 dimensões (latitude, longitude, nível e tempo). O GraDS possui um rico conjunto de funções embutidas. O usuário pode adicionar suas próprias rotinas externas escritas em qualquer linguagem de programação. http://www.iges.org/grads/
R	R: Pacote para análise estatística de dados com interface por linha de comando. Técnicas como análise de covariância, componentes principais, correlação, *Coordenadas Paralelas* podem facilmente serem usadas. http://cran.r-project.org/
Octave	Octave: Clone do MATLAB. Trabalha facilmente com matrizes e dados estatísticos e possui praticamente todas as classes do MATLAB. http://www.octave.org/

Entende-se por *Sistemas Interativos,* sistemas nos quais o usuário pode alterar parâmetros e de forma interativa ver as alterações realizadas. Estes não

exigem programação e, em geral, usam recursos de *Bibliotecas Gráficas* no seu desenvolvimento. Os *softwares* são desenvolvidos com o intuito de que usuários possam visualizar seus dados sem ter conhecimentos de programação. A Tabela 4 mostra alguns destes *Sistemas Interativos*.

Tabela 4 - Sistemas Interativos

Sistemas interativos	Informações
Amira	Amira – *Visualize, Analyse, Present*: *Software* de visualização e analise de dados, bastante usado na área biológica é médica. http://www.amiravis.com
DataViewer	DataViewer: É um sistema de visualização de dados para PCs, desenvolvido com base na ferramenta VTK que possui uma interface gráfica que permite controlar diversos parâmetros dos algoritmos de visualização fornecendo ao usuário maior liberdade de interação com os dados sob investigação. http://rbv.cesec.ufpr.br/
Paraview	ParaView: *Software* open source e multi-plataforma. Permite aplicações para visualizar conjuntos de dados de tamanho variável, de pequeno a grande. Usa a biblioteca VTK para gerar as visualizações e possui interface gráfica Qt. http://www.paraview.org
Radvis	RadVis – *Radar Visualisation*: *Software* desenvolvido pelo SIMEPAR[3] para analise visual de dados de radar meteorológico. Permite visualizar locais e concentrações de chuva além de fazer animações. O RadVis foi desenvolvido utilizando a tecnologia Web Start do Java e a bibliotéca gráfica VisAD para gerar suas visualizações. http://www.simepar.br/radvis/
Velocity	Velocity: *Software* de visualização que inclui funcionalidades para visualizar dados 2D, 3D e 4D. Avaliado para MacOS e Windows. http://www.improvision.com/products/volocity/volocity_le/
XmdvTool	XmdvTool: *Software* livre para interação e exploração visual de dados multivariáveis. Desenvolvido usando as bibliotecas gráficas OpenGl e VTK. Incluem técnicas de visualização da informação como *Scatterplots, Star Glyphs, Parallel Coordinate, Dimensional Stacking, Pixel-oriented Display*. http://davis.wpi.edu/~xmdv/

[3] **SIMEPAR – Instituto Tecnológico do Paraná:** responsável pela execução das atividades de monitoramento e previsão de tempo, elaboração de laudos meteorológicos e fornecimento de dados hidrometeorológicos, bem como pela disseminação dos mesmos. (www.simepar.br)

Sistemas interativos	Informações
xgobi, rgobi, ggobi, xgvis	xgobi, rgobi, ggobi, xgvis: Sistema interativo para visualização de dados multivariados. Incluem *Coordenadas Paralelas e Scatterplots, Dimension Stacking*. http://www.research.att.com/areas/stat/xgobi/index.html
MDV	MDV – *Multidimensional Visualisation*: Programa desenvolvido para análise de dados multidimensionais. Possui uma série de técnicas da *Visualização da Informação* implementadas. (Artero,2005)
CViz	CViz: Ferramenta projetada para analise visual de dados de alta dimensionalidade, em geral, conjuntos de dados complexos. CViz facilmente carrega os conjuntos de dados e exibe os fatores mais importantes relacionados com a agregação dos registros. http://www.alphaworks.ibm.com/tech/cviz
SisRaios	SisRaios: *Software* desenvolvido no SIMEPAR para monitoramento em tempo real da localização de ocorrências de descargas elétricas atmosféricas. Desenvolvido em Java com auxílio da Biblioteca OpenMap. http:// www.simepar.br
SatVis	SatVis - *Satellite Visualisation*: Aplicativo desenvolvido pelo SIMEPAR para visualizar imagens do Satélite GOES e NOAA. Foi implementado em linguagem Java e usa a biblioteca gráfica VTK para apresentação das imagens. http://www.simepar.br
ParVis	ParVis: Ferramenta para análise visual de dados multidimensionais a partir da técnica de Visualização da Informação *Coordenadas Paralelas*. O ParVis permite alterar as posições dos eixos de forma interativa. http://home.subnet.at/flo/mv/parvis/
HCE	HCE – *Hierarchical Clustering Explorer*: Sistema para análise exploratória de grandes conjuntos de dados. Incluem técnicas de clusterização de forma que todas possam ser vistas numa mesma tela facilitando a extração do conhecimento por comparação entre elas. http://www.cs.umd.edu/hcil/hce/
KLIMT	KLIMT – *Klassification - Interactive Methods for Trees*: *Software* interativo para analise de dados com foco na classificação e regressão por árvores de decisão. Desenvolvido em Java. http://stats.math.uni-augsburg.de/KLIMT
Mondrian	Mondrian: Sistema de visualização de dados. Diversas técnicas estatísticas e de visualização estão disponíveis. http://stats.math.uni-augsburg.de/Mondrian/
GAUGUIN	GAUGUIN – *Grouping And Using Glyphs Uncovering Individual Nuances*: *Software* para análise interativa de dados multivariados usando *Glyphs*. http://stats.math.uni-augsburg.de/software/

Sistemas interativos	Informações
CASSATT	CASSATT – *Coordinate Analysing Statistical Software Applying Tandem Transformation*: *Software* para análise exploratória de dados a partir de Coordenadas Paralelas. http://stats.math.uni-augsburg.de/CASSATT/

O sexto tipo de *Sistema de Visualização* surgiu da necessidade de explorar visualmente gráficos 3D em ambientes interativos na Internet. Na *Visualização Através da Internet*, o serviço de visualização é fornecido usando tecnologias *web,* das quais duas são bastante conhecidas: *Java-Applets* (programas desenvolvidos em linguagem Java que podem ser incluídos em uma página hipertextual em HTML) e *VRML* (*Virtual Reality Modeling Language*). A tecnologia *Java-Applets* permite executar / processar informações diretamente no cliente. Já na tecnologia *VRML*, a visualização é devolvida após um pedido do cliente via *Internet*, cujas informações são processadas num SV do servidor.

Portanto, a escolha de uma técnica de visualização é fortemente dependente das características dos dados que serão analisados. Em outras palavras, a dedução de exigências de SV para uma correta visualização deve começar com a análise dos dados. Após esta análise inicial, deve-se buscar desenvolver um protótipo para experimentar as funcionalidades desejadas para a parte visual do sistema e definir qual a linguagem ou ambiente de desenvolvimento e as bibliotecas que serão usadas, não deixando de considerar o público que se pretende atingir.

2.6 Considerações Finais

Inicialmente, neste capítulo foram apresentados os principais conceitos sobre *Visualização* em suas áreas como *Visualização Científica* e *Visualização da Informação*. O termo *Mineração Visual de Dados* é usado com o objetivo de

auxiliar o processo de descoberta de informações em base de dados *(Processo KDD)* na tentativa de unir as áreas de VI e MD na extração do conhecimento.

A *Visualização* é baseada em técnicas que transformam os dados numéricos em imagens bi ou tri-dimensionais. Imagens em três dimensões tornam possível um maior grau de interatividade com o usuário, permitindo o “passeio” no interior dos dados através de rotações, translações e outras operações de visualização. Parâmetros visuais, como transparência, luminosidade, também podem ser exploradas.

Foi apresentado também um modelo de visualização denominado *Fluxo de Dados* proposto por Walton (1993), que através de transformações aplicadas nos dados (filtro, mapeador e renderização) transformam as informações dos dados em imagem, devendo levar em consideração o público para o qual será feita a visualização, verificando se o(s) usuário(s) está(ão) familiarizado(s) com os conceitos científicos e com as técnicas de visualização a serem exploradas.

Finalmente, foi apresentado um levantamento de *Sistemas de Visualização* com uma classificação sugerida em seis tipos, sendo eles: *Linguagens de Programação Visual*, *Ambientes de Programação Visual*, *Bibliotecas Gráficas*, *Sistemas Interpretativos*, *Sistemas Interativos*, *Sistemas de Visualização Através da Internet*.

3 DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS

3.1 Considerações Iniciais

Empresas e instituições de uma forma geral estão podendo armazenar dados em grandes quantidades das mais variadas fontes. Isso se tornou possível graças ao avanço da informática. A partir da década de 80, diversos estudos começaram a ser realizados para extrair informações valiosas “escondidas” nestes dados. Um dado se transforma em informação quando este passa a ter algum significado para seu utilizador.

A exploração de dados na busca de informações é um conjunto de atividades contínuas que compartilham o conhecimento descoberto. A este processo, deu-se o nome de *Knowledge Discovery in Databases* (*KDD*).

O termo *KDD* que foi criado em 1989, refere-se ao amplo processo de descoberta de informação em banco de dados, na qual se enfatiza a aplicação de alto nível do método particular *Mineração de Dados* (MD). Enquanto que a etapa de MD se destaca pela extração de padrões escondidos nos dados, o processo completo *KDD* é mais amplo e abrange todos os processamentos (seleção, pré-processamento, e transformação dos dados) necessários para que isso ocorra, tornando possível, após técnicas de MD, avaliar e interpretar os resultados obtidos.

Na etapa *Mineração de Dados*, principal etapa do *KDD*, diversas tarefas podem ser realizadas, como análise de regras de associação e análise de agrupamentos. Para cada tarefa, diversas técnicas podem ser aplicadas. Dentre as principais técnicas utilizadas em mineração de dados, têm-se técnicas

estatísticas, técnicas de aprendizado de máquina, técnicas baseadas em crescimento-poda-validação e técnicas visuais conhecidas por *Visualização da Informação* (VI).

Neste capítulo serão vistos os conceitos envolvidos no *Processo KDD* para cada uma de suas etapas. Além de integrar ao processo à visualização usada como apoio à extração do conhecimento, procedimento conhecido por *Mineração Visual de Dados* (MVD).

3.2 Etapas do KDD

O processo de descoberta do conhecimento (*KDD - Knowledge Discovery in Databases*) é considerado, segundo alguns autores, como sendo uma análise inteligente dos dados, pois extraem das bases de dados informações triviais e não triviais que podem ser desconhecidas e potencialmente úteis.

"*Se os especialistas elaborarem uma norma (ou regra), a interpretação do confronto entre o fato e a regra constitui um conhecimento*" (GIMENES, 2007). Sendo assim, o principal objetivo do *Processo KDD* é obter o conhecimento de informações "escondidas" nos dados que sejam úteis nas tomadas de decisões. Essa tarefa possui natureza interativa e iterativa, de tal forma que não se pode esperar obter conhecimento útil pelo simples fato de introduzir uma grande quantidade de dados em um determinado programa ou sistema.

Por ser um processo interativo, o *KDD* envolve profissionais que devem possuir uma boa comunicação, viabilizando a troca de informações. Estes profissionais possuem diferentes especialidades e cada um com seu papel dentro do processo.

Para Batista (2003) e Lourdes (2007), o analista de dados é a pessoa que tem conhecimento sobre o funcionamento dos algoritmos e das ferramentas utilizadas no processo, mas não conhece o domínio aos quais os dados pertencem. Já o especialista no domínio, é aquele que tem conhecimento na área onde será aplicado o *Processo KDD*, não necessariamente precisa conhecer as

técnicas, responsáveis pela manutenção, programação e limpeza nos bancos de dados. E o usuário, é quem estabelece os critérios de avaliação e decide se o conhecimento será utilizado em alguma decisão (diretores, gerentes, pessoas de nível gerencial e executivo). O usuário deve participar de todo o processo, isso lhe dará a confiabilidade necessária dos resultados obtidos.

O *Processo KDD* envolve áreas relativas ao *Aprendizado de Máquina*, *Reconhecimento de Padrões*, *Bases de Dados*, *Estatística e Matemática*, aquisição de conhecimento para *Sistemas Especialistas* e *Visualização de Dados*. Este processo utiliza métodos, algoritmos e técnicas oriundos destas diversas áreas, com o objetivo principal de extrair conhecimento a partir de grandes bases de dados. A figura 8 mostra a relação entre estas diversas áreas.

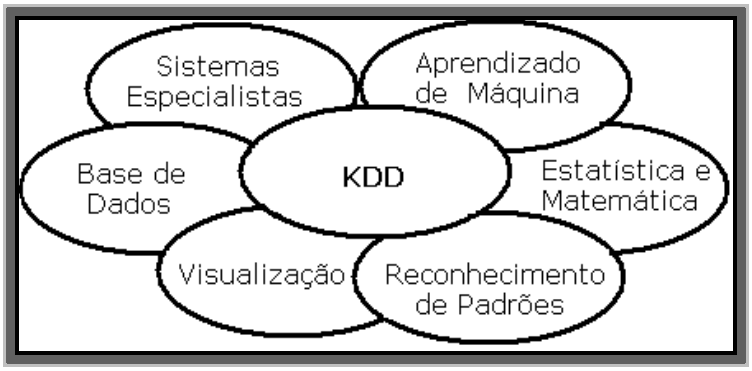


Figura 8 - Relação das áreas do *Processo KDD* (GIMENES, 2000)

O *Aprendizado de Máquina* é a área onde são utilizados modelos cognitivos ou estratégias de aprendizado de máquina, bem como os paradigmas para a aquisição automática de conhecimento.

Na área *Reconhecimento de Padrões* concentram-se estudos sobre as teorias e os algoritmos para extração de padrões e modelos.

Na área *Bases de Dados* existem tecnologias específicas, bem como uma série de pesquisas que objetivam melhor explorar as características dos dados a serem trabalhados. Tem-se, por exemplo, pesquisas que trabalham interativamente com bases de dados relacionais de clientes em atividades de *marketing*.

Modelos Matemáticos ou Estatísticos podem ser construídos para determinar regras, padrões e regularidades. No caso específico da Estatística, essa disponibiliza um grande número de procedimentos técnicos e testes para as tarefas de *Mineração de Dados*. Algumas serão apresentadas no capítulo 4.

Os *Sistemas Especialistas* são programas complexos de *Inteligência Artificial*[4] criados para resolver problemas do mundo real. Inicialmente, estes sistemas ofereciam apenas mecanismos para a representação do conhecimento, raciocínio e explicações. Posteriormente foram incorporadas ferramentas para a aquisição do conhecimento.

Finalmente, a *Visualização* que inclui diversas técnicas que podem ser usadas para apoiar ou serem apoiadas na interpretação e na geração dos resultados. Além de poder visualizar várias informações numa mesma imagem de forma interativa, permitindo navegar e selecionar áreas importantes. Em se tratando de dados oriundos de tabelas sem comportamento espacial, diversas técnicas de *Visualização da Informação* serão estudadas no capítulo 5.

Para Fayyad (1996), o *Processo KDD* é um conjunto de atividades contínuas que são compostas por cinco etapas, *Seleção dos Dados*, *Pré-processamento e Limpeza*, *Formatação ou Transformação*, *Mineração de Dados* e *Interpretação* (ver figura 9).

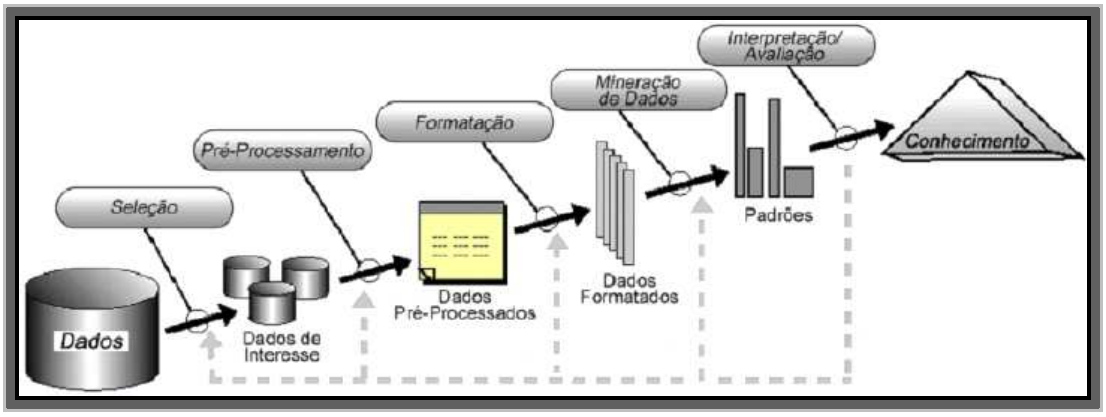


Figura 9 - Etapas do *Processo KDD* (FAYYAD, 1996)

[4] **Inteligência Artificial:** área de estudo que tem por objetivo imitar os comportamentos da natureza. A técnica rede neural, por exemplo, é uma tentativa de imitar a capacidade que os seres humanos têm para processar e armazenar informações de forma inteligente.

Inicia-se o processo com o entendimento do domínio da aplicação e dos objetivos a serem atingidos. Então uma seleção poderá ser realizada nestes dados a fim de trabalhar com os dados de interesse. Logo em seguida vem a etapa de limpeza, através de um pré-processamento dos dados. Os dados pré-processados podem ainda ser modificados, a fim de facilitar o uso das técnicas de *Mineração de Dados*. As etapas de pré-processamento e transformação podem levar até 80% do tempo necessário para todo o processo (SILVER, 1996).

Prosseguindo no processo, chega-se a etapa de *Mineração de Dados*, onde diversos métodos podem ser usados para extração de informações as quais são apresentadas a última etapa do *KDD*, a interpretação destes resultados, onde o conhecimento é adquirido.

Se o resultado final não for satisfatório, todo o processo pode ser realimentado alterando algumas informações as quais podem ser reprocessadas nas etapas anteriores.

3.2.1 Seleção

Uma vez definido o domínio de trabalho para o qual se deseja adquirir o conhecimento, deve-se selecionar e coletar o conjunto de dados ou variáveis necessárias, ou seja, escolher os dados que realmente serão úteis no processo *KDD*. A escolha dos dados a serem estudados é uma operação importante na obtenção dos resultados finais, pois todo o processamento seguinte é baseado unicamente neles.

A *Seleção* pode ser uma fase crítica, pois muitas vezes os dados podem não estar no formato desejado, ou estarem em outros bancos de dados ou ainda estarem ausentes nestes bancos. As causas que levam à situação de ausência de dados são a não disponibilidade do dado ou a inexistência do mesmo. Para estes tipos de situações, cuidados especiais devem ser levados em consideração e serão tratados na etapa pré-processamento dos dados.

3.2.2 Pré-Processamento de Dados

A etapa *Pré-processamento de Dados* é a atividade pela qual os ruídos, dados estranhos ou inconsistentes são tratados e onde são estabelecidas as estratégias para resolver os problemas de ausência de dados. Estes dados são então pré-processados e armazenados em forma de tabelas.

As estratégias para tratar as ausências de dados podem ser feitas das seguintes maneiras, podendo ser através da análise dos dados brutos, utilizado algoritmos apropriados:

- A primeira é excluir os registros que apresentam algum dado não preenchido. Entretanto, é necessário verificar quantos registros se encontram nesta situação para avaliar a perda de informação decorrente.
- A segunda alternativa consiste em preencher os campos incompletos, usando uma interpolação, realizada a partir de uma análise dos demais registros. A média e mediana dos valores também podem ser utilizadas no caso de atributos quantitativos. Uma análise em séries temporais também pode ser usada para determinar tais dados.
- A terceira maneira de tratar o problema consiste em excluir aqueles dados que apresentam muitos registros com valores não preenchidos, o que consiste de fato em uma nova seleção de dados. Esta escolha precisa ser feita com bastante critério, considerando a análise a ser realizada.

A etapa de pré-processamento tem conseqüências cruciais nas demais etapas do *Processo KDD*, e os resultados ao final do processo podem sugerir novas tentativas com diferentes configurações. Por fim, as decisões mais indicadas são aquelas que minimizam a perda de informações.

3.2.3 Transformação de Dados

Nesta fase, os dados deverão ser modificados de acordo com o algoritmo que será aplicado, representando-os mais adequadamente para as manipulações a serem realizadas pelas técnicas de *Mineração de Dados*, por exemplo, mapeando dados categóricos em valores numéricos. Ainda nessa fase podem ser feitas combinações de atributos para reduzir a dimensionalidade dos dados.

Em geral, campos representando datas e horários precisam ser convertidos em formas numéricas. Para estes valores, que possuem uma ordem natural, uma conversão bastante utilizada, consiste em obter, no caso de datas no formato *dd/mm/yyyy*, o valor inteiro que corresponde ao número de dias transcorridos desde o dia 01/01/1900 (datas anteriores resultam em números negativos). Uma conversão similar pode ser aplicada a dados que informam horários.

3.2.4 Mineração de Dados

A *Mineração de Dados* (MD) é uma tecnologia usada para revelar informações “escondidas” em grandes massas de dados. É usada em diversas áreas, como análise de riscos, *marketing* direcionado, controle de qualidade, análise de dados científicos.

Mineração de Dados define o processo automatizado de captura e análise de enormes conjuntos de dados, para então extrair um significado. Sua utilização permite avanços tecnológicos e descobertas científicas.

A maioria dos métodos de MD é baseada em conceitos de aprendizagem de máquina, reconhecimento de padrões, estatística, classificação, clusterização, visualização.

As tarefas de *Mineração de Dados* consistem na especificação do que se está querendo buscar nos dados e que tipo de regularidades ou categoria de padrões tem interesse em encontrar.

3 DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS

Várias são as tarefas de mineração proposta na literatura científica. Dentre elas, podem-se citar *Regras de Associação* (AGRAWAL; SRIKANT, 1994), *Classificação* (MEHTA; AGRAWAL; RISSANEN, 1996; LU; SETIONO; LIU, 1995), Análise de *Clusters* (NG; HAN, 1994; ESTER *et al*, 1996; GUHA; RASTOGI; SHIM, 1998), *Análise de Outliers* (KNORR; NG, 1998) e *Análise de Padrões Seqüenciais* (*Análise Evolutiva*) (AGRAWAL; SRIKANT, 1995; AGRAWAL; SRIKANT, 1996; BREJOVA *et al*, 2000). A seguir são descritas de forma sucinta estas principais tarefas de mineração:

- *Análise de Regras de Associação*: Uma regra de associação é um padrão da forma $X \rightarrow Y$, onde X e Y são conjuntos de valores. Como exemplo, tem-se a regra de associação {pão, leite} $\rightarrow$ {café}. Esta regra diz que os clientes que compram pão e leite têm uma tendência de também comprarem café. Uma regra de associação reflete um padrão de comportamento dos clientes do supermercado. Estas regras podem ser úteis para melhorar a organização das prateleiras, facilitar (ou dificultar) as compras do usuário ou induzi-lo a comprar mais.
- *Classificação e Predição*: Classificação é o processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos, com o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificados. Clientes da faixa econômica baixa, com idade entre 50 e 60 anos são maus compradores, este é um exemplo desta tarefa de mineração.
- *Análise de Clusters*: Consiste em determinar agrupamentos ou identificar classes de objetos. Como exemplo, clientes que moram na zona sul fazem compra no mercado A, enquanto que clientes da zona norte fazem suas compras no mercado B.
- *Análise de Outliers*: *Outliers* são dados que não apresentam o comportamento geral da maioria em uma base de dados. Como exemplo prático, o uso fraudulento de cartões de crédito é determinado identificando compras em valores extremamente altos (*outliers*), que fogem do padrão habitual de gastos do cliente.

- *Análise de Padrões Seqüenciais*: Um padrão seqüencial é uma expressão da forma $<I_1,...I_n>$, onde cada I_i é um conjunto de itens. Conforme a ordem em que estão alinhados, estes conjuntos refletem a ordem cronológica em que aconteceram os fatos representados por estes conjuntos. Como exemplo, tem-se o seguinte padrão seqüencial, {[carro], [pneu, toca-fitas]}, ou seja, clientes que compram carro, tempos depois compram pneu e toca-fitas.

Diversas técnicas são estudadas para cada uma destas tarefas de mineração que podem ser realizadas de forma visual. Em particular, no capítulo 4, algumas destas técnicas serão estudadas.

3.2.5 Interpretação e Avaliação

O analista de dados verifica os resultados obtidos analisando o grau de satisfação e a consistência dos resultados com base no tempo de processamento e taxa de erro. Nesta etapa é aconselhável mais de um especialista no domínio e todos os profissionais envolvidos no processo. As Interpretações dos resultados podem ser feitas através de uma análise numérica dos resultados ou de forma visual usando técnicas de *Visualização da Informação.*

Essa fase também pode realimentar todo o processo, oferecendo novas informações que podem ser novamente processadas nas etapas anteriores, em uma tentativa de refinamento dos resultados.

3.3 Integração de Visualização e o Processo KDD

Para Rezende *et al* (2003), a *Visualização* é um processo indispensável na etapa de *Mineração de Dados*. A *Visualização da Informação* permite ao usuário adquirir percepções sobre os dados, podendo provocar o surgimento de novas hipóteses (KEIM, 1979). Este mesmo autor acrescenta dizendo que quando

comparada às técnicas automáticas de *Mineração de Dados* como às *Estatísticas* e *Máquinas de Aprendizagens*, a exploração visual dos dados apresenta vantagens excedentes: lida mais facilmente com dados altamente heterogêneos e ruídosos, é intuitiva, e não requer maior entendimento de complexos algoritmos ou parâmetros da matemática ou estatística.

As técnicas de *Visualização* podem ser usadas para dar suporte ao processo de decisão quando as técnicas de *Mineração de Dados* requerem grande interação com o usuário de forma complexa. A integração destas áreas, *Mineração de Dados* e *Visualização* dá origem ao que hoje é conhecida na literatura por *Mineração Visual de Dados* (WONG, 1999). De acordo com Wong (1999), há duas formas de integrar estas áreas:

- *Acoplamento Forte*: Aproveita os pontos fortes de cada área, unindo-as numa única ferramenta.
- *Acoplamento Fraco*: As técnicas das duas áreas são intercaladas, possibilitando um aproveitamento parcial de cada uma.

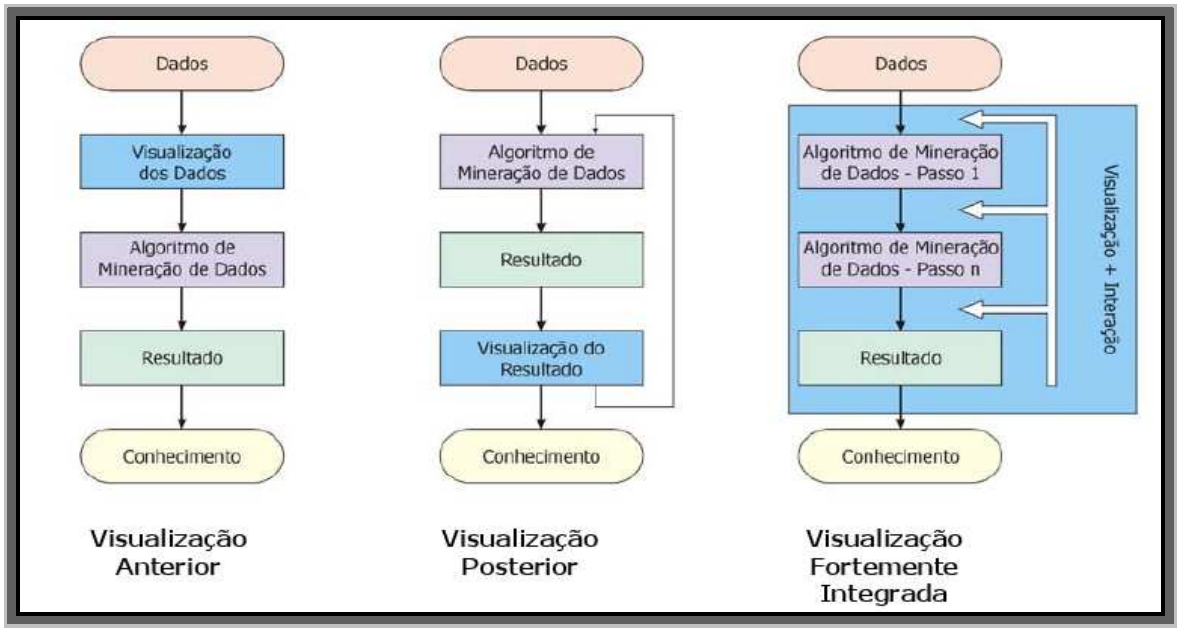


Figura 10 - Processo de integração da visualização ao *Processo KDD* (ANKERST, 2001)

A figura 10, segundo a proposta de Ankerst (2001), mostra como as técnicas de *Visualização* podem ser integradas ao processo de *Mineração de*

Dados. Nela, observa-se que a *Visualização* pode ser usada antes (*Visualização Anterior*) ou depois (*Visualização Posterior*) dos algoritmos de MD ou após cada interação (*Visualização Fortemente Integrada*).

Ankerst (2000) definiu *Mineração Visual de Dados*, como sendo um passo no *Processo KDD*, utilizando a *Visualização* como um canal de comunicação entre o computador e o usuário. Nesta abordagem, a *Visualização* seria empregada principalmente na etapa de *Mineração de Dados* e na de *Avaliação*. Sendo assim, a etapa de MD passa a ser um dos passos em que o usuário pode introduzir seu conhecimento ao invés de ser um passo meramente automatizado. Em Ankerst (2001), conforme a figura 11, o *Processo KDD* pôde ser estendido de forma que o usuário pudesse inserir seu conhecimento em todas as etapas do processo, no que foi chamado de *Processo de Descoberta de Conhecimento Centrado no Usuário*.

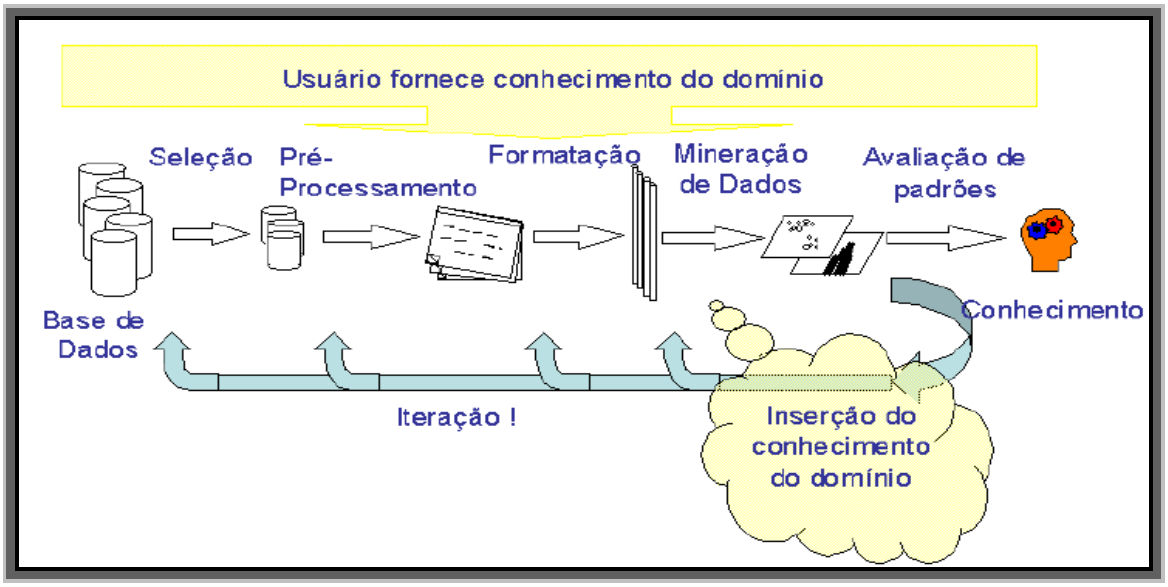


Figura 11 - *Processo KDD* centrado no usuário (ANKERST, 2001)

Portanto, a *Visualização* assume um papel importante já que esta pode ser usada em todo o processo de descoberta facilitando a sua interação com o ser humano. Por exemplo, uma mineração inicial poderá ser feita através de técnicas de *Visualização da Informação* na busca de filtrar a dimensão dos dados

trabalhados. A visualização de dados pode ser usada também para extrair informações servindo como apoio ou serem apoiadas para as técnicas de *Mineração de Dados* ou na fase final do *Processo KDD* onde é feita a interpretação dos resultados.

3.4 Considerações Finais

As etapas do *KDD*, analisadas neste capítulo, são fundamentais no processo de descoberta do conhecimento. Caso o resultado seja insatisfatório, o processo de extração do conhecimento poderá ser repensado e uma nova seleção com outro conjunto de dados poderá ser o suficiente.

Fayyad, Shapiro e Smyth (1996) salientam que todas as etapas do processo de descoberta são bastante interativas, principalmente, porque a experiência do analista é fundamental para a obtenção de bons resultados. Na etapa de mineração, normalmente, cabe também ao analista definir, com base em sua experiência, a técnica mais indicada em cada situação.

Além de poderem ser aplicadas nas demais etapas do *Processo KDD*, as técnicas visuais, conhecidas como técnicas de *Mineração Visual de Dados*, quando aplicadas na *Mineração de Dados* podem ser usadas para aumentar a extração de informações escondidas nos dados.

O resultado final deve ser compreensível para que os profissionais que tomam as decisões saibam onde, como e quando o conhecimento será aplicado.

4 TRATAMENTO DE DADOS MULTIDIMENSIONAIS

4.1 Considerações Iniciais

As compreensões de fenômenos físicos e sociais geralmente envolvem uma grande quantidade de dados. Estes dados podem ser analisados através do *Processo KDD* (capitulo 3), que começa pela seleção e tratamentos iniciais dos dados até a aplicação de métodos, dentre os quais podem ser estatísticos ou de visualização, na busca de informações úteis. A etapa de *Mineração de Dados*, principal etapa do *Processo KDD,* é responsável pela aplicação destes métodos que em geral lidam com um número grande de variáveis e também chamados de multivariadas ou multidimensionais.

"Estabelecer relações, encontrar ou propor leis explicativas é o papel próprio da ciência. Para isso é necessário controlar, manipular, medir as variáveis que são consideradas relevantes ao entendimento do fenômeno analisado." (MOITA NETO, 2007)

Segundo Landim (2007), a *Análise Multivariada* (AM) ajuda o pesquisador na redução de dados e na simplificação estrutural. Além de ser muito usada na identificação de agrupamentos, em dados amostrais ou experimentais, em investigações de dependência entre variáveis, na predição de variáveis a partir do estudo de outras variáveis e na construção e tese de hipóteses.

Os modelos multivariados, em geral, permitem o pesquisador testar ou induzir uma hipótese de um determinado fenômeno. Porém a sua utilização adequada depende do conhecimento das técnicas e das suas limitações.

4 TRATAMENTO DE DADOS MULTIDIMENSIONAIS

As técnicas multivariadas têm sido aplicadas em várias investigações científicas, nas áreas de Biologia, Física, Sociologia, Ciências Médicas, Engenharias e Meteorologia. Na Biologia, por exemplo, a Estatística é usada na seleção de plantas que serão os genitores da próxima geração, cujo objetivo é maximizar o ganho genético em um espaço mínimo de tempo. Uma série de característica das plantas pode ser convertida para um índice através do estudo multivariado nos dados (COELHO, 2007).

Na engenharia, Buzzi (2007) propôs uma metodologia para analisar as relações existentes entre as variáveis nas estruturas de barragens. O estudo foi desenvovido com os dados da barragem da Hidroelétrica de ITAIPU e teve como principal objetivo determinar relações entre os diversos instrumentos, cada um com uma funcionalidade diferente, presentes nos blocos da barragem. Para isso, foi usado *Análise de Correlação Multivariada*. Esta mesma idéia vai ser usada neste trabalho através de técnicas de *Mineração Visual de Dados* e será vista com mais detalhes no capítulo 6.

Ainda no capítulo 6, será visto como as técnicas para tratamento de dados multivariados podem apoiar as técnicas visuais. Uma aplicação baseada em *Redes Neurais* será usada para tratamento de imagens de dados provindos de radares meteorológicos em busca de uma filtragem daqueles que não representam chuva.

Assim, neste capítulo, nas próximas seções, será realizado um detalhamento das técnicas para *Análise Multivariada* de dados, incluindo *Redes Neurais*. Este estudo se torna necessário para critério de comparação e validação dos métodos visuais. Além de poderem ser usados, como apoio a modelos visuais.

4.2 Organização dos Dados

Em se tratando de dados multidimensionais, geralmente uma enorme quantidade de dados deve ser armazenada. Muitas vezes é necessário organizar os dados em tabelas de forma que estes sejam facilmente compreendidos pelo usuário. Neste trabalho, as linhas (tuplas) representarão as amostras e as colunas as variáveis conforme mostrado na tabela 5.

Tabela 5 - Organização das Variáveis

Amostras	Variáveis					
	1	2	...	j	...	m
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1m}
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2m}
⋮	⋮	⋮	⋱	⋮	⋱	⋮
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{im}
⋮	⋮	⋮	⋱	⋮	⋱	⋮
n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{nm}

A organização de dados em tabelas envolve algumas fases do *Processo KDD*. As três primeiras etapas do *KDD*, que são, seleção, processamento e tratamento dos dados, devem ser levadas em consideração durante a montagem da tabela. Estas etapas, respectivamente permitem, por exemplo, selecionar da base de dados àqueles de interesse, filtrar dados faltantes e converter dados categóricos (como datas e horários) em formatos numéricos.

Com a tabela criada, torna-se possível o uso de técnicas de *Mineração de Dados* na extração do conhecimento, o qual é obtido através da análise dos resultados.

4.3 Análise de Correlação Multivariada

A *Análise de Correlação Multivariada*, como o próprio nome já diz, envolve a relação existente entre diversas variáveis. Este método é usado para encontrar o grau de relacionamento entre as variáveis.

O coeficiente de correlação é uma medida que, através de um único número, identifica o nível de correlação entre variáveis. Para Johnson e Wichern (1998) e Triola (1999), este coeficiente é a medida de relacionamento entre pares de variáveis.

O coeficiente de correlação varia no intervalo [-1, 1] e, sendo assim, quanto mais próximo dos extremos (-1 e 1), maior é a relação entre os dados. Quando os dados são próximos de "-1", as variáveis são inversamente correlacionáveis, ou seja, os valores crescentes de uma das variáveis estarão associados aos valores decrescentes da outra, e quando este coeficiente é próximo de "1" estas variáveis possuem comportamentos próximos. Coeficientes próximos de zero sugerem que as variáveis não possuem relação (KACHIGAN, 1986).

Graficamente, em técnicas de *Visualização da Informação*, em particular a técnica *Gráficos de Dispersão*, vista com mais detalhes no capítulo 5, permite analisar a correlação entre variáveis através da dispersão dos pontos em torno de uma reta (TOLEDO; OVALLE, 1995).

Uma das grandes vantagens do coeficiente de correlação é a facilidade com que as variáveis podem ser relacionadas estando em escalas completamente diferentes e em diferentes unidades (KACHIGAN, 1986).

Quando se trabalha com várias variáveis, combinações entre estas podem ser usadas para calcular a correlação entre todos os pares de variáveis. Estes dados podem ser armazenados em forma de matriz (matriz de correlação) onde os valores da diagonal principal são iguais a "1" (correlação perfeita) e representam o coeficiente de correlação de uma determinada variável consigo mesma. Além disso, esta matriz é simétrica em relação a sua diagonal, ou seja, $C_{ij} = C_{ji}$ e quadrada. Cada linha da matriz representa a relação de uma variável com as demais (ver figura 12). Sendo assim, para *"K"* variáveis, o número total de

células será *"K^2"* das quais *"K"* pertencem às diagonais e *"K^2 – K"* são células não diagonais.

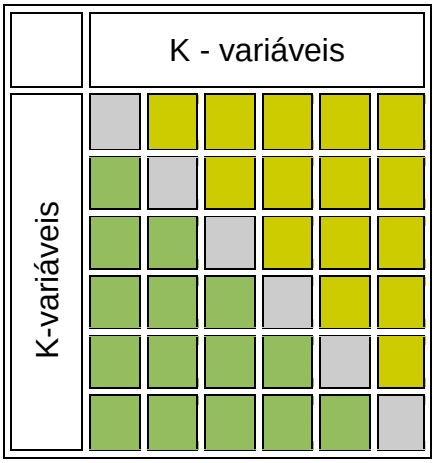


Figura 12 - Representação da matriz de correlação conforme suas propriedades. Valores das células acima da diagonal (amarelas) são iguais aos valores das células abaixo da diagonal (verdes). Células da diagonal principal (cinzas) possuem valores iguais a 1

A correlação entre pares de variáveis pode ser determinada através da seguinte equação, onde o coeficiente de correlação *(r)* é determinado através dos conjuntos de valores das variáveis *x* e *y* indicando o quão relacionadas estão estas variáveis.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}, \quad 1 \le i \le n \tag{1}$$

onde:

$$\bar{x} \text{ representa a média do conjunto de valores de } x \text{ e é definido por}: \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_i \tag{2}$$

$$\bar{y} \text{ representa a média do conjunto de valores de } y \text{ e é definido por}: \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_i \tag{3}$$

$$(x_i - \bar{x}) \text{ representa o desvio entre o } x_i \text{ e a média do conjunto } \bar{x} \tag{4}$$

$(y_i - \bar{y})$ *representa o desvio entre o* y_i *e a média do conjunto* $\bar{y}$ (5)

$$S_x \text{ representa o desvio padrão do conjunto } x: S_x = \sqrt{\frac{(x_i - \bar{x})^2}{(n-1)}} \quad (6)$$

$$S_y \text{ representa o desvio padrão do conjunto } y: S_y = \sqrt{\frac{(y_i - \bar{y})^2}{(n-1)}} \quad (7)$$

4.4 Análise de Agrupamentos

Para Soukup e Davidson (2002), a *Análise de Agrupamento* (AA) (*cluster analysis*) é um procedimento que consiste na divisão de uma população de objetos em grupos que apresentem similaridades entre os itens que os compõem.

Sendo assim, a análise de agrupamentos é o termo usado para descrever diversas técnicas numéricas cujo objetivo é classificar os valores de uma matriz de dados sob estudo de grupos discretos. Os métodos, em geral, buscam uma formulação de hipóteses à procura de agrupamentos de itens representados por pontos do espaço *n*-dimensional em um número conveniente de grupos relacionados a partir de coeficientes de similaridade. (LANDIM, 2007; DYAS; RAGAN, 2007)

Os coeficientes de similaridade podem ser gerados através das distâncias entre pares de pontos, ou de correlação entre pares de valores ou ainda através de associação entre pares de caracteres qualitativos. Existem na literatura diversos estudos que discutem estes diversos tipos de coeficientes, dentre as quais: Sneath e Sokal (1973), Everitti (1982), Prentice (1980), Gordon (1981), Greig-Smith (1983) e Pielou (1984).

Os métodos para *Análise de Agrupamentos* podem ser enquadrados nos seguintes tipos (DAVIS, 1986):

- *Métodos de Partição*: procura classificar regiões no espaço, definido em função de variáveis, que sejam densamente ocupados em termos de observações daqueles com ocupação mais rala.

- *Métodos com Origem Arbitrária:* procuram classificar as observações segundo *"k"* conjuntos previamente definidos. Neste caso, *"k*" pontos arbitrários servirão como centróides iniciais e as observações irão se agrupando, por similaridade, em torno desses centróides para formar agrupamentos.
- *Métodos por Similaridade Mútua*: procuram agrupar observações que tenham uma similaridade comum com outras observações. Inicialmente uma matriz *n x m* de similaridades entre todos os pares da observação é calculada. Em seguida, as similaridades entre colunas são repetidamente recalculadas. Colunas representando membros de um único agrupamento tenderão a presentar correlações próximas a "1" e valores menores como não membros.
- *Métodos por Agrupamentos Hierárquicos*: são as técnicas mais comumente usadas. Para o seu desenvolvimento parte-se de uma matriz simétrica de coeficientes de associação entre itens e para a combinação dos mesmos, segundo níveis hierárquicos de similaridade, utiliza-se de um procedimento aglomerativo de tal modo que cada ciclo de agrupamento obedeça a uma ordem sucessiva no sentido do decréscimo de similaridade. Embora diversas medidas de similaridade tenham sido propostas, somente duas são geralmente usadas: coeficiente de correlação e coeficiente de distância.

Método K-means, Método de Kohonen, Método Fuzzy-K-Means, Método Hierárquico Aglomerativo, Método Hierárquico Divisível, Análise dos Componentes Principais, Método de Ward, Ligação Média, Ligação Completa (vizinho mais distante) e Ligação Simples (vizinho mais próximo) são algumas das técnicas usadas para encontrar *clusters* em dados multidimensionais.

Uma forma de analisar o resultado final dos agrupamentos gerados por estes métodos é através de formas visuais em duas dimensões conhecida por *dendrograma,* um tipo específico de diagrama ou representação icônica que organiza determinados fatores e variáveis, organizando de forma hierárquica seus

agrupamentos. Em termos gráficos se assemelha aos ramos de uma árvore que vão se dividindo noutros sucessivamente.

A figura 13(a) mostra o gráfico dos pontos em sistemas cartesianos para quatro variáveis sendo $C_1(1,2)$, $C_2(5,7)$, $C_3(2,2)$ e $C_4(7,5)$, na qual é possível observar, baseando-se na distância Euclidiana, a formação de dois *clusters* $\{C_1(1,2), C_3(2,2)\}$ e $\{C_2(5,7), C_4(7,5)\}$. A figura 13(b) mostra um *dendrograma* destas mesmas variáveis, donde se extrai as mesmas conclusões.

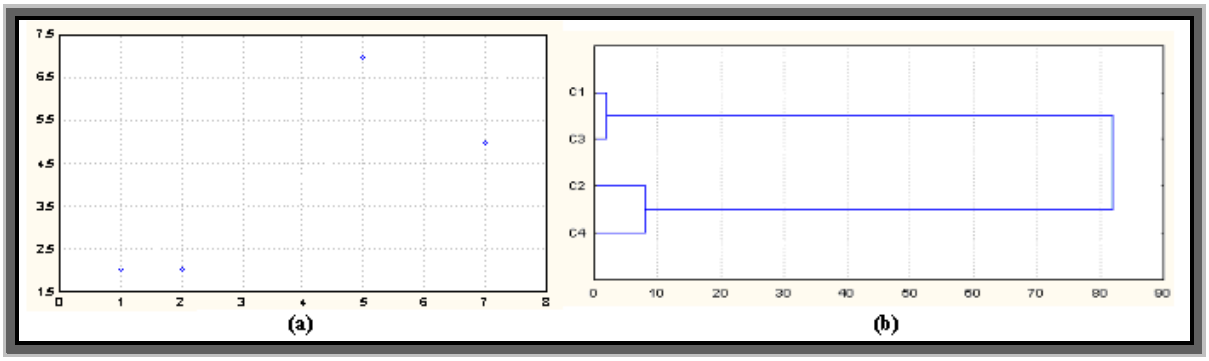


Figura 13 - Análise de agrupamentos: (a) gráfico dos pontos em coordenadas cartesianas e (b) uso de *dendrogramas* para formação de *cluster*

A *Análise de Agrupamentos* é controversa entre pesquisadores, pois pouco se sabe dos pressupostos estatísticos dos seus vários métodos. O que se têm são vários testes limitados que verificam a significância dos resultados. Dentre os quais, segundo Johnson (1998), está o *T de Hotteling* e a técnica da MANOVA (*Análise de Variância Multivariada*).

4.5 Classificação de Dados – Redes Neurais

O cérebro humano é capaz de processar informações mais rápido que qualquer outro processador conhecido. Ele é constituído de aproximadamente dez bilhões de neurônios (células nervosas) responsáveis pela transmissão de informações relacionadas a todas as funções e movimentos do organismo. A comunicação entre os neurônios é feita através de sinapses, que transmitem

estímulos através de diferentes concentrações de sódio (Na+) e potássio (K+). Os neurônios, juntos formam uma enorme rede, chamada *Rede Neural* (RN), que proporciona uma fabulosa capacidade de processamento e armazenamento de informações.

As células nervosas (neurônios) são constituídas por componentes responsáveis por determinadas funções. Os dendritos são responsáveis por receber as informações vindas de outros neurônios; o corpo da célula (soma) faz a coleta e combinam as informações vindas dos outros neurônios e finalmente, o axônio, constituído por fibras tubulares que podem chegar a alguns metros é responsável pela transmissão das informações para outras células (ver figura 14).

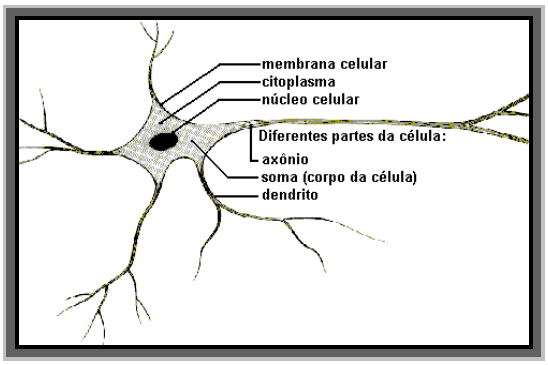


Figura 14 - Constituintes das células nervosas

As *Redes Neurais* possuim algumas características peculiares: altamente interconectada; apresenta paralelismo maciço, ou seja, muitos neurônios operam ao mesmo tempo; o processamento é distribuído, de modo que um fato pode corresponder à atividade de uma série de neurônios; admite tolerância a falhas; e a aprendizagem é exibida pelo ajustamento do efeito do acoplamento de dois neurônios.

Baseada nestes fatos, estudos têm sido realizados na tentativa de imitar o processo básico do aprendizado humano por meio da qual as novas informações são absorvidas e se tornam disponíveis para referências futuras. Esta área conhecida por *Inteligência Artificial*, que segundo Rich e Knight (1994), é a área da ciência da computação destinada ao conhecimento, construção e validação de

sistemas inteligentes, isto é, exibindo alguma forma ou características associadas à inteligência, abrange modelos capazes de aprender na tentativa de fazer com que os computadores possam realizar tarefas e pensar inteligentemente como seres humanos.

Estes estudos vêm se aprimorando a partir de 1940. Mc Culloch e Pitts, em 1943, propuseram um modelo para uma célula nervosa, chamado de neurônio artificial e mostraram que uma coleção de neurônios artificiais eram capazes de calcular certas funções lógicas. Em 1959, baseadas nas idéias iniciais de McCulloch e Pitts, e ainda após Hebb ter apontado o significado das conexões entre as sinapses e ter desenvolvido uma idéia de aprendizagem básica em 1949, Rosenblatt descreveu o primeiro modelo de *Rede Neural Artificial*, o *Perceptron*, que permitiu a aprendizagem de funções lógicas a partir de um arranjo nos neurônios artificiais numa rede com topologia particular e modificações nas conexões entre as sinapses. Em 1962, Windrow desenvolveu um tipo diferente de RN baseada numa poderosa estratégia de aprendizagem. E em 1974, após o estudo de *Redes Neurais Artificiais* quase ter sido abandonado por completo, por força do trabalho de MinsKy e Papert que expuseram as limitações do *Perceptron*, Werbus conseguiu o maior progresso nos estudos de redes neurais, lançando as bases do algoritmo *Back-Propagation*.

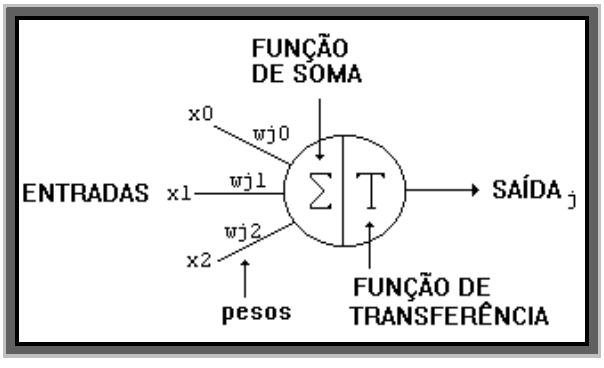


Figura 15 - Modelo de neurônio artificial. Fonte: adaptado de MCCULLOCH e PITTS (1943, p. 115-133).

Os neurônios artificiais, conforme figura 15, sofreram algumas modificações em relação aos biológicos. Os dendritos foram substituídos por

entradas, cujas ligações com o corpo celular artificial são realizadas através de elementos, chamados de pesos (simulando as sinapses). Os estímulos captados pelas entradas são processados pela função de soma, e o limiar de disparo do neurônio biológico foi substituído pela função de ativação ou função de transferência.

A *Função de Ativação* (FA), também chamada de *Função de Transferência*, é uma função matemática que, aplicada à combinação linear entre as variáveis de entrada e pesos que chegam a determinado neurônio, retorna o seu valor à saída. De acordo com Másson e Wang (1990), a *Função de Ativação* corresponde a um limiar que restringe a propagação do impulso nervoso à transposição de um certo nível de atividade, mapeando o potencial da unidade de processamento para um intervalo pré-especificado de saída. Existem diversas funções matemáticas que são utilizadas como FA. As *Funções de Ativação* mais comumente usadas são: *linear*, *degrau*, *rampa* e a *sigmóide* (ver figura 16).

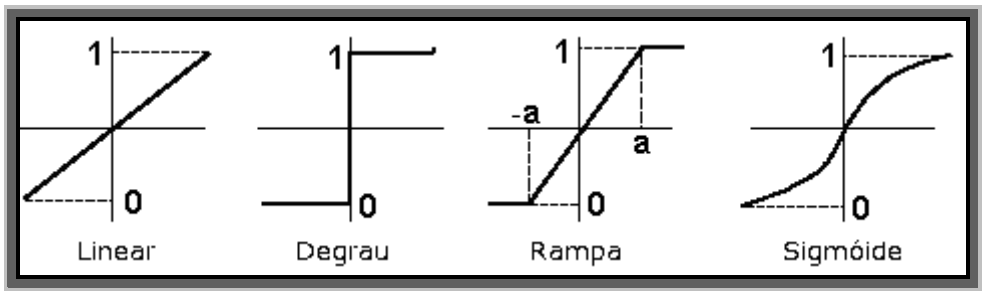


Figura 16 - Tipos de funções de ativação

O treinamento em uma *Rede Neural* pode ser supervisionado ou não supervisionado. A primeira é usada quando se tem o conhecimento da saída desejada, o sucesso é obtido quando se obtém a correta saída para uma determinada entrada. Os algoritmos de *Perceptron* e o *Back-Propagation* fazem parte deste grupo. Na segunda, o treinamento não supervisionado é usado quando se tem apenas o conjunto de entradas conhecidos e deseja-se classificá-los. Neste caso, o algoritmo busca extrair quaisquer informações estatísticas do

interior dos dados. Algoritmos de *Kohonen* e a *Rede de Hopfield* são modelos baseados nesta filosofia de aprendizagem.

Quanto ao fluxo de dados em uma rede neural, estes podem ser classificados como *Feed-Forward* (propagação dos dados unidirecional) ou *Feed-Back* (propagação dos dados nos dois sentidos).

São muitos os modelos de *Redes Neurais*, sendo que os mais estudados são: o *Perceptron*, *Redes Lineares*, e *Redes de Múltiplas Camadas*. Neste trabalho a ênfase será dada a este último, por ser um modelo de grande aplicação e usado no capítulo 6.

A figura 17 mostra o esquema geral para uma *Rede Neural* com múltiplas camadas do tipo *Feed-Forward*.

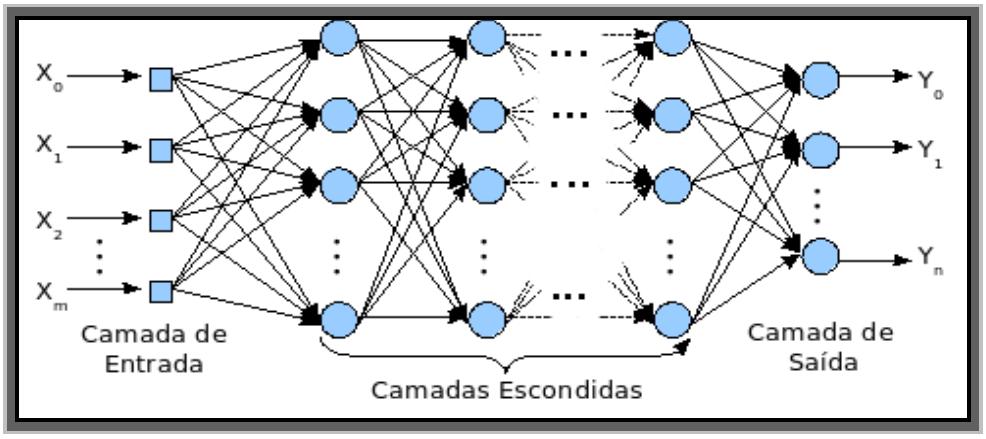


Figura 17 - Modelo de rede com múltiplas camadas

Ao aplicar *Redes Neurais* há dois parâmetros iniciais que devem ser determinados: o número de camadas e o número de neurônios em cada camada. Redes com apenas duas camadas (a de entrada e a de saída) são pouco utilizadas devido a sua limitação.

O aumento do número de camadas melhora o desempenho da rede, aumentando sua capacidade de aprendizagem, ou seja, o aumento do número de camadas aumenta a precisão com que a rede delimita as regiões de decisão. A figura 18 mostra como esse aumento de camadas influencia no treinamento.

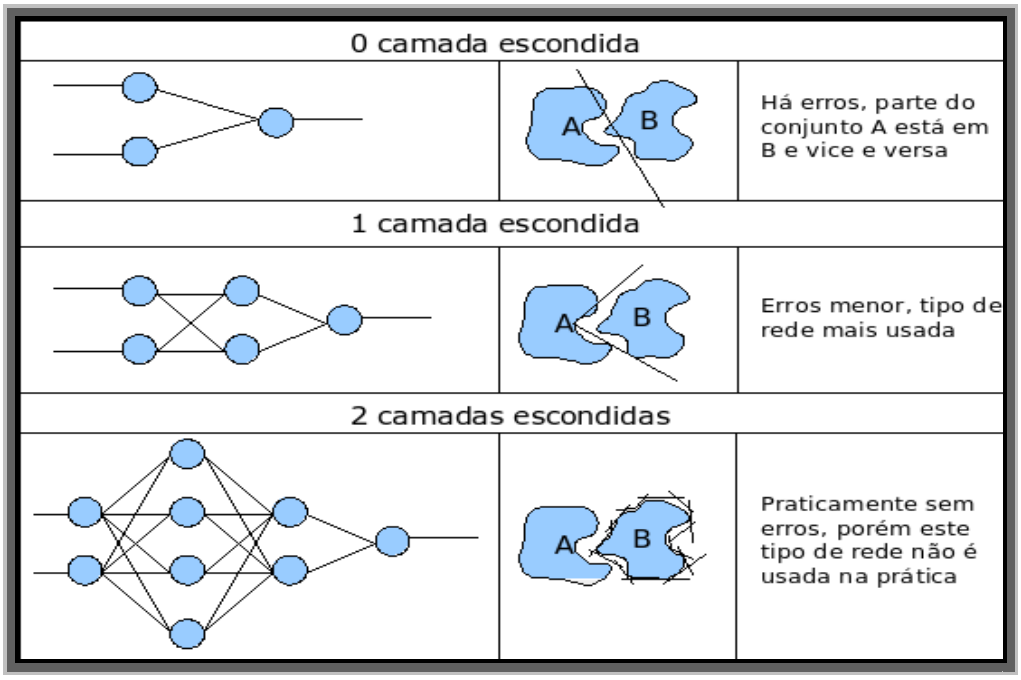


Figura 18 - Comportamento da rede ao aumentar o número de camadas escondidas. Fonte: adaptado de GORNI (1993)

O uso de muitas camadas escondidas, embora tenha um desempenho melhor, exige muito processamento e não são muito utilizados na prática por força do teorema de Kolmogorov (HECHT-NIELSEN, 1991) que afirma que o uso de uma RN com apenas uma camada escondida é suficiente para calcular uma função arbitrária qualquer a partir dos dados fornecidos.

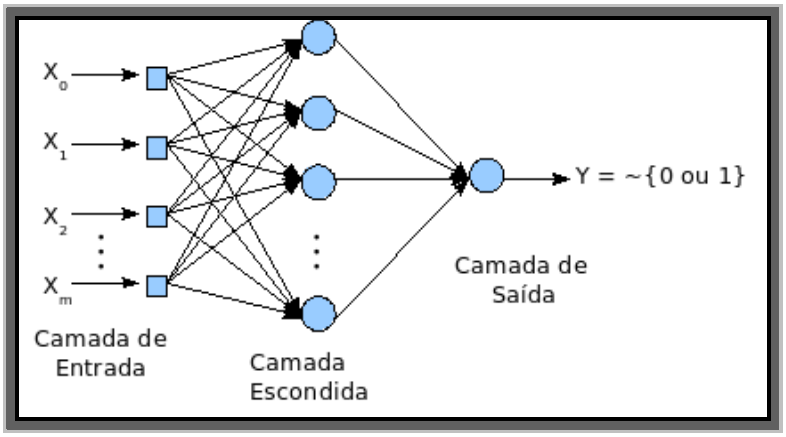


Figura 19 - Modelo de rede neural usado nas aplicações deste trabalho com uma camada escondida

4 TRATAMENTO DE DADOS MULTIDIMENSIONAIS

Neste trabalho optou-se pelo o uso de *Redes Neurais* com uma camada de entrada, com número de entradas variáveis, uma escondida com número de neurônios variáveis e uma de saída que diz se o dado é bom (valores próximos de 1) ou ruim (próximos de 0). Este modelo pode ser visto na figura 19.

As redes de múltiplas camadas podem ser criadas e treinadas pelo algoritmo *Back-Propagation* (ver Figura 20), que é constituído de duas fases: *Propagação Forward* e *Propagação Backward*. Vetores de entradas e os correspondentes vetores de saídas são usados para treinar a rede até que ela possa aproximar uma função que classifique os vetores de entradas de maneira apropriada.

É o vetor de pesos *(w)* que determina como a rede responderá a uma entrada arbitrária, é nele que são armazenadas todas as informações do treinamento da rede. Um conjunto inicial de pesos (solução inicial) deve ser apresentado à rede. Estes valores mudam a cada iteração do algoritmo.

Outros parâmetros devem ser considerados na aplicação do algoritmo *Back-Propagation*, como, a taxa de aprendizagem (γ) e o momento (α) que são valores que auxiliam a performance de uma rede neural. Alguns autores sugerem um declive gradual da taxa de aprendizagem a medida que evolui (GORNI, 1993). Outros optaram pela fixação destes valores enfatizando a necessidade de estarem no intervalo (0,1) (HAYKIN, 1994).

O parâmetro θ é um valor *thresshold* adicionado a soma ponderada, e em alguns casos é omitido, enquanto que em outros é considerado como o valor peso cujo correspondente valor de entrada é sempre igual a “1”. O papel de θ, também chamado de *bias* ou *vício*, é aumentar o número de graus de liberdade disponíveis no modelo, permitindo que a RN tenha maior capacidade de se ajustar ao conhecimento a ela fornecido.

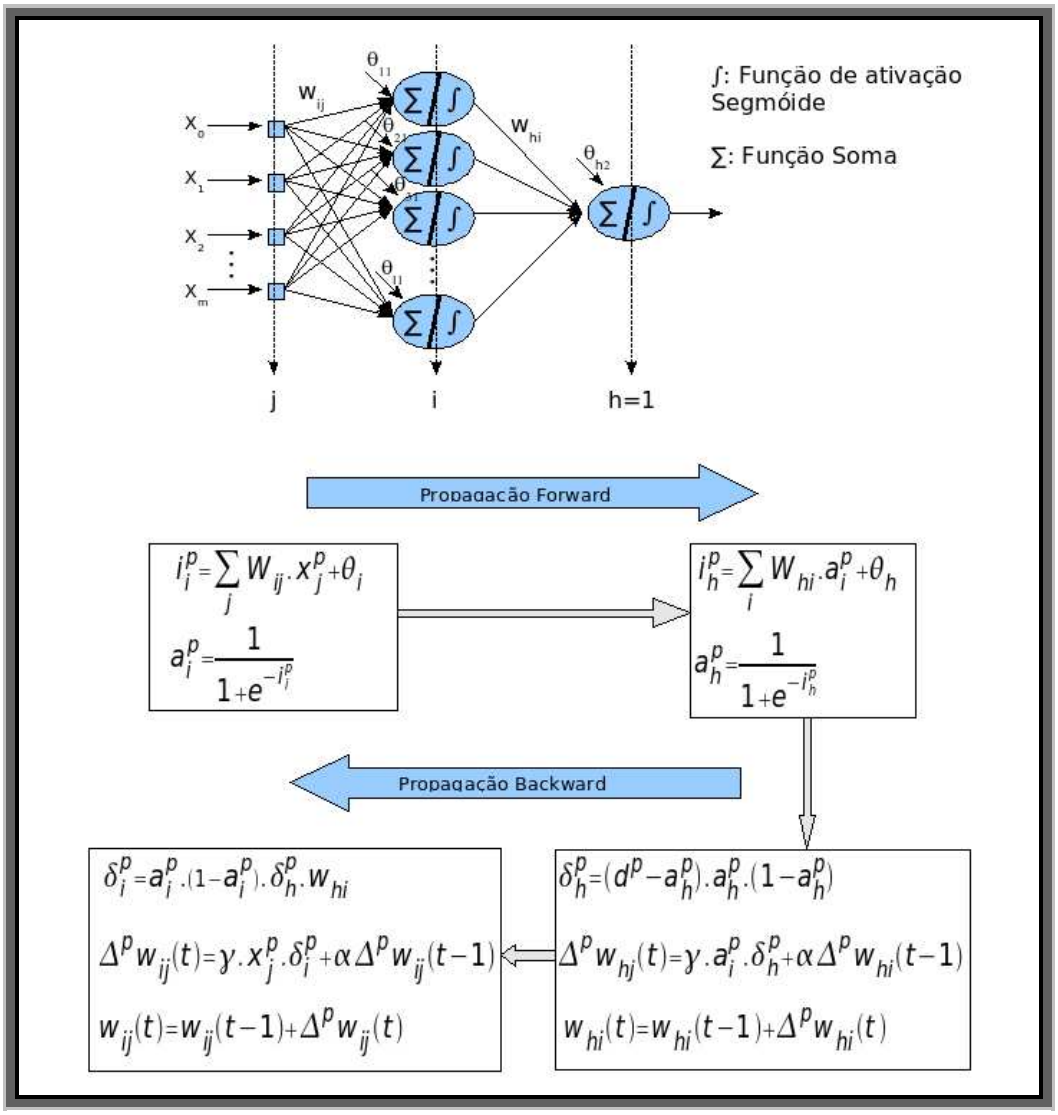


Figura 20 - Algoritmo *Back-Propagation* usando função de ativação *sigmóide*

A figura 21 mostra a importância do termo momento num comparativo com diferentes valores para taxa de aprendizagem. Na trajetória (a), o termo momento é adicionado e a taxa de aprendizagem é pequena. Leva bastante tempo (interações) para chegar a solução. Na trajetória (b), o termo momento não é considerado e a taxa de aprendizagem é alta. Neste caso, o mínimo nunca será alcançado devido às oscilações. Na trajetória (c), a taxa de aprendizagem é alta e o termo momento é considerado. O mínimo é alcançado rapidamente. (KRÖSE; VAN DER SMAGT, 1993)

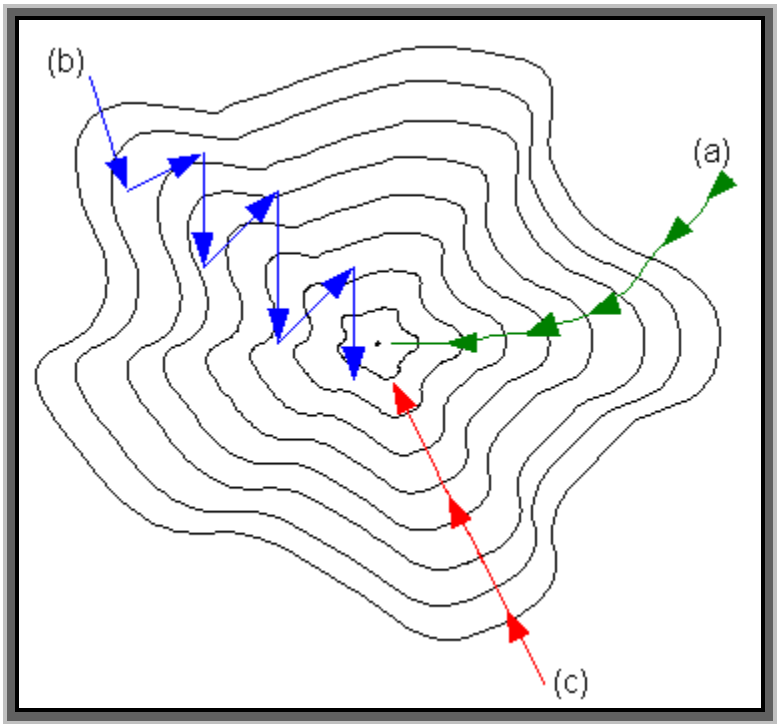


Figura 21 - Desempenho de uma rede neural conforme a variação da taxa de aprendizagem e a taxa de momento. Fonte: adaptado de KRÖSE e VAN DER SMAGT (1993)

O treinamento pode conduzir a um mínimo local ao invés de global, porém se este for um resultado insatisfatório, um novo treinamento pode ser realizado mudando o número de neurônios, ou de camadas ou até mesmo o número dos pesos iniciais ou ainda a taxa de aprendizagem e o termo momento. A escolha destes não é uma determinação simples e pode variar na busca de uma solução mais adequada.

4.6 Considerações Finais

No capítulo 3, foram vistas as etapas do *Processo KDD*, em particular, a etapa *Mineração de Dados*. A MD é o termo usado para tratar dados multidimensionais extraindo destes informações importantes “escondidas” em seu interior. O tratamento dos dados multivariados ou multidimensionais requerem

conhecimentos em diversas áreas como Estatística, Visualização, Inteligência Artificial, Banco De Dados, nas quais diversos métodos são estudados para este fim.

Na prática geralmente é importante agrupar dados conforme algumas características ou então analisar as relações existentes entre variáveis ou ainda classificar uma entrada como sendo boa ou ruim. Estes problemas podem ser resolvidos utilizando técnicas de tratamento de dados multidimensionais, dentre as quais estão, respectivamente, as de *Análise de Agrupamentos*, *Análise de Correlações* e as *Redes Neurais*.

Neste contexto, o objetivo deste capítulo foi dar a base para as técnicas existentes em *Mineração de Dados*, e não fazer um estudo aprofundado delas, posto que estas serão usadas como apoio ou para comparação e validação das técnicas de *Mineração Visual de Dados* que serão vistas nos próximos capítulos.

5 VISUALIZAÇÃO DA INFORMAÇÃO

5.1 Considerações Iniciais

Antes mesmo do uso de computadores para criar visualizações, a visualização de dados de duas ou três dimensões já era realizada, e suas técnicas têm sido usadas por muitos anos (TUFTE, 1983; TUFTE, 1990). Quando os computadores começaram a ser usados para criar visualizações, também começou o desenvolvimento de muitas técnicas novas, bem como a extensão de técnicas existentes. Tornou-se possível o tratamento de grandes volumes de dados além de permitir interação.

Dentre as técnicas de *Visualização* para exploração de dados multidimensionais, existem as técnicas tradicionais em duas e três dimensões, como os *Gráficos de Linhas e Dispersão* (CLEVELAND, 1993; BERTIN, 1981), e técnicas mais sofisticadas, que permitem uma interação maior com usuário e uma exploração em bancos de dados ainda mais robustos.

Neste capítulo, além dos conceitos da *Visualização da Informação*, uma breve discussão sobre as vantagens e desvantagens no contexto de aplicação será mencionada para cada uma das técnicas de *Visualização* a serem estudadas.

5.2 Técnicas de Visualização da Informação

A *Visualização da Informação* (VI), conforme visto no capítulo 2, é o nome dado a área de estudo voltada à visualização de dados não inerentemente espaciais, ou seja, grandes massas de dados, geralmente armazenados em enormes tabelas ou banco de dados, sem características físicas (como temperatura, pressão) e busca por informações que possam ser úteis para análise dos dados.

A visualização de dados tem sido usada como meio de comunicação desde os primórdios da humanidade. A primeira forma gráfica que se tem notícias é uma representação de uma cidade da Babilônia encontrada na região de Kirkuk no Iraque há 6200 a.C. (figura 22).

Figura 22 - Representação gráfica de uma cidade da Babilônia há 6200 a.C. (FRIENDLY, 2007)

Após alguns séculos, o uso de visualização para analisar o comportamento de variáveis deu início ao que hoje é conhecida por *Visualização da Informação*, a figura 23 ilustra uma das primeiras visualizações realizadas a mostrar variáveis, representando um estudo das órbitas planetárias ao longo do tempo no século X.

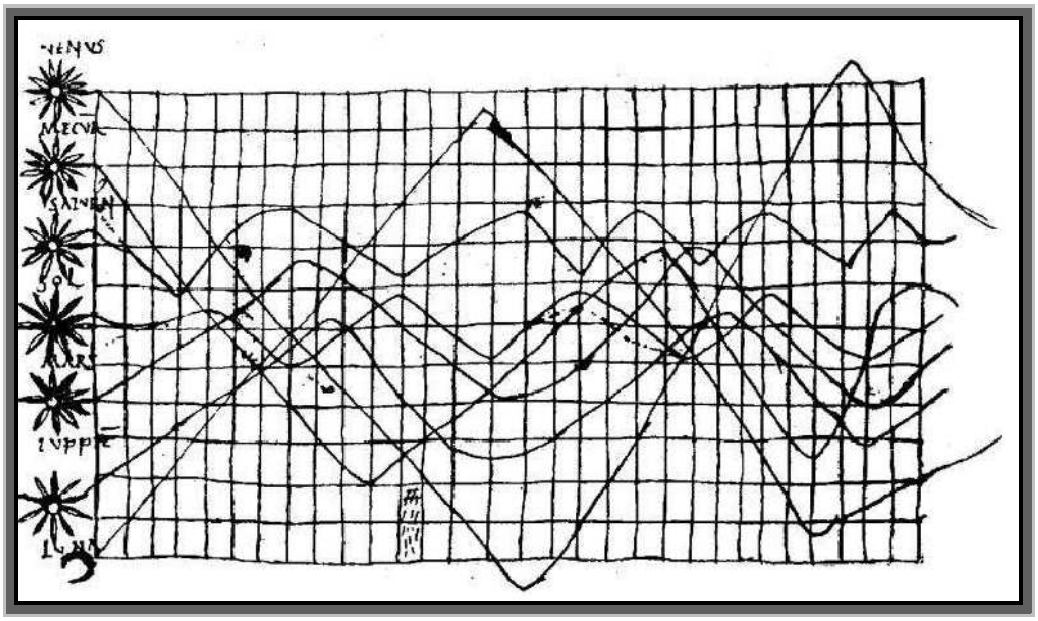


Figura 23 - Inclinação das órbitas planetárias ao longo do tempo – ano 950 (FRIENDLY, 2007; FUNKHOUSER, 1936, p. 261)

William Playfair[5], considerado o criador das primeiras técnicas de VI, fez várias visualizações entre 1770 e 1782, quando a *Visualização* se tornou importante no mercado de negócios. A figura 24 mostra a representação de dados de exportação e importação para este período.

Em 1701, Edmund Halley[6] faz uma das primeiras visualizações utilizando contornos (isolinhas) (ver figura 25).

[5] **Willian Playfair** (1759-1823): um economista inglês. Em 1786, William Playfair publicou os primeiro gráfico em seu livro **The Commercial and Political Atlas.** Esse livro é repleto de gráficos estatísticos que representam a economia no século XVIII na Inglaterra usando gráficos de barra (PLAYFAIR, 2007).

[6] **Edmund Halley** (1656 – 1742): Foi um astrônomo e matemático britânico. Halley foi o primeiro a descobrir um cometa periódico, que subseqüentemente passou a ser chamado cometa de Halley. Aplicou o método de Newton para calcular órbitas de cometas. Halley publicou os resultados de suas observações em 1705, na obra A **Synopsis of the Astronomy of Planets**. Halley também dedicou uma parte de seu tempo aos assuntos relativos à economia, engenharia naval e diplomacia, exercendo papel de destaque na publicação dos Principia, de Newton. (HALLEY, 1701)

5 VISUALIZAÇÃO DA INFORMAÇÃO

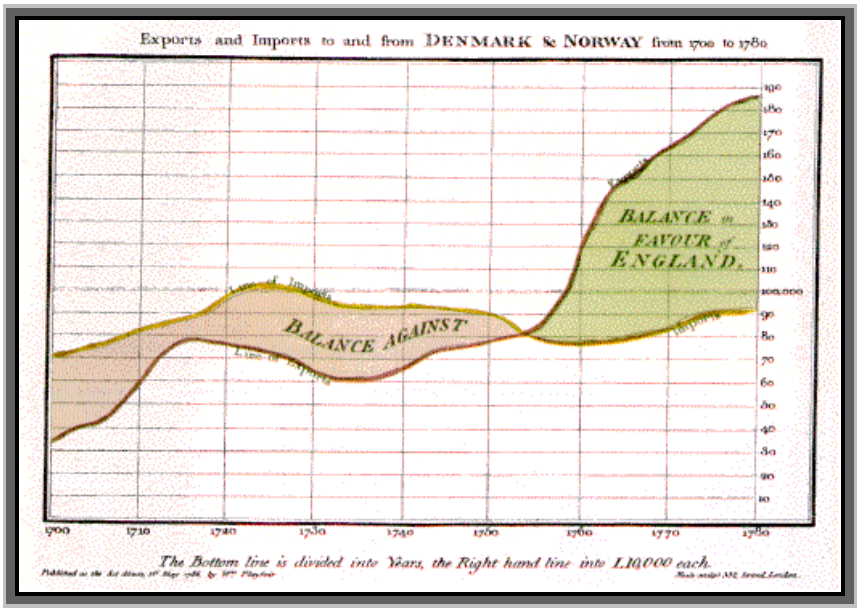


Figura 24 - Importação e Exportação entre 1770 e 1782 (FRIENDLY, 2007; FRIENDLY, 2005)

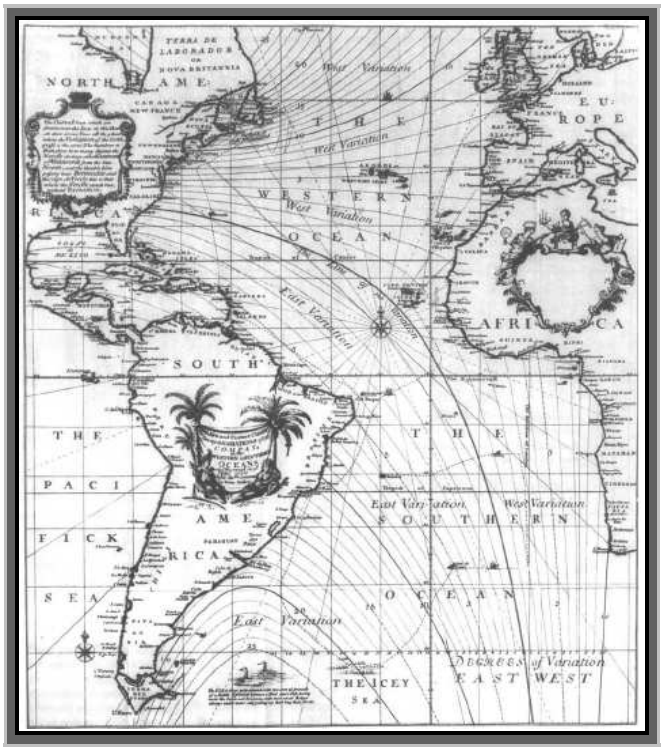


Figura 25 - Declinação Magnética (FRIENDLY, 2007; HALLEY, 1701; PALSKY, 1996)

5 VISUALIZAÇÃO DA INFORMAÇÃO

No século XX, uma das primeiras visualizações de dados multidimensionais é apresentada por Charles Minard[7] na Campanha para conquistar a Rússia. Um exemplo pioneiro de como muita informação pode ser sintetizada para se tornar mais inteligível. Nesse gráfico (ver figura 26) há quatro variáveis diferentes que contribuem para demonstrar o fracasso da campanha (em apenas uma representação bidimensional):

- A distância e direção que percorreram;
- A altitude que as tropas atravessaram;
- A variação no número de soldados à medida que as tropas morriam de fome e dos ferimentos;
- As baixas temperaturas que enfrentaram.

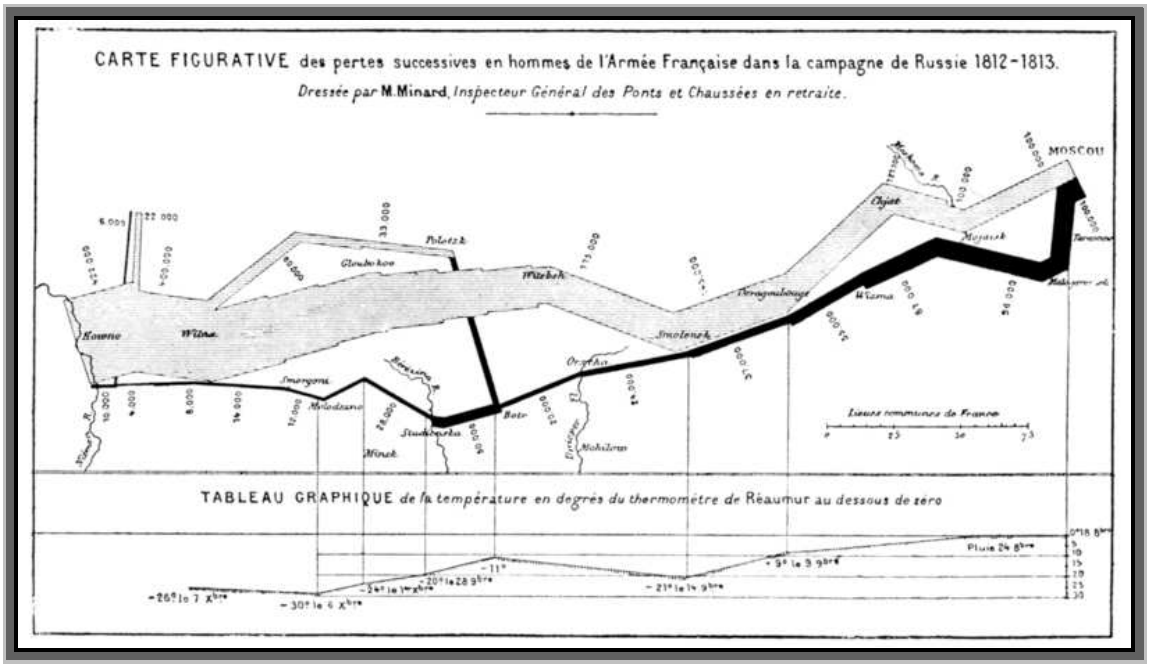


Figura 26 - Infográfico de Charles Minard sobre a marcha de Napoleão (FRIENDLY, 2007; FRIENDLY, 2002; FRIENDLY, 2005)

[7] **Charles Joseph Minard** (1781 - 1870): Engenheiro francês e pioneiro na visualização multidimensional. Ficou conhecido pela sua representação visual que mostra a marcha do exército de Napoleão durante a campanha de 1812 para conquistar a Rússia. (MINARD, 2007)

Dr. Snow[8] mostra através da *Visualização* a descoberta da causa do surto de cólera (figura 27).

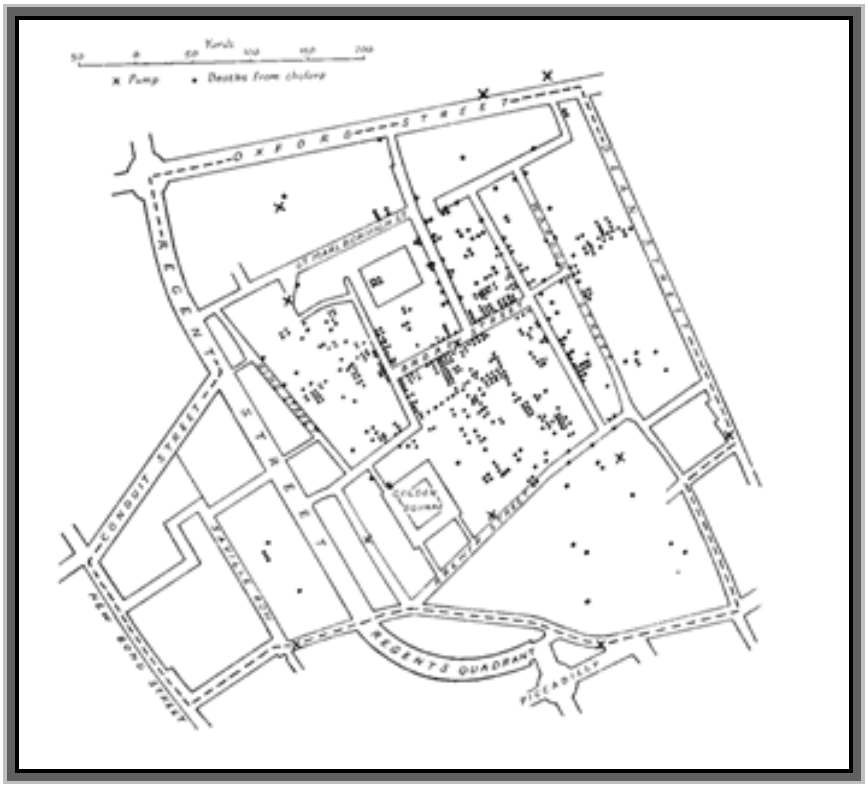


Figura 27 - Mapa de Londres com casos de cólera (pontos) e poços de água (cruzes) (FRIENDLY, 2007; GILBERT, 1958)

Na década de 80 do século XXI, o aparecimento de computadores pessoais de baixo custo e as estações de trabalho (*workstations*) trouxeram nova vida dentro da análise gráfica dos dados multidimensionais.

Atualmente, o avanço da tecnologia fez surgir computadores significativamente mais poderosos com possibilidades de gerar gráficos complexos 3D. Paralelamente houve um grande aumento na acessibilidade da

[8] **John Snow** (1813 - 1858) - Médico britânico e líder na adoção da anestesia e da higiene médica. Também é considerado um dos pais da epidemiologia, por ter identificado a cadeia de transmissão do *vibrio cholerea*, o responsável pela cólera. Antes de Snow, acreditava-se que a contaminação da cólera ocorria através do ar. (FRIENDLY, 2007)

informação e na ligação a redes de comunicação. Estes fatores tornaram possíveis novas formas de apresentar visualmente.

Os parâmetros visuais, como cor, tamanho, forma, posição, foram incorporados no *Sistemas de Visualização* e são muitos usados para representar características e propriedades dos dados em *Visualização da Informação* em duas dimensões (2D).

Em VI em três dimensões (3D), por trabalhar com uma dimensão a mais, torna possível uma representação mais eficiente do espaço limitado disponível. Esta dimensão adicional permite novos parâmetros visuais como tipo de material, luminosidade, transparência e novas técnicas de interação, como rotações geométricas e "passeio" (*walkthrough*) através dos dados, que convidam os utilizadores a explorar e manipular sistemas de informações grandes e complexos (capítulo 2, seção 2.3).

Baseado nestas formas de interação e navegação, diversas técnicas de *Visualização da Informação* estão sendo desenvolvidas com o objetivo de facilitar a interpretação de dados.

Determinar a técnica a ser usada para visualizar um conjunto de dados de uma determinada aplicação, não é uma tarefa fácil. Uma caracterização dos dados é uma das considerações iniciais na escolha de uma técnica de visualização. Diversos autores, na tentativa de padronizar as técnicas de visualização de dados multidimensionais, sugerem classificações que são feitas de diferentes maneiras e seguindo diferentes critérios.

Schneiderman (1996), por exemplo, classificou as técnicas segundo os tipos de dados e as tarefas a serem realizadas pelo usuário. Segundo ele, os dados podem ser: temporais, unidimensionais (1D), bidimensionais (2D), tridimensionais (3D), multidimensionais (nD) e dirigidos à visualização de hierarquias e de relacionamentos (grafos).

Freitas e Wagner (1995) apresentam uma proposta de caracterização de dados baseada em critérios como: classe (tipo) de informação, tipos de valores, e natureza e dimensão do domínio (vide resumo na tabela 6).

Tabela 6 - Caracterização de dados baseada em critérios, exemplos de domínios diferentes (FREITAS; WAGNER, 1995)

Critério	Classe	Exemplo
Classe de Informação	Categoria	Gênero
	Escalar	Temperatura
	Vetorial	Grandezas físicas associadas a dinâmicas dos fluídos
	Tensorial	
	Relacionamento	*Link* num hiperdocumento
Tipos de Valores	Alfa-numérico	Gênero
	Numérico	Temperatura
	Simbólico	*Link* num hiperdocumento
Natureza do Domínio	Discreto	Marcas de automóveis
	Contínuo	Superfícies de um terreno
	Contínuo-discretizado	Anos (tempo discretizado)
Dimensão do Domínio	1D	Fenômeno ocorrendo no tempo
	2D	Superfície de um terreno
	3D	Volume de dados médicos
	nD	Dados de uma população

De forma geral, os valores assumidos pelas variáveis podem ser classificados nos formatos básicos nominal e quantitativo. O primeiro apresenta valores claramente distintos, discretos e enumeráveis. O segundo representa valores numéricos, contínuos, sobre os quais podem ser aplicadas operações aritméticas.

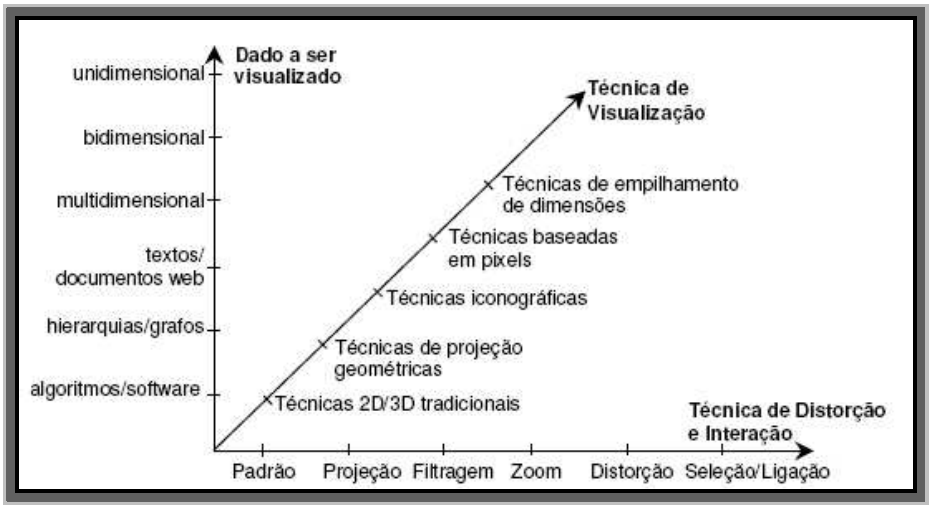


Figura 28 - Classificação das Técnicas de Visualização (Keim, 2002)

5 VISUALIZAÇÃO DA INFORMAÇÃO

Keim (2002) sugere uma classificação segundo três critérios, ilustrada na figura 28: a natureza do dado a ser visualizado, a abordagem de mapeamento adotada pela técnica e os métodos de interação e distorção usados para manipular a representação visual.

Quanto à natureza dos dados, estes podem ser unidimensional (1D), bidimensional (2D), multidimensional (nD), texto e hipertexto, hierarquias/grafos e algoritmos/*softwares*. A abordagem para interagir com os dados visualmente podem ser projeções, filtragem, zoom, distorção e Seleção/Ligação. Já as Técnicas de *Visualização* estão classificadas como *Técnicas 2D/3D Tradicionais*, *Técnicas de Projeções Geométricas*, *Técnicas Iconográficas*, *Técnicas Baseadas em Pixel* e *Técnicas de Empilhamento de Dimensões*.

Na tentativa de agrupar melhor as técnicas de *Visualização*, neste trabalho uma nova classificação para os grupos de técnicas foi adotada. Este novo agrupamento foi gerado a partir da literatura encontrada de diversos autores e embora algumas técnicas possam pertencer a dois ou mais grupos aqui foi considerado àquele que melhor se caracteriza devido à natureza dos dados. Além das técnicas de *Visualização* adotadas por Keim (2002), neste trabalho consideraram-se ainda três novos grupos, a saber: *Técnicas Hierárquicas*, *Técnicas Dinâmicas* e *Técnicas Híbridas*. As *Técnicas Empilhamento de Dimensão* foram colocadas juntas às *Técnicas Hierárquicas* devido à natureza hierárquica dos dados.

A escolha de uma técnica para visualizar um conjunto de dados se torna mais fácil quando é considerada a caracterização dos dados. Porém, diversas técnicas devem ser aplicadas na tentativa de determinar qual delas transmitirá um nível mais satisfatório da informação que se procura. A comparação entre as técnicas para um determinado conjunto de dados poderá ser útil para uma maior exploração da informação.

Nas próximas seções serão vistas as principais técnicas de visualização onde se procurou mencionar as vantagens e desvantagens para cada técnica em questão.

5.2.1 Técnicas 2D e 3D Tradicionais

Técnicas como *Gráfico de Pizza*, *Dispersão*, *Linhas*, *Barras* entre outros são amplamente utilizados na exibição de dados em duas e três dimensões. Este grupo abrange todas estas técnicas de *Visualização* de dados. O programa Excel, ferramenta do Microsoft, possibilita o uso de diversos gráficos para visualizar dados. A figura 29 mostra algumas destas técnicas.

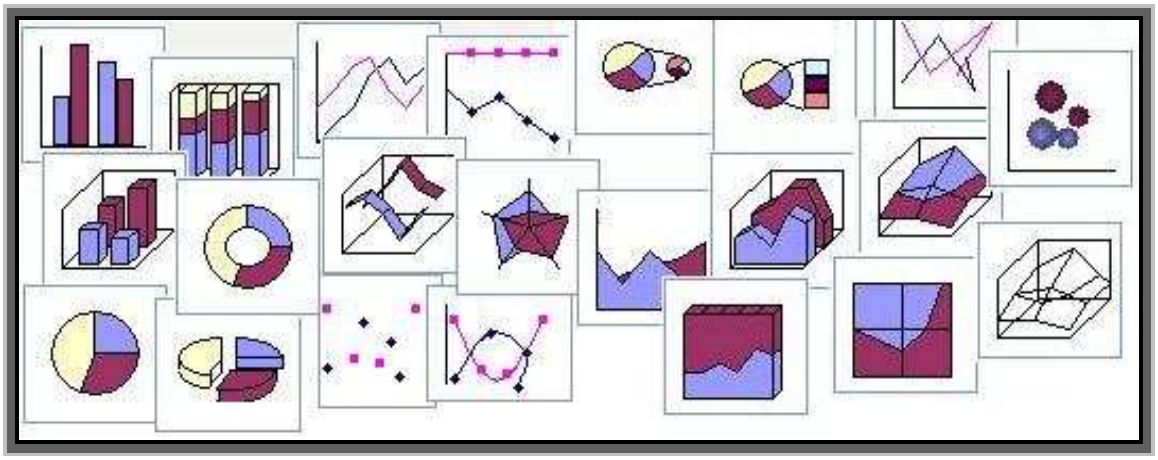
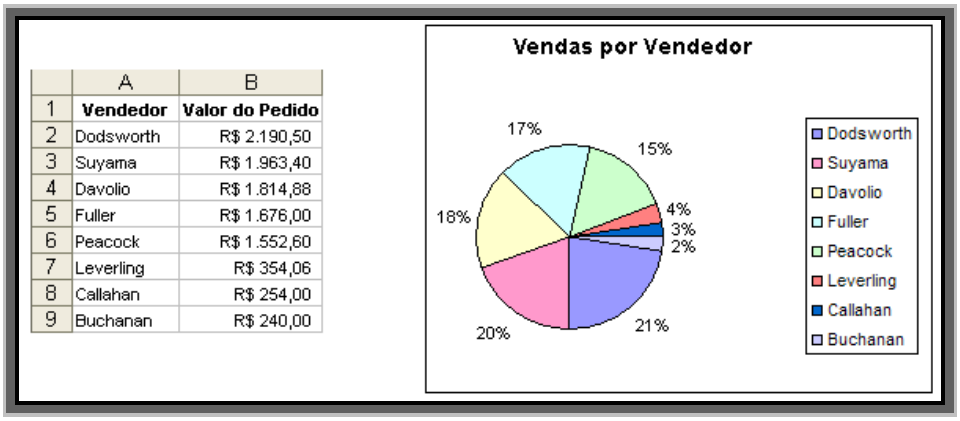


Figura 29 - Técnicas 2D e 3D comumente utilizadas

Os *Gráficos de Pizza*, por exemplo, são excelentes para a exibição de pontos de dados[9] como porcentagens de um todo. Entretanto, quando vários pontos de dados equivalem a menos de 5% da pizza, fica difícil distinguir as fatias. Por exemplo, um *Gráfico de Pizza* sobre os dados de vendas a seguir (figura 30) contém três fatias que ficam abaixo de 5%, dificultando a visualização e conseqüentemente a sua interpretação.

[9] **Pontos de dados**: valores individuais plotados em um gráfico e representados por barras, colunas, linhas, fatias de pizza ou rosca, pontos e diversas outras formas chamadas marcadores de dados. Os marcadores de dados da mesma cor constituem uma série de dados.



	A	B
1	**Vendedor**	**Valor do Pedido**
2	Dodsworth	R$ 2.190,50
3	Suyama	R$ 1.963,40
4	Davolio	R$ 1.814,88
5	Fuller	R$ 1.676,00
6	Peacock	R$ 1.552,60
7	Leverling	R$ 354,06
8	Callahan	R$ 254,00
9	Buchanan	R$ 240,00

Figura 30 - Representação por gráfico de pizza sobre dados de venda, evidenciando a dificuldade de interpretar os dados no caso de fatias pequenas

Neste Caso, para tornar as fatias menores mais visíveis em um *Gráficos de Pizza*, alguns recursos podem ser usados. O Excel, oferece os subtipos de *Gráficos de Pizza de Pizza* (figura 31) e de *Barra de Pizza*. Cada um desses subtipos separa as fatias menores do *Gráficos de Pizza* principal e as exibe em um *Gráficos de Pizza* adicional ou *de barras* empilhadas, conforme mostrado na próxima imagem.

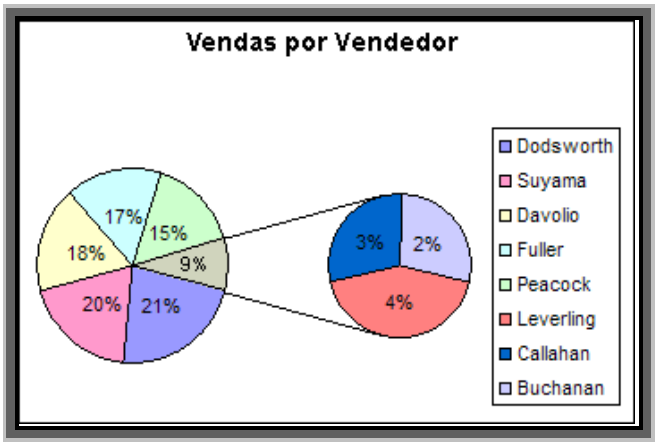


Figura 31 - Representação por gráfico de pizza de pizza dos dados de vendas.

Observe que os rótulos de percentagens no *Gráfico de Pizza* secundário exibem os mesmos números que o *Gráfico de Pizza* comum. Os números

representam apenas as fatias individuais que foram movidas para o gráfico secundário; eles não totalizam 100%.

Os *Gráficos de Barras* são gráficos nos quais os itens de dados são representados sob a forma de barras retangulares. As barras podem ser verticais ou horizontais, e assim como no *Gráfico de Pizza*, os dados podem se distinguir pela cor ou por algum tipo de sombreado ou padrão.

Dados qualitativos, particularmente quando as categorias são ordenadas, são usualmente bem ilustrados num simples *Gráfico de Barras* onde a altura da barra é igual à freqüência.

Quando as barras são apresentadas em três dimensões, então se dá o nome de *Cityscapes* à técnica. Os dados são mapeados nos atributos das barras e colocados no plano 2D horizontal (figura 32). Este conjunto usa como metáfora os “arranha-céus” de uma cidade. Estilos arquiteturais, cores, altura, transparência podem ser usados para representar os atributos nestes “arranha-céus”.

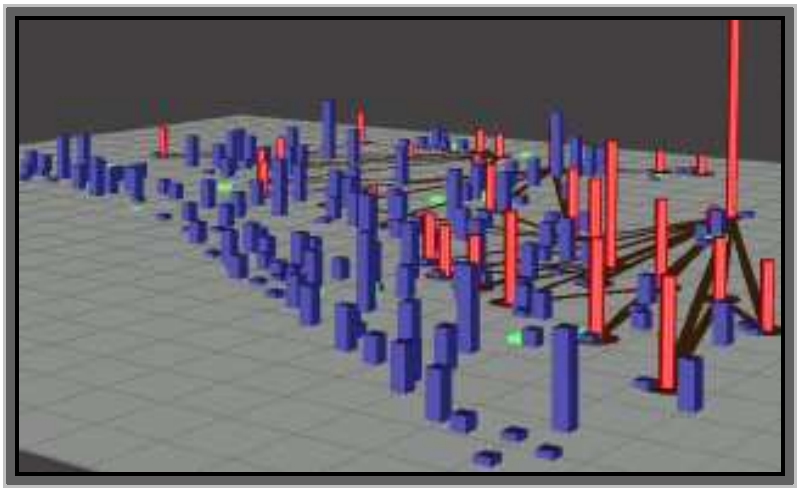


Figura 32 - Exemplo de uma visualização do tipo *Cityscape* (CHUAH *et al*, 1995; SANTOS; GROS; ABEL, 1999)

Santos, Gros e Abel (1999) afirmam que *Diagrama de Superfícies* é a representação 3D para os populares grafos bidimensionais. Os dados são traçados ao longo dos três eixos de coordenadas x, y, z; estes então são ligados

de modo a formarem uma malha que podem ser coloridas a partir de uma escala de cores conforme os valores dos atributos. Esta técnica permite facilmente identificar picos (valores mínimos e máximos), além de extrair padrões.

Outra técnica bastante conhecida que pode ser considerada como técnica tradicional é a *Paredes de Perspectivas*, desenvolvida por Mackinlay (1991) e usada para visualizar grandes volumes de dados ordenados ao longo de uma única dimensão.

Esta técnica surgiu a partir do amadurecimento da idéia proposta por Spence e Apperley (1982). Na técnica *Bifocal Display* (SPENCE; APPERLEY, 1982), os itens de informação são apresentados em três áreas distintas, sendo a central aquela que contém a informação em foco, em destaque, e as outras informações do contexto geral são apresentados nas laterais da região focal (figura 33).

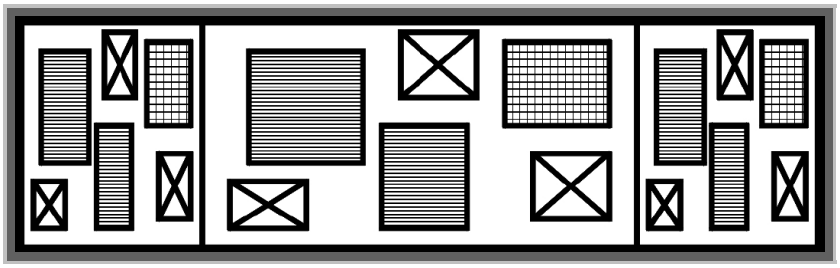


Figura 33 - Representação visual da técnica Bifocal Display

Nas *Paredes de Perspectivas* as informações são mapeadas no espaço bidimensional utilizando uma parede 3D (FREITAS *et al*, 2001). A figura 34 mostra o uso das paredes de perspectivas para representar arquivos de acordo com a data da última alteração.

As *Paredes de Perspectivas*, segundo Freitas *et al* (2001), embora permita ter uma visão global numa única visualização, além de integrar uma vista detalhada e reter o contexto da informação, uma das suas insuficiências está no fato de só conseguir lidar com bases de informação ordenada ao longo de uma única dimensão.

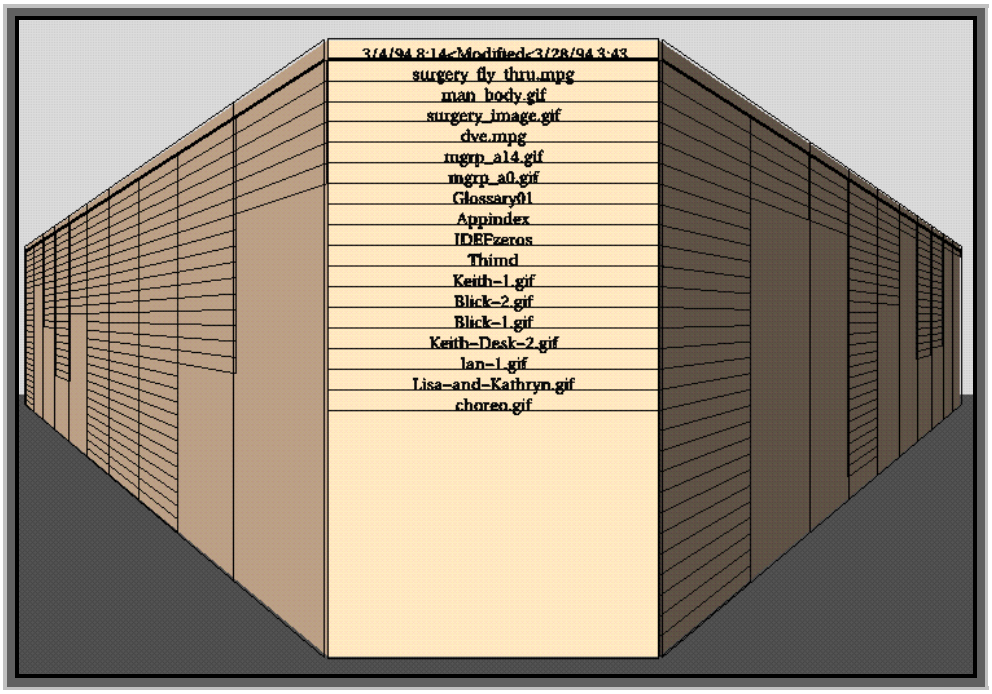


Figura 34 - Exemplo de uma parede de perspectiva (MUKHERJEA; FOLEY; HUDSON, 1995).

As *Técnicas Tradicionais 2D e 3D*, embora simples, algumas dessas, como os *Gráficos de Dispersão*, oferecem apoio eficaz para a análise visual, permitindo detectar distribuição, correlação entre os atributos, agrupamentos e outras informações.

Graficamente, a técnica *Gráficos de Dispersão*, plota o comportamento das variáveis no espaço bidimensional, permitindo analisar a dispersão dos dados. Quando estes dados estão dispersos aproximando-se de uma reta, diz-se que os variáveis são altamente correlacionáveis. Se esta reta for crescente, as variáveis possuem correlação positiva, caso contrário, se ela for decrescente, a sua correlação será próxima de “-1”, ou seja, serão inversamente correlacionáveis. Caso os dados estejam dispersos, não se aproximando de uma reta, então estes possuem correlações próximas de “0” identificando um baixo índice de relação entre as variáveis.

Em geral, problemas como os citados para o *Gráfico de Pizza*, embora possam ser contornados, são comuns para estes tipos de técnicas. Segundo

Artero (2005), uma outra desvantagem para estas técnicas está na limitação quanto ao número de atributos que podem ser apresentados simultaneamente.

5.2.2 Técnicas Orientadas a Pixels

Neste grupo, os atributos dos dados são mapeados em pixels coloridos, conforme os valores dos atributos representados. Cada atributo é apresentado em uma janela individual de forma que para exibir *m* atributos a janela deverá ser dividida em *m* janelas (KEIM; KRIEGEL, 1994; KEIM; KRIEGEL, 1996). Cada pixel da janela é a representação visual de um dos registros dos dados para um determinado atributo (figura 35). Este pixel é colorido conforme um mapa de cores previamente fixado de acordo com a faixa de possíveis valores do atributo.

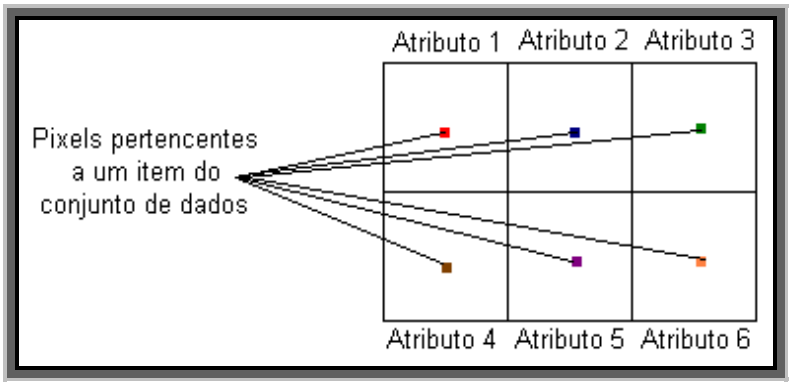


Figura 35 - Representação por janelas de 6 atributos de um item do conjunto de dados (KEIM, 2000)

Esta técnica, geralmente é usada para determinar padrões nos dados (*clusters*), ou correlações e dependência funcional entre atributos (KEIM, 1996; KEIM, 2000). Neste caso, uma análise de regiões correspondentes em atributos distintos deverá ser feita. A figura 36 exemplifica o caso da correlação existente entre as dimensões *MinAngle* e *RightAngle*.

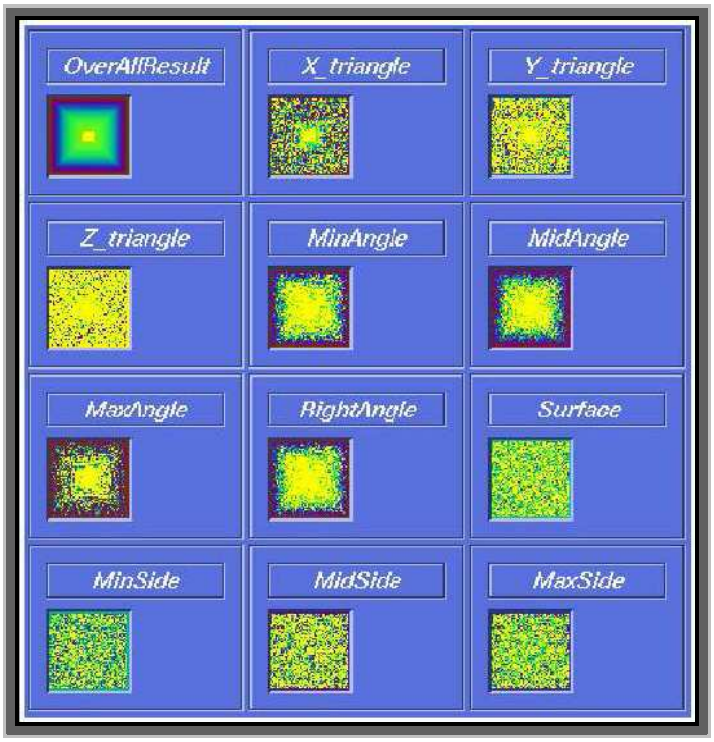


Figura 36 - Identificação de correlação e dependências funcionais no VisDB (KEIM, 1996)

Segundo Branco (2003), a principal desvantagem deste método é quando o número de atributos é muito grande. Isso se deve ao fato da dependência direta em relação à resolução da tela, pois quanto maior a dimensionalidade dos dados, maior será o número de janelas e, conseqüentemente, menor será o número de atributos que poderão ser vistos simultaneamente. Por outro lado, se um único atributo é representado em uma janela de resolução de 1280 x 1024, é possível exibir mais de um milhão de valores simultaneamente.

Keim e colaboradores têm tido um papel importante no desenvolvimento e aplicação de técnicas nesta categoria (KEIM; KRIEGEL, 1994; KEIM; KRIEGEL, 1996; KEIM, 2000; ANKERST; KEIM; KRIEGEL, 1996).

A técnica *Segmentos Circulares* (*Circle Segments*) é outra técnica bastante conhecida proposta por (ANKERST; KEIM; KRIEGEL, 1996). Esta técnica, conforme o próprio nome já diz, mapeia os dados em pixels coloridos em segmentos circulares. A figura 37 (a) mostra um arranjo para a distribuição dos

dados, em (b) mostra como estes dados são mapeados e em (c) a representação de um conjunto de dados contendo cinqüenta (50) atributos.

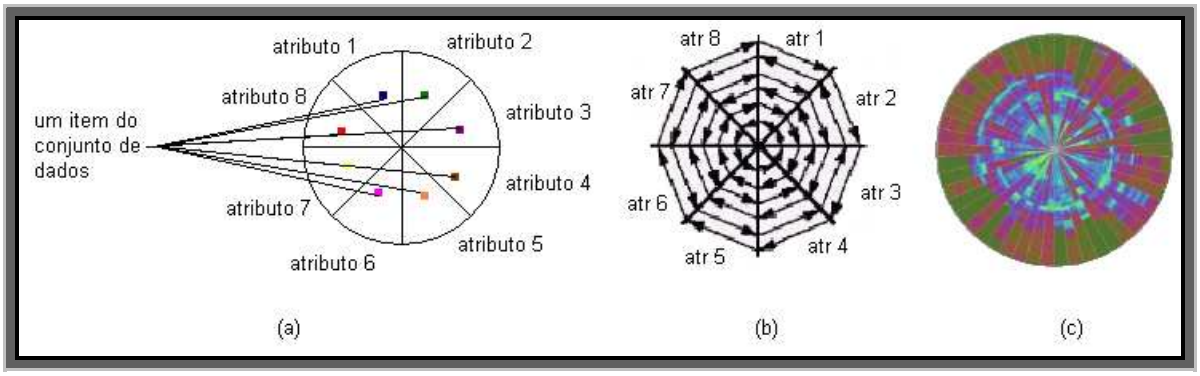


Figura 37 - Técnica segmentos circulares. (a) Distribuição dos dados. (b) Mapeamento dos dados. (c) Representação de um conjunto de dados (ANKERST; KEIM; KRIEGEL, 1996)

Esta técnica apresenta limitações para conjuntos de dados muito grandes, pois o número de segmentos aumenta à medida que aumenta a dimensão, diminuindo o espaço disponível para representação do conjunto de dados.

5.2.3 Técnicas de Projeção Geométrica

As *Técnicas de Projeções Geométricas* projetam os dados multidimensionais em um espaço bidimensional, buscando apresentar projeções interessantes dos conjuntos de dados. Em particular, uma técnica bastante utilizada desta categoria é a *Coordenadas Paralelas* (*Parallel Coordinates*), outra técnica como *Matrizes de dispersão* (*Scatterplot Matrices*), *Gráfico Estrela* (*Star Graph*), *Visualização Radial* (*Radial Visualization - RadVis*) e *Tubo de Dados* (*Data Tube*), também fazem parte deste grupo.

Matrizes de Dispersão é uma generalização para a técnica *Gráficos de Dispersão* (seção 5.2.1 - *Técnicas 2D e 3D Tradicionais*). Enquanto esta se preocupa com o mapeamento individual, aquela busca comparar diversos atributos simultaneamente dois a dois mapeando os dados para um espaço bidimensional. Informações como correlações e dispersões dos dados podem ser

extraídas neste tipo de visualização contribuindo para um maior entendimento dos dados além de apoiar no procedimento de redução de dimensionalidade.

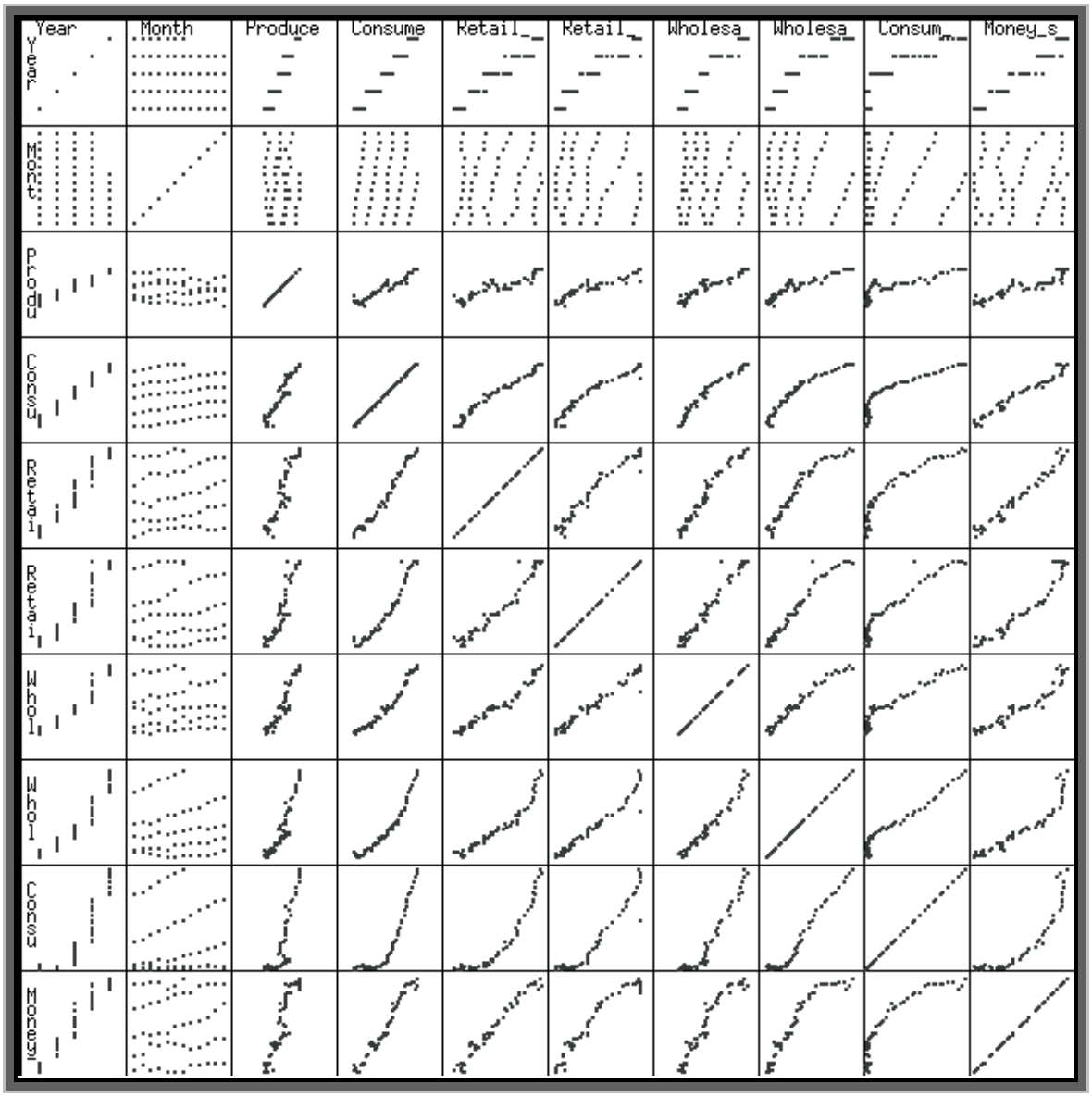


Figura 38 – Representação da técnica matriz de dispersão para um conjunto de dados de 10 atributos (WARD, *et al*, 2007)

A figura 38 mostra a aplicação da técnica matriz de dispersão para dados financeiros ao longo de cinco anos; observe-se na imagem que as variáveis consumo e produção são altamente correlacionáveis, ou seja, o aumento no consumo implica num aumento da produção. Esta conclusão pode ser extraída

pois no gráfico *Consumo x Produção* os dados estão dispersos se aproximando de uma reta crescente.

Assim como nas técnicas orientadas a pixels, visto na seção 5.2.2, a alta dimensionalidade prejudica a visualização dos dados, reduzindo a área para cada dispersão (*scatterplots*). Problemas como estes podem ser minimizados usando técnicas de interação como, por exemplo, o *zoom*.

Outra metodologia para visualização geométrica em *n*-dimensões para problemas multivariáveis bastante conhecida são as *Coordenadas Paralelas* (CARVALHO, 2001) cuja idéia inicialmente foi apresentada por Alfred Inselberg na Universidade de Illinois em 1959, que tem trabalhado nela desde então. É uma técnica de *Visualização* onde as dimensões são representadas como uma série de eixos paralelos uns aos outros e com igual espaçamento entre eles nos quais os valores estão representados (INSELBERG, 1999; INSELBERG; AVIDAN, 1999).

Artero (2005) define *Coordenadas Paralelas* como um espaço de dimensão *n* mapeado para um espaço bidimensional usando *n* eixos eqüidistantes e paralelos a um dos eixos principais. Cada eixo representa um atributo e, normalmente, o intervalo de valores de cada atributo é mapeado linearmente sobre o eixo correspondente. Cada item de dado é exibido como uma linha poligonal que intercepta cada eixo no ponto correspondente ao valor do atributo associado.

A figura 39 ilustra a aplicação da técnica *Coordenadas Paralelas* para os mesmos dados da figura 38. Nesta, selecionou-se os dados do ano 1 para analisar a variação dos valores ao longo das variáveis. Informações de relacionamento entre variáveis podem ser extraídas analisando pares consecutivos de atributos. Um grupo de linhas projetadas bastantes próximas uma das outras e sem muitos cruzamentos, indica um grau de relacionamento positivo entre as tuplas que as compõem (ver o comportamento das tuplas dos atributos produção e consumo, por exemplo).

Outra vantagem deste método de visualização é que a representação de todos os vetores em um mesmo gráfico nos permite efetuar comparações visuais entre vetores.

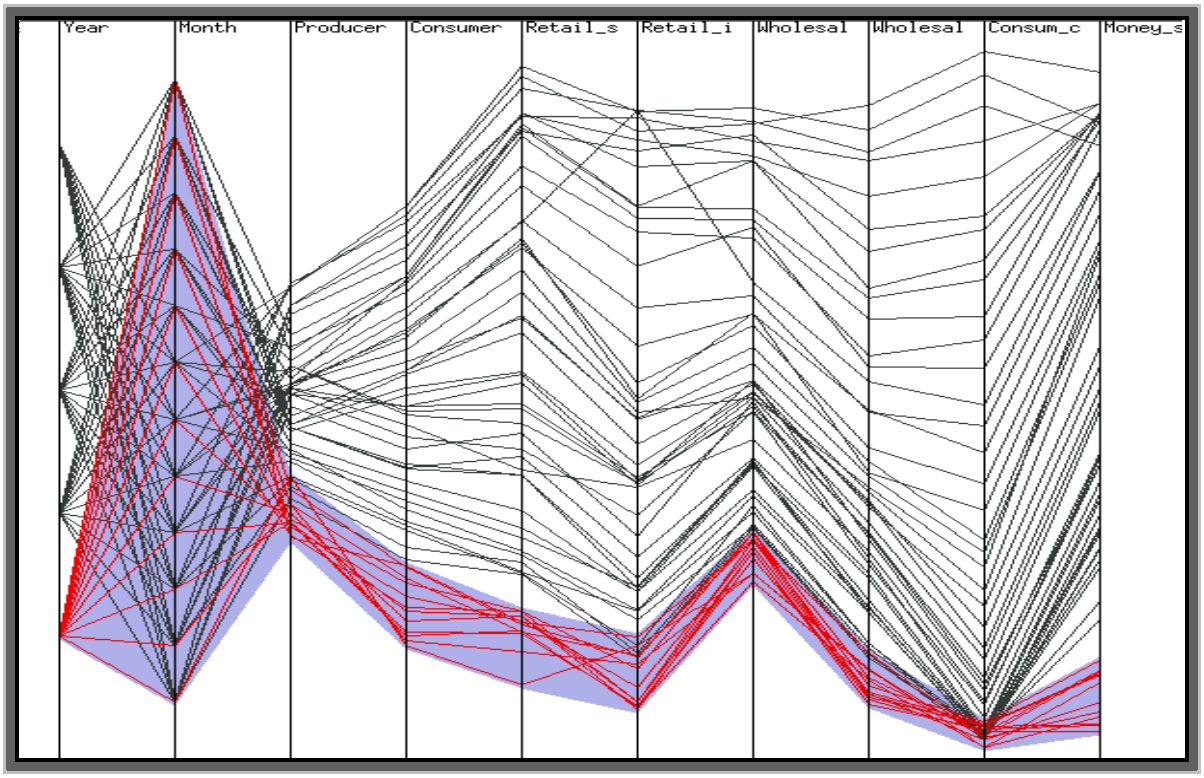


Figura 39 – Visualização por *Coordenadas Paralelas* do conjunto de dados financeiros ao longo de 5 anos, onde cada eixo é rotulado pelo nome correspondente à variável (WARD *et al*, 2007)

Esta técnica permite ainda encontrar *clusters*. A clusterização é facilmente visualizada quando conjuntos de dados saem de um mesmo ponto e seguem para as demais variáveis. A figura 40 mostra a clusterização para dados de dimensão 5D. Neste grupo foram encontrados sete *clusters* distintos pela coloração.

Uma desvantagem para esta metodologia está na representação de muitas variáveis, causando sobreposição de linhas dificultando a extração de qualquer tipo de informação, nem mesmo dedutiva, a respeito dos dados.

Outra desvantagem está na limitação da resolução horizontal da tela, ou seja, a medida que o número de dimensões cresce, os eixos vão se aproximando um dos outros dificultando a interpretação dos resultados.

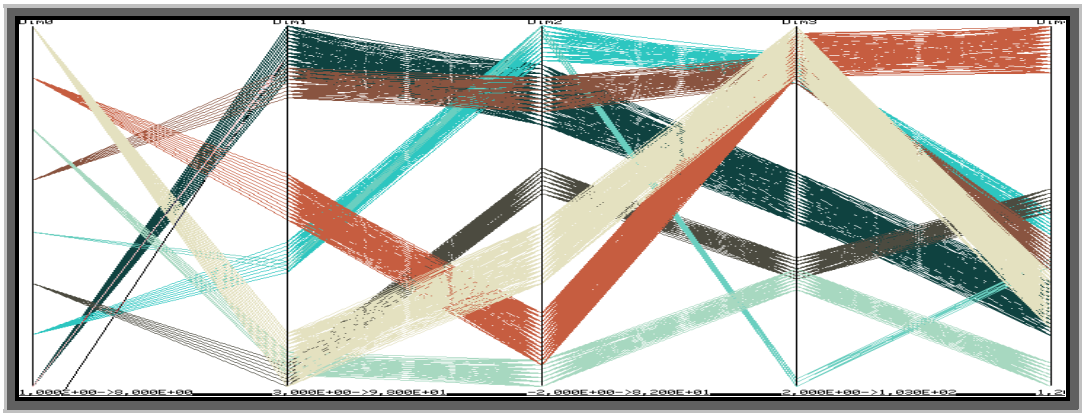


Figura 40 - *Coordenadas Paralelas* na análise de agrupamentos

Quando se deseja analisar correlações entre pares de variáveis, é necessário que estas estejam em seqüências, em alguns casos pode ser feita a ordenação dos eixos conforme os valores de suas correlações como proposto por Carvalho (2001) ou então ordená-los de forma interativa conforme intuição e necessidade do usuário, alguns programas como o ParVis permitem este tipo de interação.

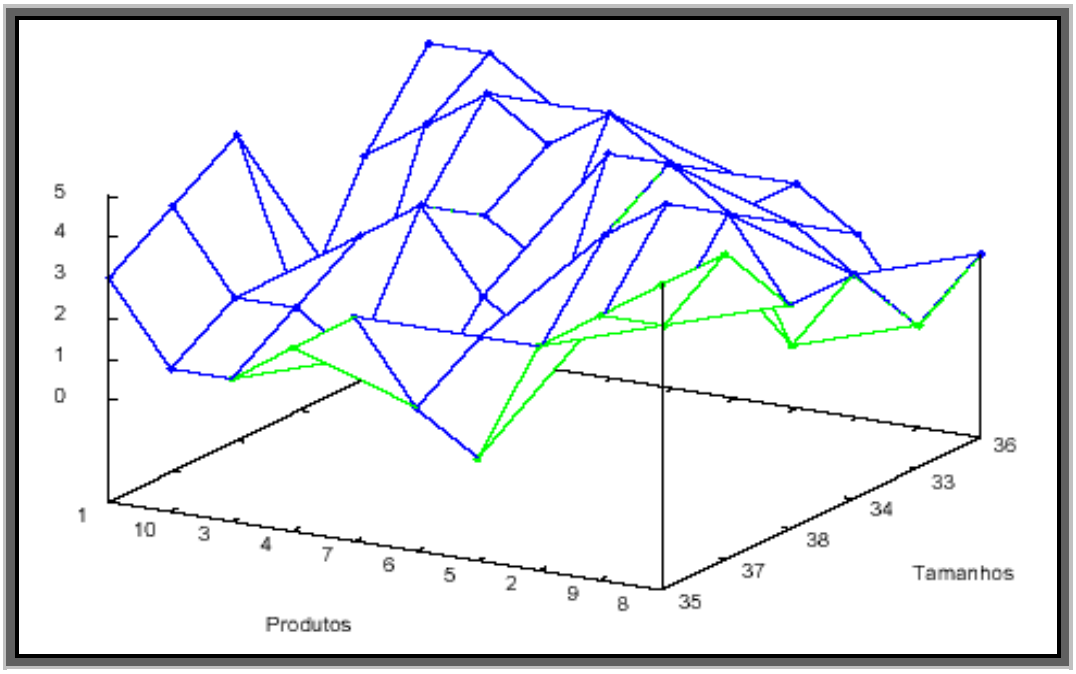


Figura 41 - Uso da técnica *Coordenadas Paralelas* em 3D (CARVALHO, 2001)

Alguns autores dedicam seus estudos às *Coordenadas Paralelas* de forma a aprimorar esta técnica. Carvalho (2001), por exemplo, propõe em seu trabalho o uso de *Coordenadas Paralelas* em três dimensões. Segundo o autor, o uso de uma dimensão a mais ajuda na interpretação dos dados, permitindo uma maior varredura das informações além de permitir técnicas de interação como rotação, *zoom* e *pan* (ver figura 41).

O *Gráfico Estrela* (SOBOL; KLEIN, 1989; PARSAYE; CHIGNELLl, 1993; HOFFMAN; 1999) é outra técnica inspirada em *Coordenadas Paralelas*, diferenciando desta pelo fato dos eixos serem arranjados em uma disposição radial conforme a figura 42 ao invés de serem paralelos. Assim como na *Coordenadas Paralelas*, os dados são representados por poligonais que interceptam cada eixo na posição correspondente aos valores dos atributos associados. Segundo Chau, Lin e Yeh (1999), quando o número de atributos é alto, os eixos ficam muito próximos na parte central, dificultando a análise.

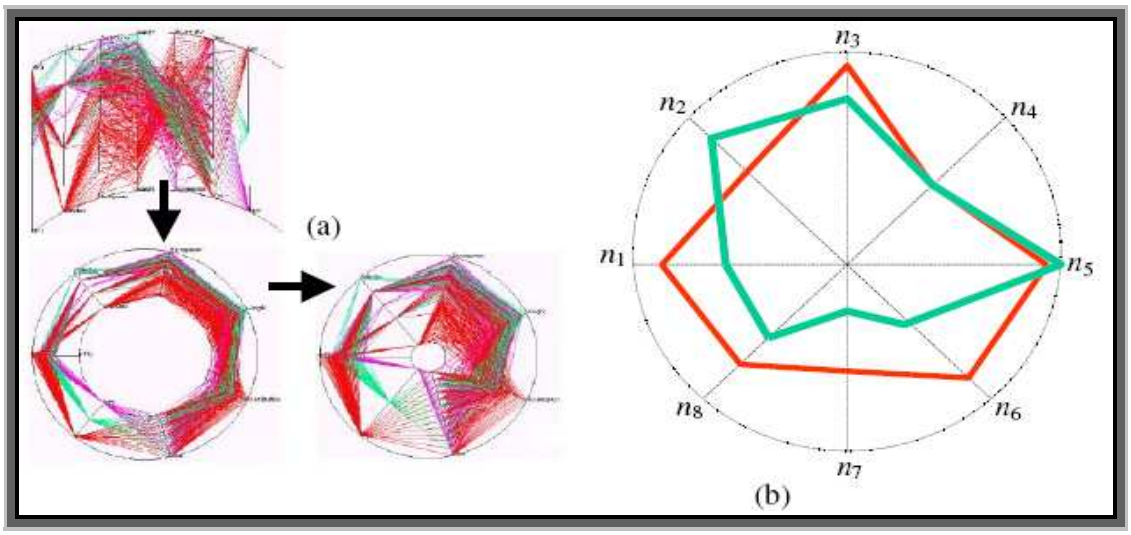


Figura 42 - (a) Obtenção da técnica *Gráfico Estrela* a partir da técnica *Coordenadas Paralelas* (Hoffman, 1999); e (b) Visualização de dois registros de dimensão oito utilizando o *Gráfico Estrela*

O *RadVis* (*Radial Coordinates Visualization*) (HOFFMAN, 1999) é outra técnica geométrica que também adota um arranjo radial. Para uma visualização *n*-dimensional, *n* linhas emanam radialmente do centro de um círculo e terminam no

seu perímetro, como ilustrado na figura 43. Para cada atributo, constantes de atração (um sistema imaginário de molas) são associadas aos valores, sendo que a posição final do marcador visual será aquela em que ocorre o equilíbrio das forças sobre o marcador. O mapeamento resultante constitui uma transformação não linear do espaço original, que preserva algumas simetrias.

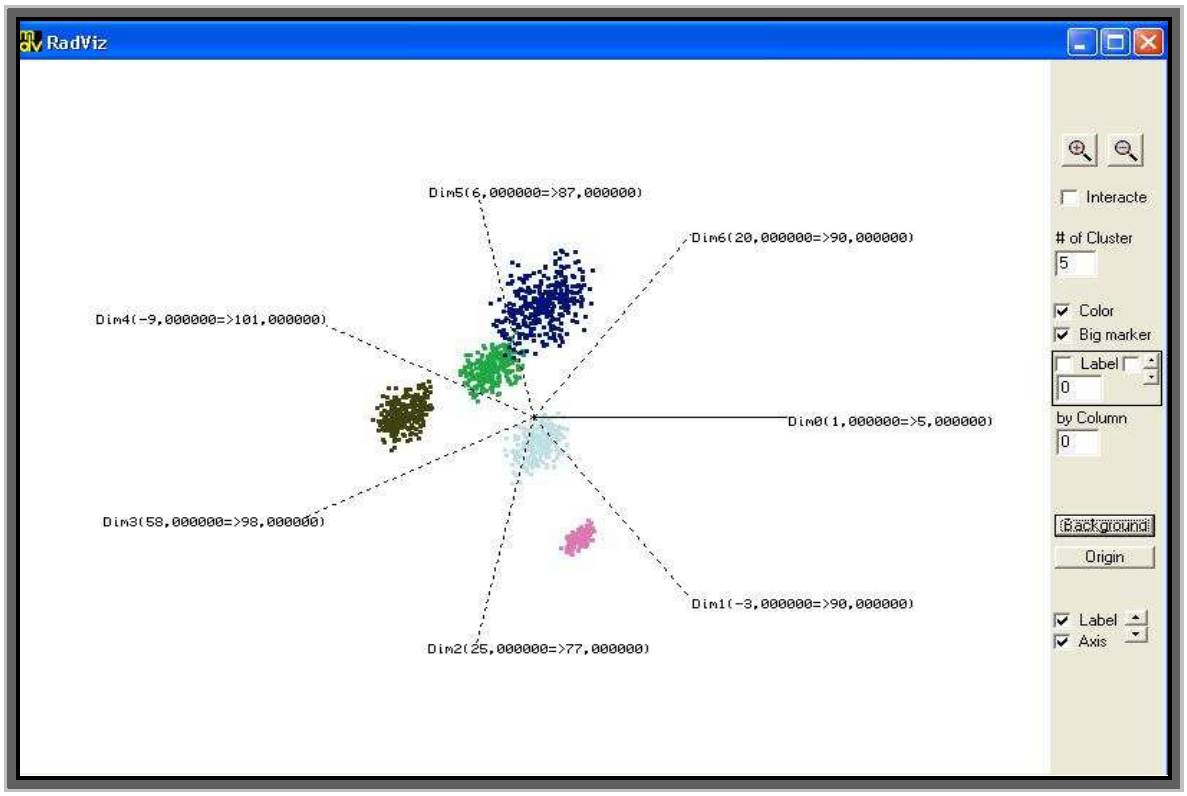


Figura 43 - Visualização de um conjunto de dados através da técnica *RadVis* (ARTERO, 2005)

Grinstein *et al* (2001) salientam que as principais características desta técnica são:

- registros cujos atributos têm valores iguais são posicionados no centro do sistema de eixos;
- registros em que um dos atributos tem valor dominante em relação aos demais são mapeados próximos ao eixo correspondente a este atributo;

- registros similares no espaço n-dimensional são mapeados próximos entre si no espaço 2D, o que favorece a identificação de agrupamentos na visualização;
- apresenta baixa complexidade computacional O(*mn*), permitindo a sua aplicação a grandes conjuntos de dados de alta dimensionalidade.

Entre as desvantagens, pode-se citar a intensa sobreposição dos marcadores e o congestionamento visual excessivo quando aplicado a grandes conjuntos de dados. No caso da análise de agrupamentos, a principal desvantagem é que registros muito diferentes entre si podem ser mapeados em posições próximas, o que precisa ser tratado de algum modo em etapas posteriores.

Outras técnicas, como *Análise dos Componentes Principais* (PCA) (PEARSON, 1901), *FastMap* (FALOUTSOS; LIN, 1995) e *Vis3D* (ARTERO, 2005), estendem o *RadVis* para uma visualização no espaço tridimensional. A extensão em 3D consegue acomodar um maior número de marcadores e contornar problemas como os de oclusão, pois permite que o usuário interaja com o modelo de maneira a observar diferentes projeções nos dados. Para dimensões muito altas a sobreposição dos marcadores é inevitável, porém seus resultados são mais satisfatórios que os das técnicas de projeções em duas dimensões, como o *RadVis* (ARTERO, 2005).

Dada a matriz de dados D_{mxn}, a técnica *Viz3D* projeta os dados *n*-dimensionais na superfície e no interior de um cilindro 3D, onde os *m* registros d_i de D em coordenadas 3D (x_i,y_i,z_i) são mapeados seguindo a equação.

$$Viz3D_i(d_{i,0}, d_{i,1}, ..., d_{i,n-1}) = \begin{cases} x_i = x_c + \frac{1}{n}\sum_{j=0}^{n-1} \frac{d_{i,j} - \min_j}{\max_i - \min_j} \cos(\frac{2\pi j}{n}) \\ y_i = y_c + \frac{1}{n}\sum_{j=0}^{n-1} \frac{d_{i,j} - \min_j}{\max_i - \min_j} sen(\frac{2\pi j}{n}) \\ z_i = z_c + \frac{1}{n}\sum_{j=0}^{n-1} \frac{d_{i,j} - \min_j}{\max_i - \min_j}, \quad i = 0,...,m-1 \quad e \quad j = 0,...,n-1 \end{cases} \tag{8}$$

sendo: x_c, y_c *e* z_c as coordenadas do centro de um sistema de eixos radiais; max_j = *Máximo*(d_k, *j)* e min_j = *Mínimo*(d_k, *j)* para *k = 0, ..., m-1*;

A figura 44 mostra como é feito o mapeamento destes dados.

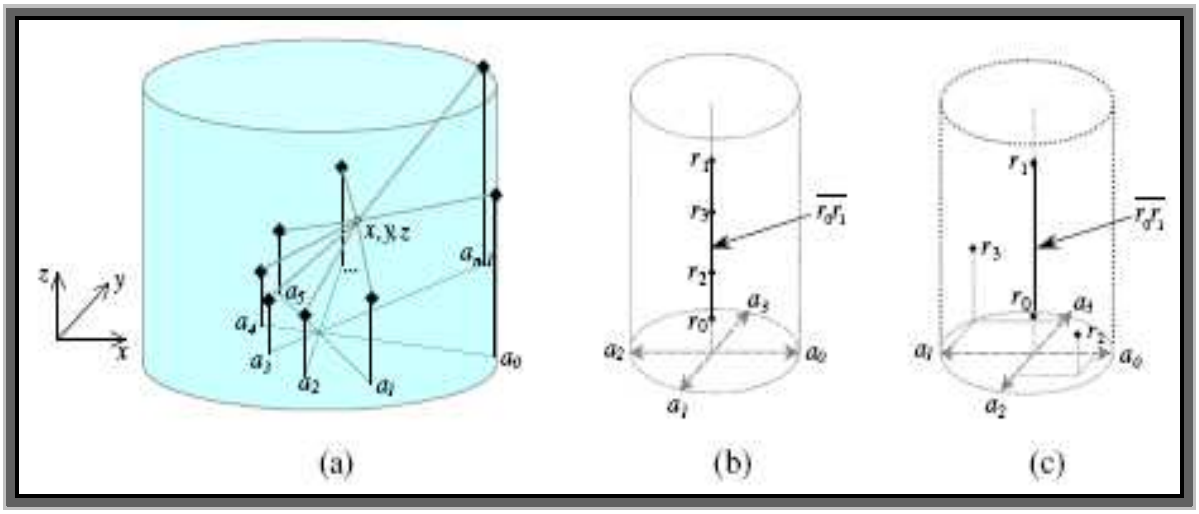


Figura 44 – (a) Projeção 3D no *Viz3D*; (b) Mapeamento dos registros r0, r1, r2 e r3 (dimensionalidade quatro) no Viz3D, adotando a seqüência de eixos a0, a1, a2 e a3; (c) Mapeamento com a seqüência de eixos a0, a2, a1 e a3.

Segundo Artero (2005), em seus testes, a técnica *Vis3D* mostrou um melhor desempenho quando comparadas às técnicas *FastMap* e *PCA*. Os testes foram aplicados para dados de dimensionalidade 11 com 10.000, 20.000 e 30.000 registros e comparados entre si pelo tempo de execução para gerar as projeções.

A figura 45 mostra o uso da técnica Vis3D na análise de agrupamentos. A imagem foi gerada a partir do programa MDV (ARTERO, 2005), que possibilita o usuário a formar os *clusters* de forma interativa e visual, selecionando regiões das quais, intuitivamente, pertence a um determinado grupo, sem usar algoritmos matemáticos ou estatísticos para este propósito.

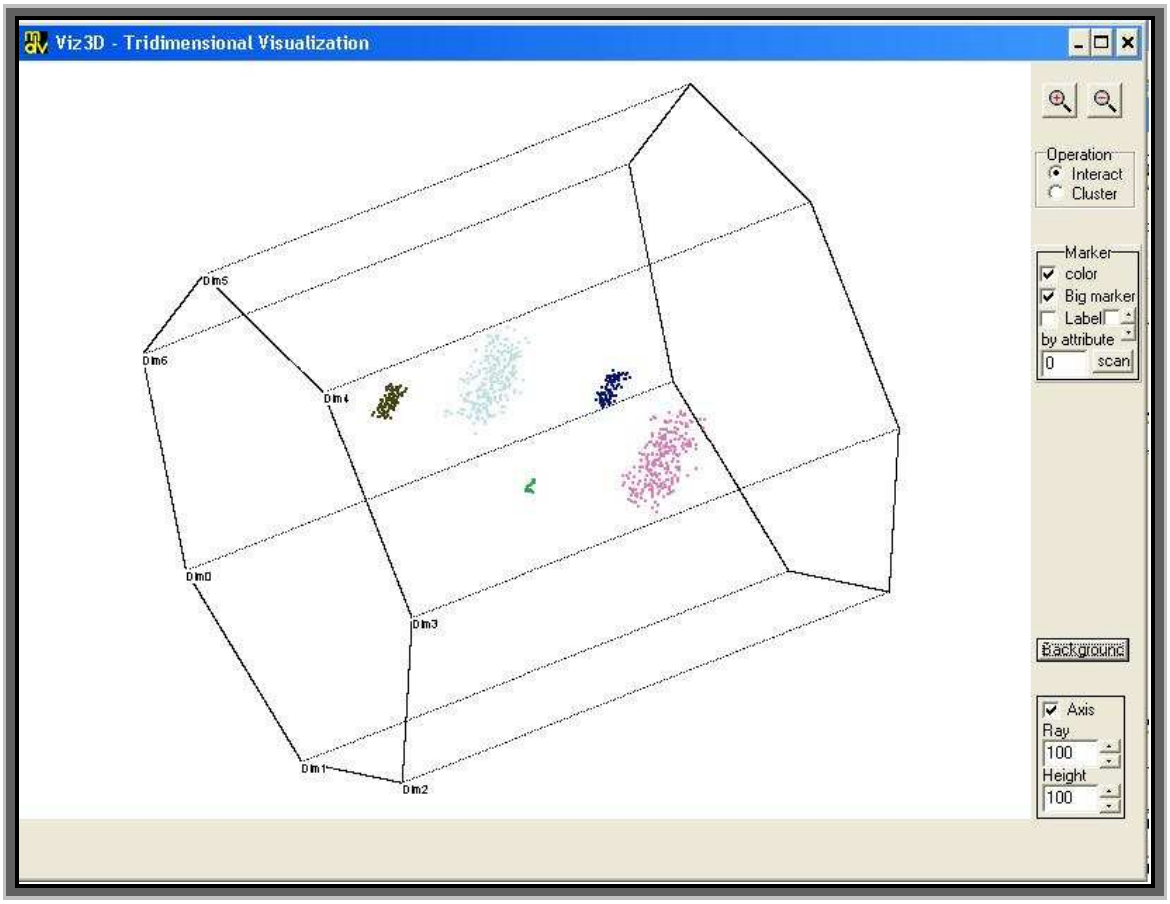


Figura 45 - Análise de *cluster* a partir da técnica *Vis3D*. Aqui cinco agrupamentos são observados (ARTERO, 2005)

A técnica *Star Coordinates* (KANDOGAN, 2001) estende o *RadVis*, permitindo ao usuário controlar a direção e o comprimento dos eixos radiais, bem como selecionar faixas de interesse sobre os eixos. A possibilidade de manipular interativamente a disposição e o mapeamento dos eixos apóia a busca por uma projeção mais adequada. Entretanto, a interação é dificultada quando o número de atributos é muito alto, o que torna impraticável a tarefa de encontrar uma boa configuração para os eixos. A figura 46 mostra a visualização de um conjunto de dados usando o *Star Coordinates.*

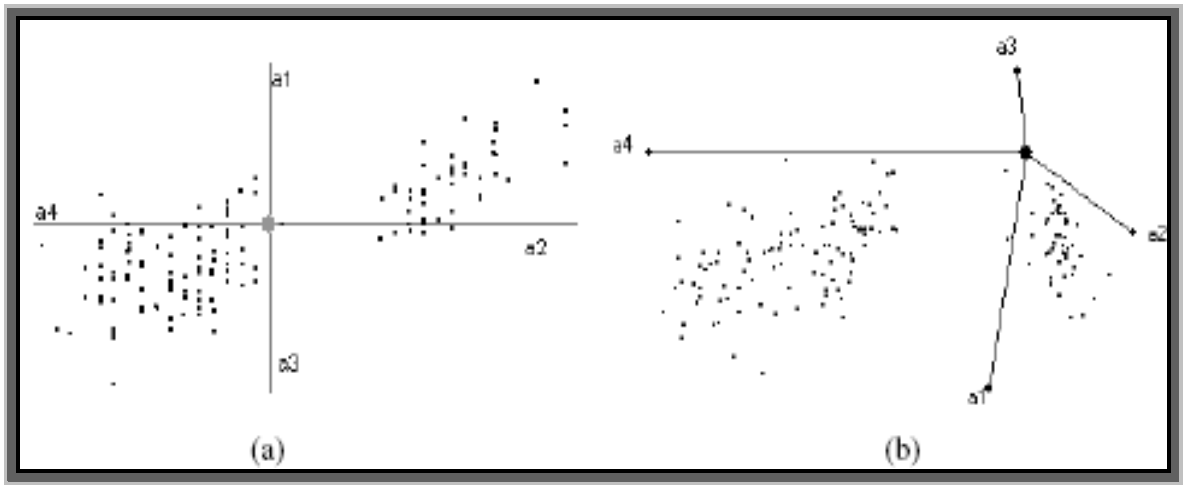


Figura 46 – (a) Visualização de um conjunto de dados com a técnica *Star Coordinates;* *(b) Visualização obtida após interação do usuário com os eixos*

Ankerst (2000) propõe uma técnica denominada *Tubo de Dados* (*Data Tube*) para a visualização de grandes conjuntos de dados, a qual é inspirada na técnica *Segmentos Circulares*. No *Tubo de Dados* os valores dos atributos são projetados no interior de um tubo (3D), conforme disposição apresentada na figura 47(a), sendo que o comprimento do tubo é determinado pelo número de registros do conjunto de dados. A visualização de um conjunto com seis atributos é ilustrada na figura 47(b).

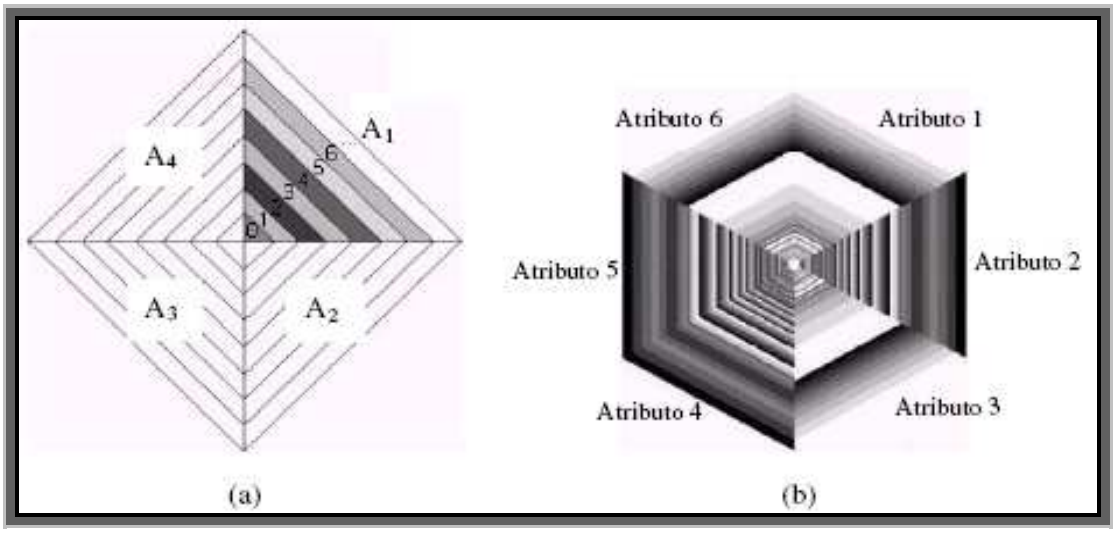


Figura 47 – (a) Disposição dos dados no *Tubo de Dados*; (b) Visualização de alguns registros de um conjunto de dados com seis atributos (ANKERST, 2000)

A técnica permite ao usuário navegar pela representação tridimensional, entrando e saindo do interior do tubo. A principal vantagem sobre a técnica *Segmentos Circulares* é a capacidade de tratar conjuntos de dados muito grandes, pois todos os registros podem ser apresentados ao usuário durante a navegação pelo seu interior.

5.2.4 Técnicas Iconográficas

Este conjunto de técnicas tem como objetivo mapear os atributos em características particulares de ícones. Cada característica do ícone representa um atributo dos dados multidimensional.

Chernoff (1973) mapeou atributos dos dados em características faciais com a intenção de utilizar conhecimentos comuns, uma vez que as pessoas estão habituadas a distinguir expressões faciais na vida diária. Dois atributos são mapeados para as duas dimensões espaciais da tela, e os demais são mapeados para propriedades visuais de um ícone na forma de uma face estilizada, como o formato da boca, nariz, olhos, etc. Valores diferentes nos atributos resultam em diferentes formas e posições dos componentes. A figura 48 exemplifica o uso da técnica *Faces de Chernoff* ao longo do tempo.

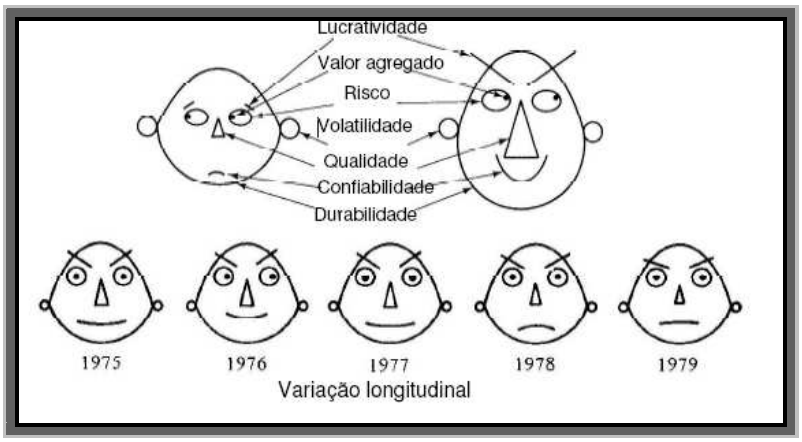


Figura 48 - Uso da técnica *Faces de Chernoff* para representação longitudinal de 8 atributos

O problema das *Faces Chernoff* é que embora sejam bastante úteis para mostrar tendências em dados multidimensionais, os valores dos dados propriamente ditos têm que ser apresentados adicionalmente, uma vez que aquelas não transmitem qualquer informação sobre os reais valores com as quais se relacionam. No entanto, a capacidade de ilustrar tendências não é desprezível, dado que pode ser usada, por exemplo, para ilustrar sobre que parte dos dados a atenção deve ser focada.

Essa abordagem explora a capacidade humana de reconhecer e analisar faces, mas vários autores (KEIM; KRIEGEL, 1996; CHOU; LIN; YEH, 1999) observam que, devido à dificuldade de distinguir diferenças muito pequenas nas imagens resultantes, ela não é adequada para a identificação de agrupamentos.

Ward (1994) usa outro ícone para fazer a representação dos dados. A técnica *Star Glyphs*, mapeia os dados em formas de estrelas onde cada atributo é representado por uma das pontas das estrelas, cujo tamanho é proporcional ao valor representado. A figura 49 exemplifica o uso desta técnica e mostra um comparativo entre seis diferentes tipos de automóveis onde cada ponta do ícone estrela representa diferentes atributos (aceleração, deslocamento, potência, MPG, peso).

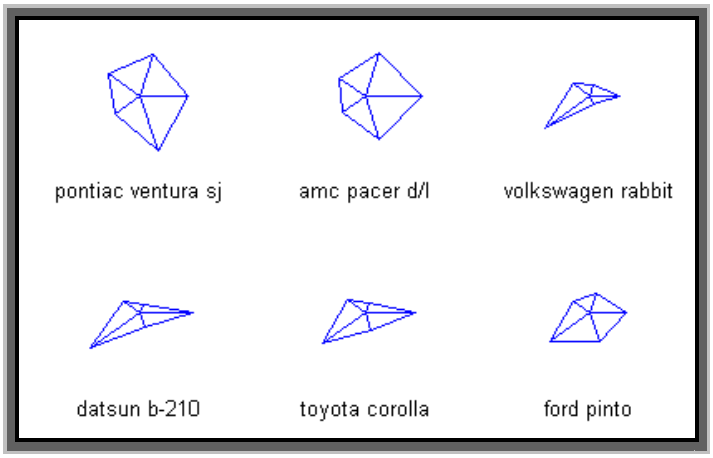


Figura 49 - Uso da técnica *Star Glyphs* para representar diferentes características de diferentes automóveis

Stick Figures é uma técnica que utiliza as duas dimensões da tela para mapear duas dimensões dos dados e as demais dimensões são mapeadas para os ângulos e/ou comprimentos de segmentos de um ícone formado por múltiplos segmentos de reta (PICKETT; GRINSTEIN, 1988). A figura 50(a) apresenta um ícone com uma configuração que apresenta cinco variáveis, na qual uma dimensão é mapeada pela inclinação do corpo do ícone, e a orientações das varetas permitem mapear outras quatro dimensões. Uma família de *Stick Figures* é ilustrada na figura 50(b), em que cada uma tem um corpo e quatro segmentos. Segundo Branco (2003), outras formas de representar dimensões nesses ícones seriam por meio da variação de cores e espessuras das varetas.

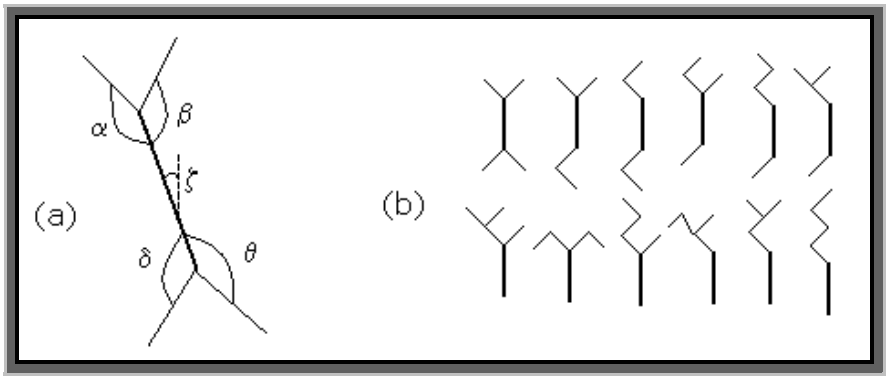


Figura 50 - *Stick Figures*. (a) Ícone representando cinco variáveis; (b) família de *Stick Figures* (WONG; BERGERON , 1997)

Quando mapeados na tela, os ícones (um para cada item de dado) formam texturas que variam de acordo com as características dos dados, permitindo identificar padrões na imagem que podem indicar dependência funcional entre os atributos visualizados (KEIM; KRIEGEL, 1996; WONG; BERGERON, 1997). A figura 51 exibe a imagem formada por esta técnica que representam cinco variáveis, exemplificando como essas texturas podem ser formadas.

Figura 51 - Uso da técnica *Stick Figures* no mapeamento de cinco variáveis (ANKERST, 2001)

Embora esta técnica permita o mapeamento de grandes quantidades de dados, a sua alta dimensão pode prejudicar na extração de informações na textura formada. Além disso, o reconhecimento de um importante padrão na imagem depende da seleção de um mapeamento adequado dos parâmetros dos dados para os parâmetros visuais. Segundo Wong (1997), o grande gargalo neste processo de visualização está neste número de possíveis mapeamentos visuais que crescem em ordem fatorial em relação ao número de dimensões mapeadas.

5.2.5 Técnicas Hierárquicas / Grafos

As técnicas hierárquicas, geralmente são aplicadas em dados cuja própria natureza apresenta uma correlação explícita entre níveis e subconjuntos (por exemplo, diretórios em sistemas de arquivos). Sendo assim, o espaço *n*-dimensional dos dados (não necessariamente hierárquicos) é então dividido em

subespaços que são organizados uns dentro dos outros e exibidos de forma hierárquica.

As técnicas *Cone Trees* e *Cam Trees* (ROBERTSON; MACKINLAY; CARD, 1991) são a representação tridimensional das populares árvores 2D. Estas técnicas se diferenciam apenas pela posição da árvore de cones formada, enquanto nesta a representação é feita na horizontal, naquela, esta representação é na vertical.

A construção da árvore é feita a partir de um nó raiz localizado no vértice de um cone. Todos seus filhos são então igualmente espaçados e posicionados na base deste cone. Este processo se repete para cada nó da árvore que possui filhos. Em cada nível a altura e o diâmetro da base dos cones são então recalculados para que toda a informação esteja visível (ver figura 52). (SANTOS; GROS; ABEL, 1999; FREITAS *et al*, 2001)

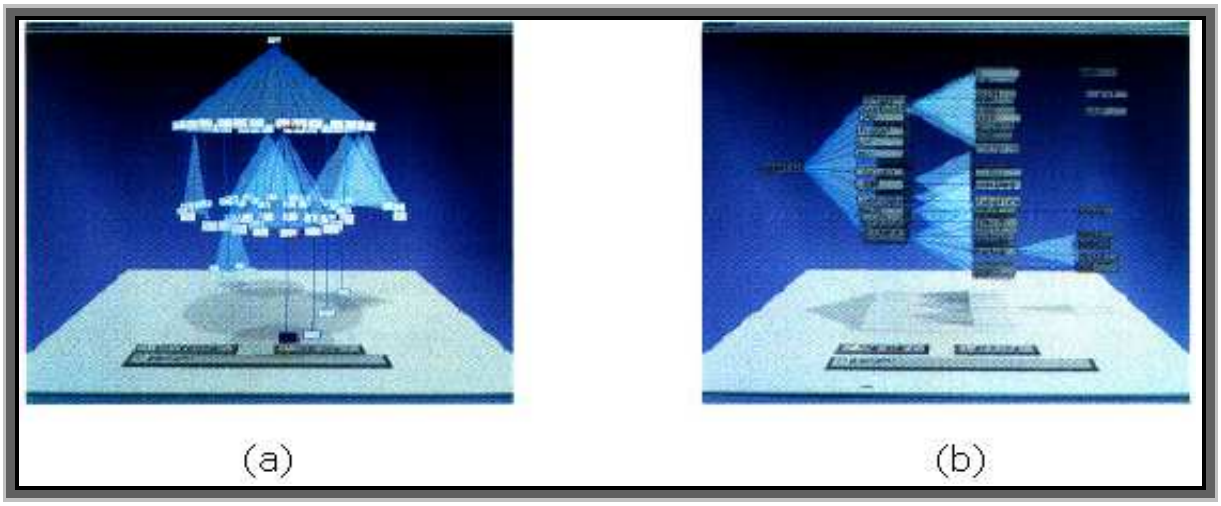


Figura 52 - Técnicas hierárquicas de visualização (a) *Cone Tree* e (b) *Cam Tree* (ROBERTSON; MACKINLAY; CARD, 1991)

Recursos tridimensionais como rotação, *zoom* e *pan,* permitem acesso rápido às informações com boa orientação para a visualização. Estes recursos permitem, por exemplo, escolher um nó que se deseja examinar mais detalhadamente de modo que o cone do nó escolhido seja apresentado mais a frente. Embora estes recursos possam ser aplicados, eles não impedem a

oclusão de nós. Para solucionar este problema uma variante desta técnica o *Recunfigurable Disc Tree* usa discos ao invés de cones permitindo que todos os nós dos cones sejam exibidos (JEONG; PANG, 1998).

Uma abordagem diferente, utilizando o espaço de tela para representar elementos de informação, ao invés de utilizar objetos geométricos, foi adotada por Johnson e Schneiderman (1991), com a técnica *Treemap*. A técnica surgiu da necessidade de saber como os arquivos estavam sendo usados e armazenados pelo grupo de estudos da Universidade de Maryland (SCHNEIDERMAN *et al*, 2007).

O *Treemap* mapeia as informações dividindo toda a tela do computador em partes para representar os diretórios e subdividindo estas partes para representar os subdiretórios e assim por diante (ver figura 53).

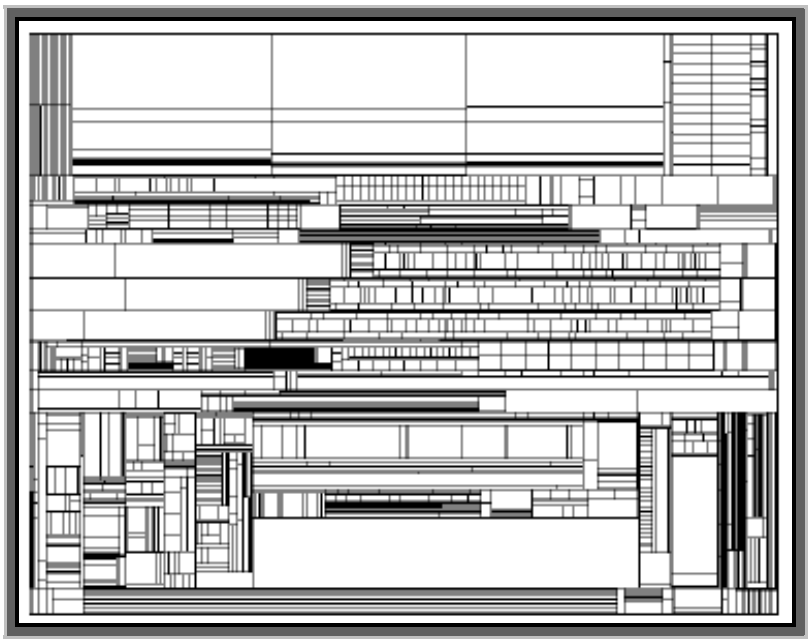


Figura 53 - Uso da técnica Treemap no mapeamento de diretórios de computadores (SCHNEIDERMAN *et al*, 2007)

Uma desvantagem do método *Treemap* está na dificuldade de identificar visualmente os diferentes níveis do *Treemap* quando a hierarquia se torna muito profunda.

Segundo Freitas *et al* (2001), esta técnica deu origem a outras como *Cushion TreeMaps* (VAN WIJK; VAN DE WETERING, 1999), *Information Slices* (ANDREWS; HEIDEGGER, 1998) e a *Interface do Sunburst* (STASKO; ZHANG, 2000). Estas técnicas, em geral, buscam facilitar a identificação dos diferentes níveis do *Treemap*.

Em uma tentativa de realçar a satisfação do usuário com a aparência dos *Treemaps* e a intuição do usuário sobre a informação da hierarquia a fim de facilitar a identificação destes níveis, Wijk e Wetering (1999) desenvolveram uma técnica denominada por *Cushion Tremaps*. Conforme pode ser observado na figura 54, o algoritmo do *Cushion Treemap* usa a iluminação dos pixels e uma função de altura para representar os diversos níveis da hierarquia. Esta função de altura é acionada para destacar retângulos menores.

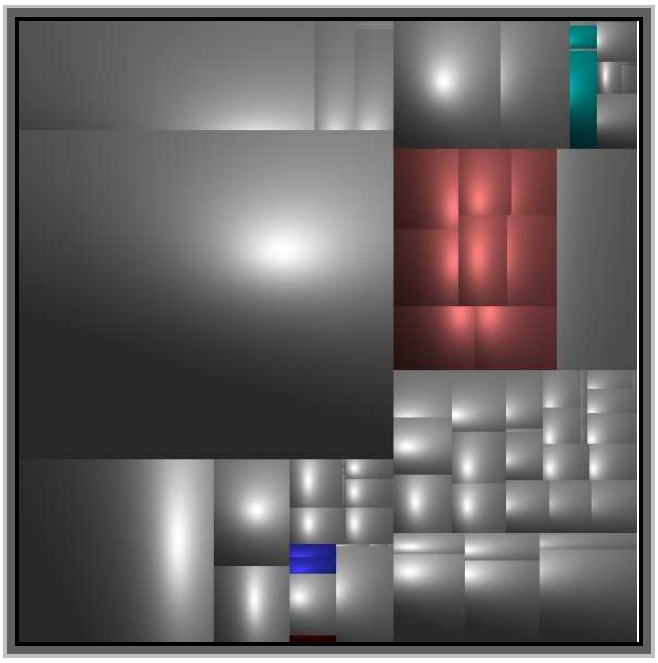


Figura 54 - Uso da técnica Cushion Treemaps. Iluminação e cores são usados para diferenciar os níveis dos diretórios (VAN WIJK; VAN DE WETERING, 1999)

A técnica *Information Slices* usa discos semicirculares (figura 55) para visualizar hierarquias. Esta técnica permite que dados sejam vistos de forma compacta com vários níveis em duas dimensões (ANDREWS; HEIDEGGER,

1998). Cada disco representa uma hierarquia de múltiplos níveis. Em cada nível da hierarquia, os filhos são dispostos de acordo com os valores dos dados. Para hierarquias maiores uma série de discos em cascatas podem ser usadas. Um segundo semicírculo pode ser usado para representar níveis com mais detalhe.

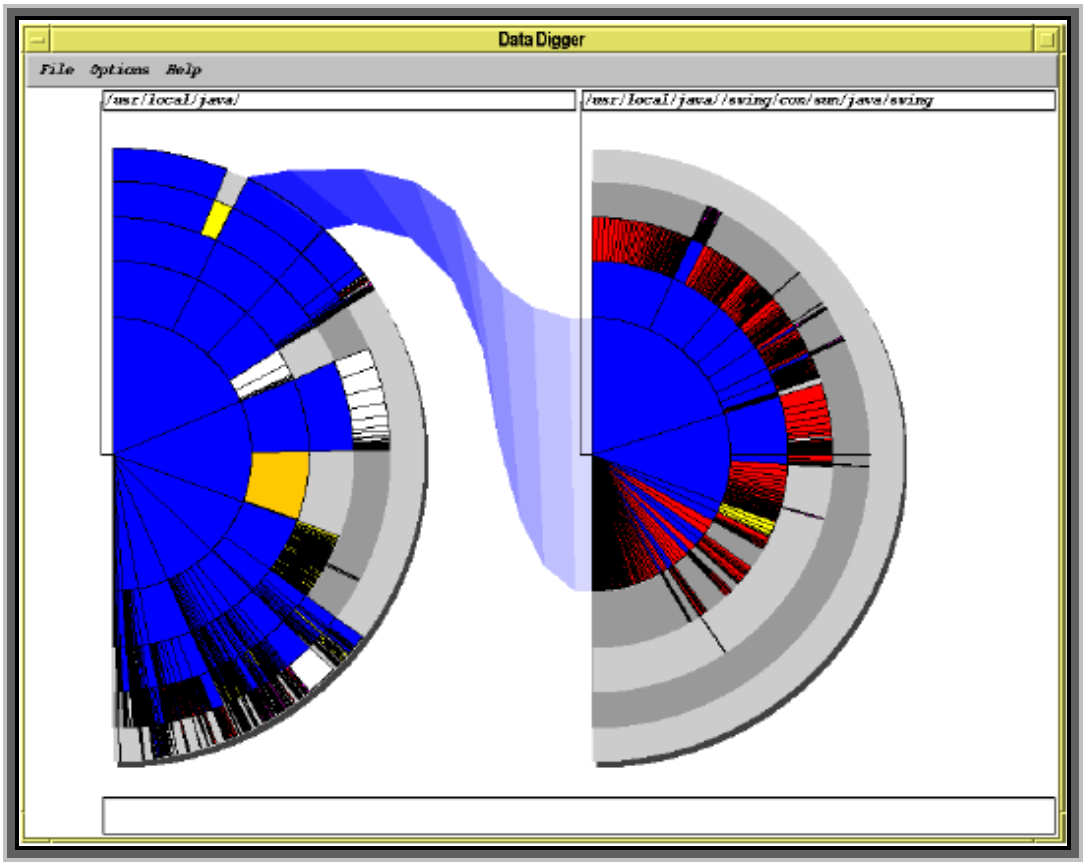


Figura 55 - Uso da técnica *Information Slices* mostrando semicírculo auxiliar para apresentar níveis com mais detalhes (ANDREWS; HEIDEGGER, 1998)

Na *Interface do Sunburst*, Stasko e Zhang (2000) usam discos completos, seguindo basicamente o mesmo conceito do *Informaton Slices*.

Na técnica *Empilhamento de Dimensão* (*Dimensional Stacking*) (LE BLANC; WARD; WITELLS, 1990), o espaço *n*-dimensional dos dados é subdividido em espaços bidimensional. Segundo Keim, Kriegel (1996) e Wong e Bergeron (1997), esta técnica não exige funções ou regras extras para plotar os dados, ao contrário das demais técnicas hierárquicas.

Um esquema conceitual para está técnica está representado na figura 56 onde quatro atributos estão sendo mapeados. A figura 57 exemplifica o uso da técnica aplicado a botânica, onde as cores representam um tipo de flor (em alguns casos a classificação pode ser mista), no eixo-x está representado o comprimento das pétalas e no eixo-y o comprimento das sépalas (quadrados menores), seguindo a mesma orientação dos eixos (quadrados maiores) estão as medidas de altura das mesmas partes da flor.

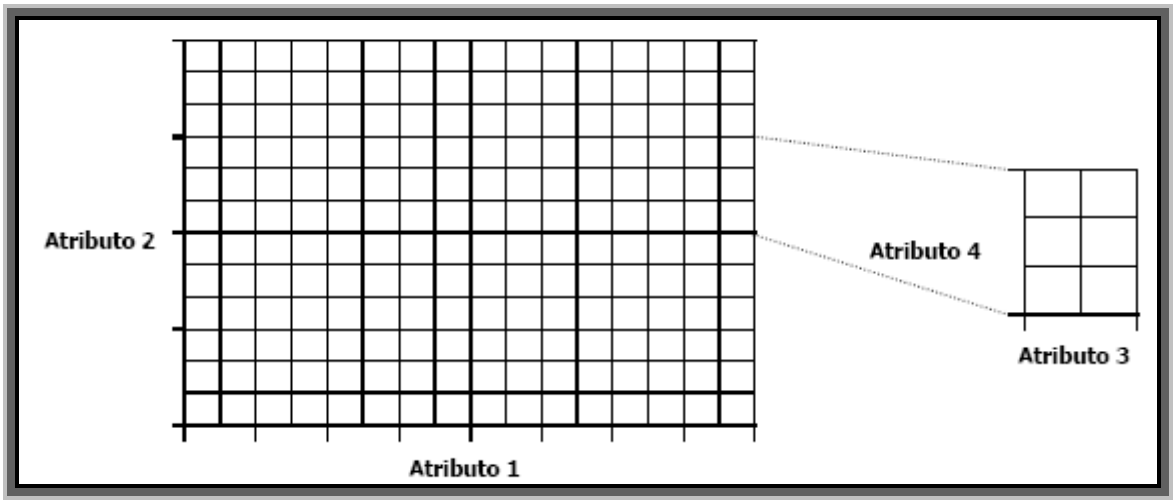


Figura 56 - Modelo conceitual da Empilhamento de Dimensão (Ankerst, 2001)

Segundo Hoffman e Grinstein (1999), esta técnica pode ser usada para determinar agrupamentos de pontos com comportamento discrepantes, e de padrões. Branco (2003), afirma que a interpretação dos resultados se torna muito difícil quando o número de atributos é muito grande, segundo o autor acima de nove atributos esta extração de informação já é bastante prejudicada.

Uma outra limitação a está técnica é o alto número de valores que um determinado atributo poderá assumir. O arranjo hierárquico dos atributos e a categorização dos dados devem ser levados em consideração, em geral, atributos de maior importância deverão ficar em níveis mais externos (KEIM; KRIEGEL, 1996; WONG; BERGERON, 1997; WARD, 1994).

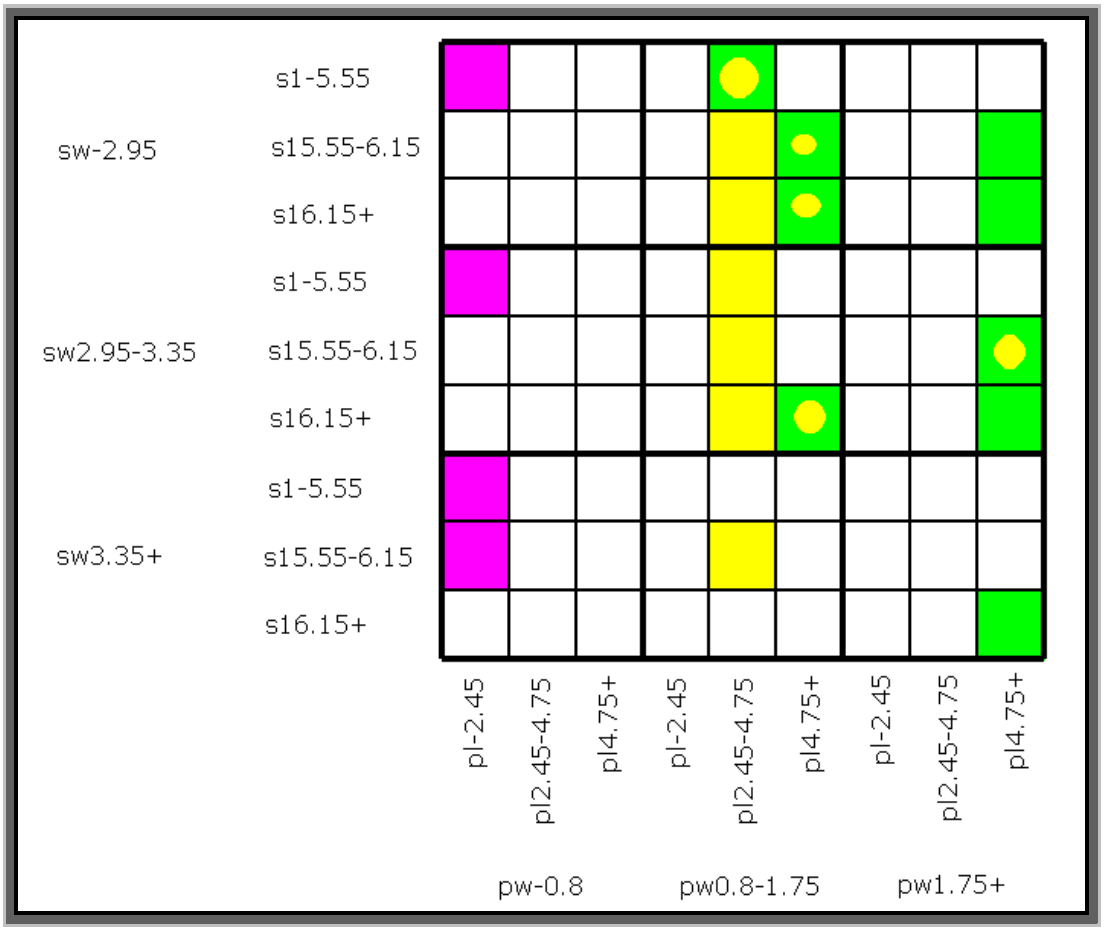


Figura 57 - Empilhamento de Dimensões aplicado à botânica, as três cores designam os três tipos de flores, em alguns casos a classificação é mista (HOFFMAN; GRINSTEIN, 1999)

Mundos dentro de Mundos (*Worlds-within-Worlds*) (BESHIERS; FEINER, 1993) é uma outra técnica baseada em hierarquias. Esta técnica mapeia mundos tridimensionais dentro de mundos tridimensionais. A função de cinco variáveis f(x1, x2, x3, x4, x5) (6 atributos) exemplifica como esta técnica pode ser aplicada. A figura 58 mostra uma superfície definida pelo valor da função f(x1, x2) (mundo interno) cuja representação está em função das demais variáveis (x3, x4, x5) (mundo externo).

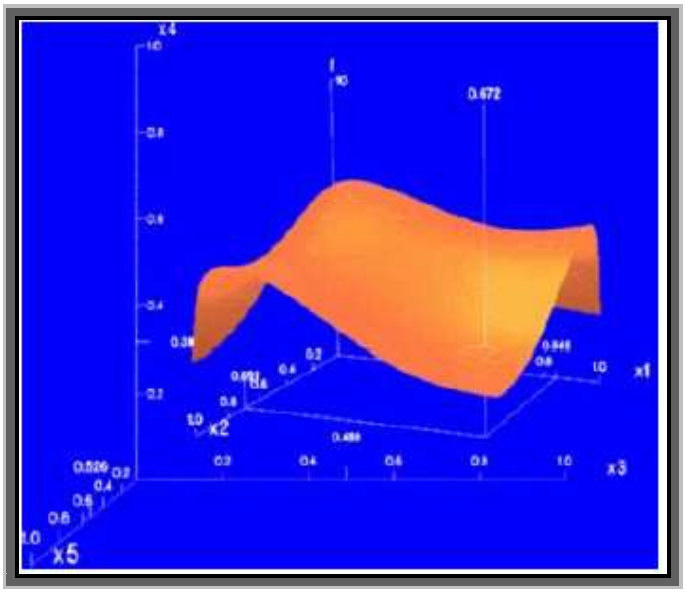


Figura 58 - Dados de dimensionalidade 6 mapeados no espaço tridimensional através da técnica *Mundo dentro de Mundos.* No caso as variáveis x_3, x_4 e x_5 são mantidas constantes (BESHIERS; FEINER, 1993)

Em se tratando de grafos, a idéia é visualizar dados volumosos usando técnicas que mapeiem as características (direcionado/não-direcionado, cíclico/acíclico, etc.) de um dado grafo, de maneira clara e rápida. As técnicas podem ser subdivididas segundo a dimensionalidade visual da representação: 2D ou 3D.

As visualizações geradas dependem de muitos fatores que refletem características inerentes aos grafos, muitas delas refletindo definições próprias da teoria dos grafos. Em relação ao desenho de grafos 2D tem-se como propriedades (existentes ou não): a planaridade (não cruzamento de linhas – arestas), a ortogonalidade (somente linhas ortogonais) e a propriedade de distribuição em grade (coordenadas dos vértices são números inteiros). As propriedades estéticas, objetivando a otimização são: número mínimo de cruzamentos, exibição ótima de simetria, exibição ótima de agrupamentos, número mínimo de curvas em grafos com poli-linhas, distribuição uniforme dos vértices, e comprimento uniforme das arestas. A figura 59 exemplifica o uso das técnicas para representação de grafos em 2D, adequando-se as propriedades

descritas acima ao tipo do grafo. Na figura 60 é mostrada a representação em grafo 3D.

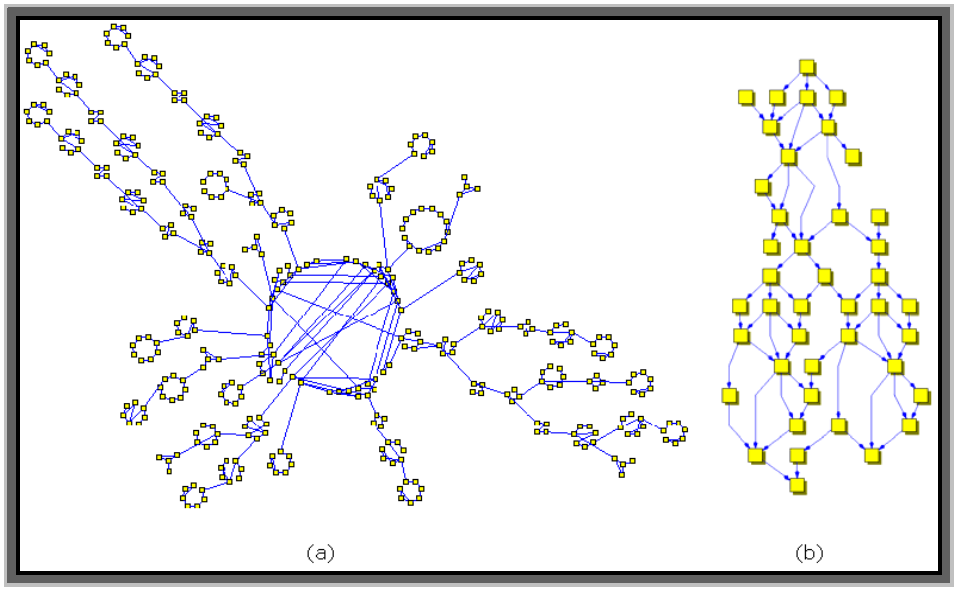


Figura 59 - Representação por grafos na visualização de dados; (a) Grafo otimizado para agrupamento; (b) Grafo acíclico direcionado (ANKERST, 2001)

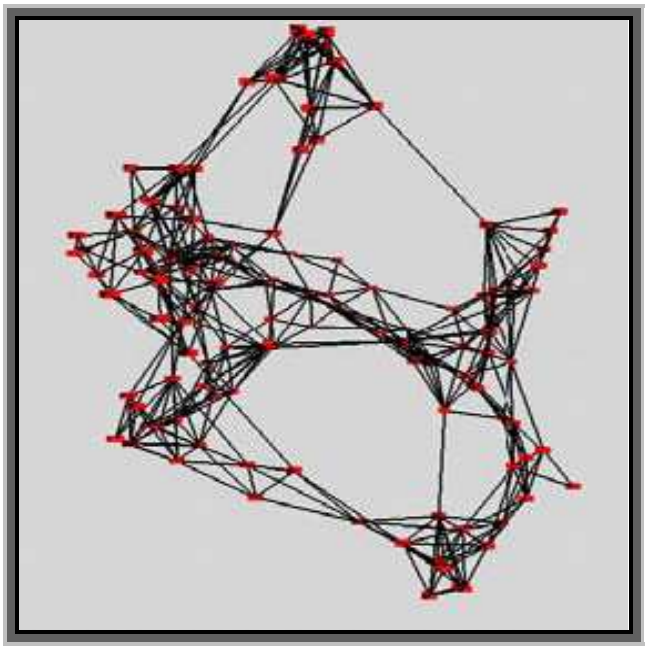


Figura 60 - Representação em 3 dimensões de um grafo otimizado para agrupamentos (ANKERST, 2001)

5.2.6 Técnicas Dinâmicas

Segundo Santos, Grol e Abel (1999), as *Técnicas Dinâmicas* são responsáveis por darem tratamentos dinâmicos às visualizações, fazendo com que estas reajam automaticamente as ações do utilizador ou a mudanças dos dados.

Em geral, estas técnicas podem ser aplicadas nas maiorias dos métodos vistos anteriormente e embora seja possível uma visualização mais detalhada de uma determinada região, as outras regiões são prejudicadas, alterando a estrutura dos dados e dificultando a distinção.

Vistas de Fisheye, por exemplo, usa a idéia de uma lente de aumento ou de *fisheye*, responsável por aumentar os objetos que estão próximos a lente e mostrando os objetos circundantes cada vez com menos detalhes.

O conceito de *Vistas de Fisheye* foi introduzido por Furnas (1986) e foram Sarker e Brow (1992) os primeiros a usarem as *Vistas de Fisheye* em um grafo 2D. Este tipo de vista permite a visualização de um grande volume de informação numa única vista, onde uma sub-região específica poderá ser visualizada com mais detalhes.

Esta técnica permite a interação direta do usuário, ou seja, conforme vai se movimentando a lente, automaticamente o foco vai mudando e toda a visualização centraliza-se neste foco, deixando-o com mais detalhes. A estrutura dos dados vão se modificando conforme a *Vista de Fisheye* vai mudando de posição. A figura 61 mostra a representação por grafo das principais cidades dos EUA e na figura 62 a aplicação da *Vista de Fisheye* com foco em *St. Louis*.

5 VISUALIZAÇÃO DA INFORMAÇÃO

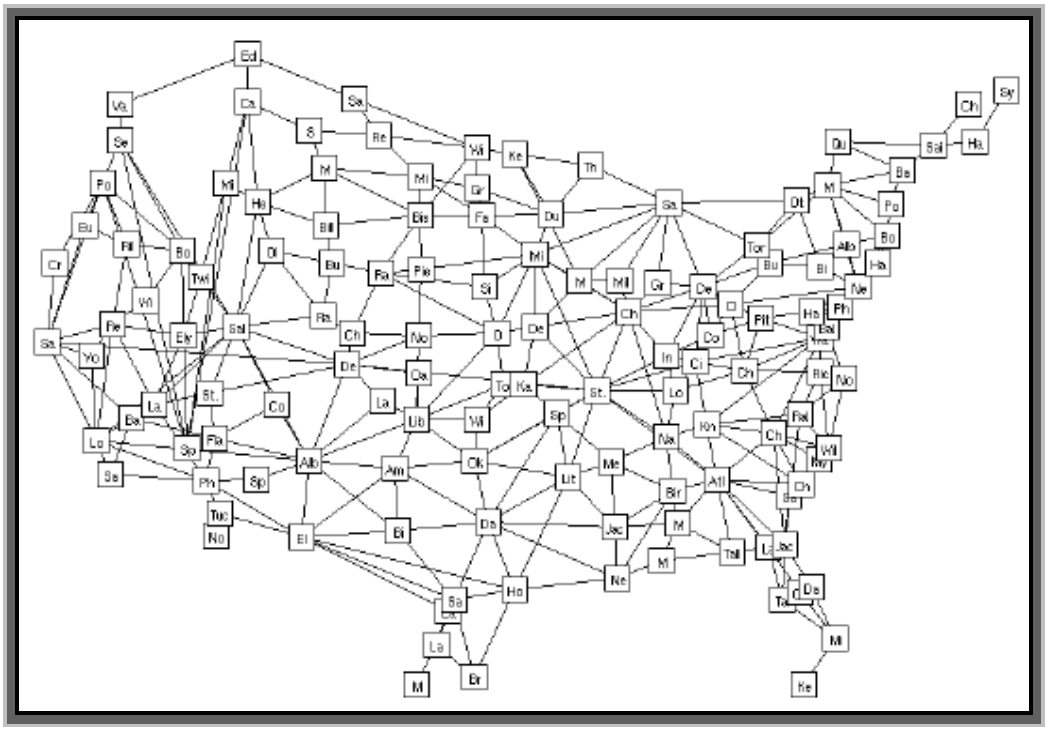


Figura 61 - Grafo representando as principais cidades dos *EUA* (SARKER; BROW, 1992)

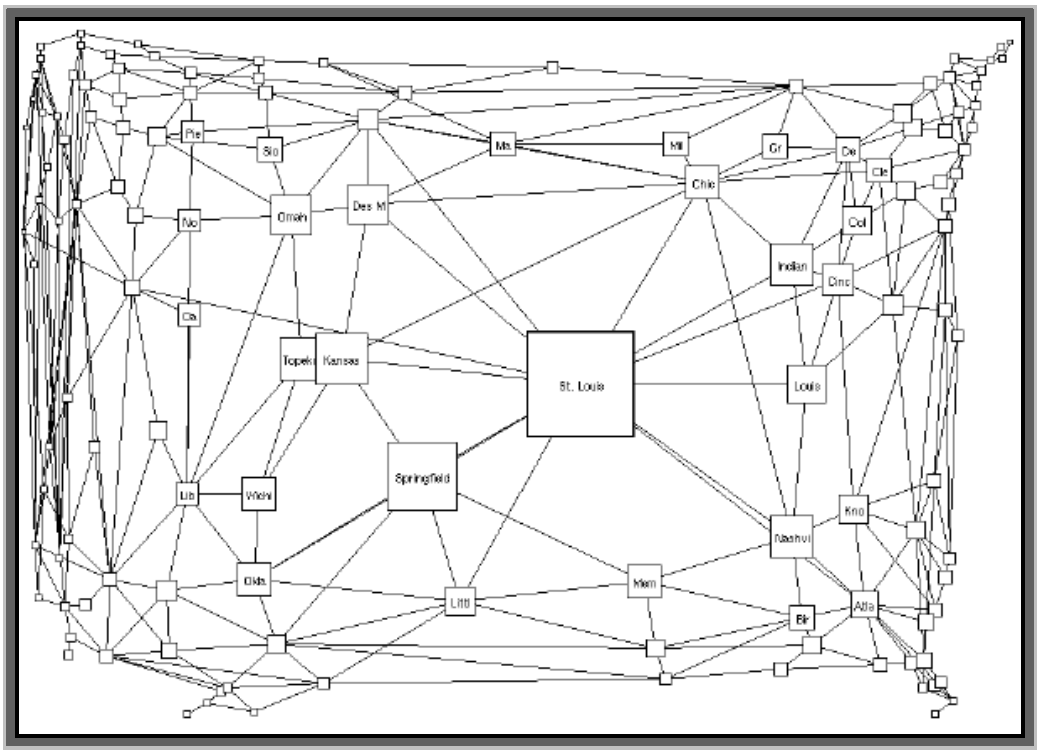


Figura 62 - Uso da técnica *Vistas de Fisheye* nas proximidades de *St. Louis* (SARKER; BROW, 1992)

Bem similar à técnica *Vistas de Fisheye,* a técnica *Rubber Sheet*, usa como metáfora uma folha de borracha. Esta folha pode então ser “esticada” e conseqüentemente os dados são movimentados e o foco, definido pelo utilizador é mostrado com mais detalhes.

Esta técnica foi desenvolvida por Sarkar, entre outros, (1993) como tentativa de aprimorar algumas técnicas de visualização da informação como as *Vistas de Fisheye* e *Paredes de Perspectivas*. A principal vantagem desta técnica quando comparadas às técnicas citadas é a possibilidade do uso de focos múltiplos além de controlar com precisão o espaço alocado a cada um dos focos (ver figura 63).

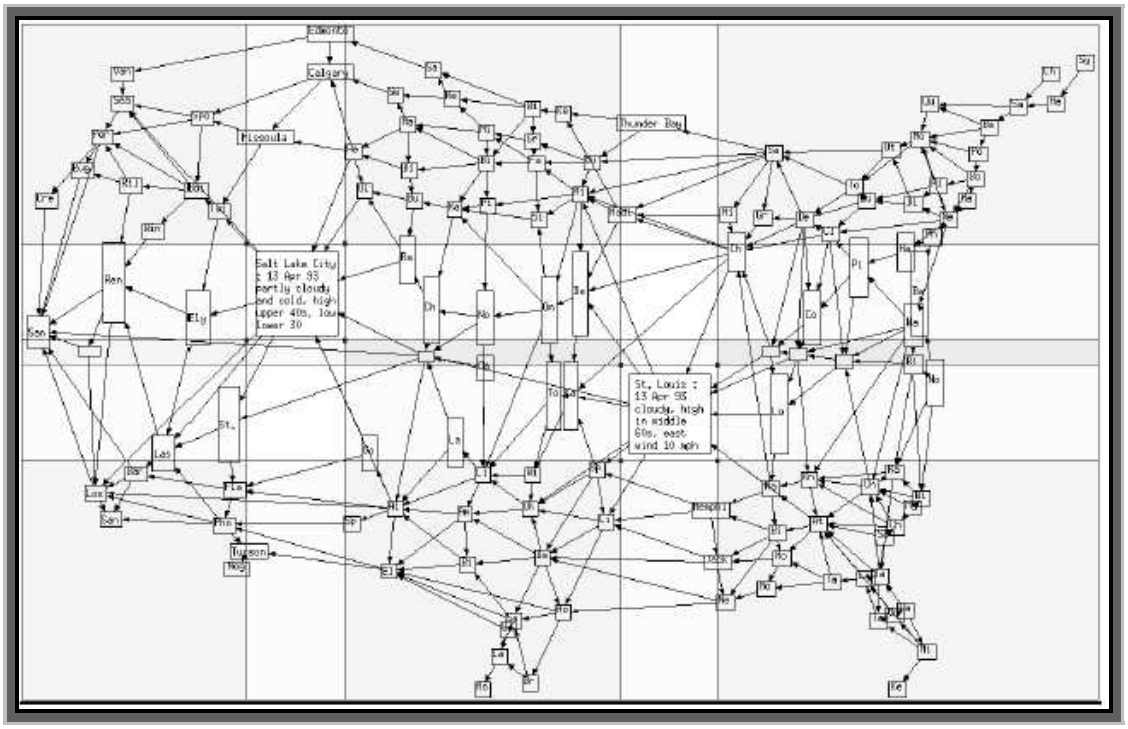


Figura 63 - Uso da técnica *Rubber Sheet* sobre o grafo das cidades dos *EUA* com focos em *St. Louis* e em *Salt Lake City* (SARKAR *et al*, 1993)

Outra técnica de caráter dinâmico são os *Icons Emotivos.* Segundo Walker (1995) os *Icons Emotivos* são *icons* que mudam dinamicamente seu comportamento com a presença de um utilizador no mundo virtual. Estes *icons*

podem, por exemplo, agir agressiva ou passivamente, avançar ou recolher, crescer ou encolher, dependendo do perfil ou do grau da importância que estes dados têm para o utilizador (SANTOS; GROL; ABEL, 1999). *Icons Emotivos* podem agir na presença de outros *icons*. Por exemplo, quando dois *icons* possuem natureza correlacionáveis, estes podem se mover ficando uns próximos aos outros, enquanto que *icons* com natureza sem correlação se afastam entre si.

Para Santos, Grol e Abel (1999), os *Icons Emotivos* podem ser utilizados para aumentar a interatividade de uma apresentação de informação, tornando-a mais dinâmica e mais fácil para o utilizador compreender a informação apresentada. O objetivo é aumentar o impacto e eficácia da interface com o utilizador numa visualização complexa.

5.2.7 Técnicas Híbridas

As *Técnicas Híbridas* são técnicas desenvolvidas a partir do uso de uma ou mais técnicas conhecidas. A técnica *Parallel Glyphs*, proposta por Fanea, Carpendale e Isenberg (2005), por exemplo, usa duas das visualizações clássicas na sua composição, a *Coordenadas Paralelas* e os *Star Glyphs*. Estas técnicas conforme vistas nos itens anteriores mapeiam as informações no espaço bidimensional, a primeira usa eixos paralelos ao longo dos quais são plotados os valores dos pontos correspondentes enquanto que a segunda as informações são formadas por polígonos que radiam eixos a partir de um ponto central.

Estas técnicas foram usadas pelos autores para desenvolver uma técnica interativa, cujos valores são mapeados para uma visualização tridimensional. A idéia é usar os *glyphs* no lugar dos eixos paralelos da técnica *Coordenadas Paralelas*. O uso destes *glyphs* dá uma representação tridimensional à visualização (ver figura 64).

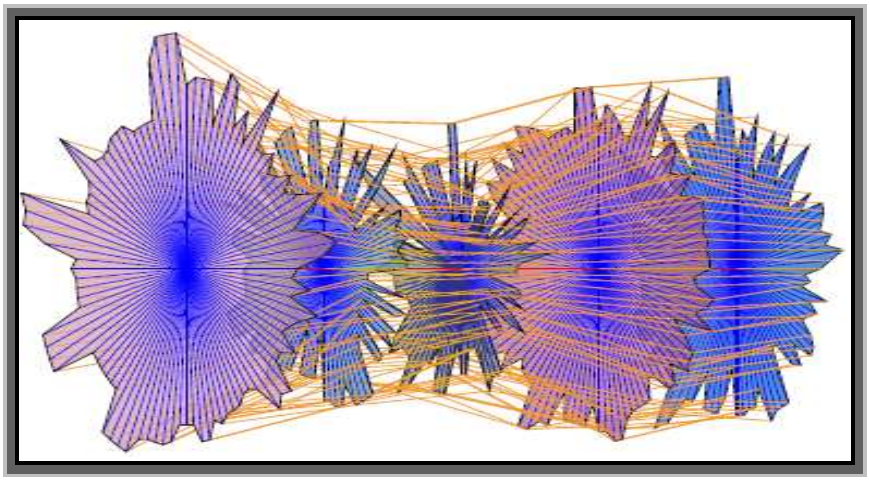


Figura 64 - Representação de dados através da técnica *Parallel Glyphs* (FANEA; CARPENDALE; ISENBERG, 2005)

Para cada coluna da tabela de dados um polígono correspondente é construído cujos vértices são plotados a uma distância proporcional ao valor da tabela. Os polígonos são dispostos a uma mesma distância uns dos outros, ligando vértices correspondentes a uma mesma linha da matriz de dados.

Esta técnica tem como característica principal evitar a sobreposição de linhas. Além de permitir que o usuário navegue sobre os dados de forma interativa. Uma das suas desvantagens, assim como acontece com outras técnicas de *Visualização da Informação*, está na representação visual de dados de alta dimensionalidade, pois quanto o maior o número de dimensões, mais vértices serão construídos nos polígonos e conseqüentemente mais linhas, que vão se tornando cada vez mais próximas umas das outras dificultando a interpretação dos resultados.

5.3 Considerações Finais

A classificação das técnicas de *Visualização da Informação* é uma tarefa difícil. Muitos autores classificam estas técnicas seguindo diversos critérios e algumas das técnicas podem pertencer a uma ou mais classificações. Com base

nas classificações destes autores, e com um levantamento de diversas técnicas existentes na literatura, uma nova classificação foi proposta neste capítulo.

Também foi discutido como técnicas de mineração podem apoiar a exploração visual de grandes conjuntos de dados, e também serem apoiadas por recursos visuais. Percebe-se que os dois problemas críticos em *Mineração de Dados* também são críticos em *Visualização de Informação*. O primeiro é a alta dimensionalidade, tipicamente tratada durante a etapa de *Transformação dos Dados* do *Processo KDD*. O segundo problema é a grande quantidade de registros contidos nas bases de dados atuais, o que demanda o uso de algoritmos cada vez mais eficientes.

Alguns destes problemas podem ser contornados com técnicas interativas, *zoom*, *pan* e rotação. Estas técnicas permitem o usuário navegar em todo o “mundo” virtual, possibilitando que uma determinada região seja vista com mais detalhes.

As técnicas de *Visualização da Informação* permitem uma análise de dados de dimensões grandes, facilitando a extração do conhecimento. Cada técnica de *Visualização* é sugerida a um determinado tipo de dados. A dificuldade de encontrar uma técnica que melhor represente os dados estudados acaba se tornando um problema, porém estas técnicas podem ser usadas em conjunto, permitindo encontrar padrões em técnicas diferentes.

6 MÉTODO DE PESQUISA E EXPERIMENTOS

6.1 Considerações Iniciais

O *Processo KDD*, responsável pela descoberta do conhecimento em banco de dados já inclui em suas etapas a *Visualização*, fundamental em todo o processo na busca do conhecimento.

Diferentes técnicas de duas grandes áreas de pesquisa, *Visualização da Informação* (VI) e tratamento de dados multidimensionais (*Mineração de Dados*) foram estudadas nos capítulos 4 e 5. Os experimentos que aqui serão apresentados permitem dar uma visão de como unir estas áreas para auxiliar no processo de descoberta e extrair de forma mais eficiente informações importantes.

A *Mineração Visual de Dados*, nome dado na tentativa de integrar estas áreas, a *Mineração de Dados* e a *Visualização da Informação*, é similar à etapa *Mineração de Dados*. Assim como esta, a *Mineração Visual de Dados* utiliza diferentes técnicas para encontrar informações úteis “escondidas” na massa de dados. Diferenciam-se uma da outra pelo fato daquela usar a *Visualização* na busca de padrões e não simplesmente em uma análise dos resultados numéricos gerados pela *Mineração de Dados*.

Em geral, as visualizações não requerem conhecimentos específicos de áreas como Matemática e Estatística e utilizam o sentido da visão do ser humano para explorar o ambiente e extrair as informações necessárias (capítulo 2). Por outro lado, a *Visualização da Informação* pode ser usada para auxiliar ou ser auxiliada por algoritmos complexos que envolvem conhecimentos mais

aprofundados destas áreas, melhorando o desempenho e facilitando a extração do conhecimento.

Muitas vezes, o uso das técnicas de VI é o suficiente para extrair as informações que se tem interesse. Porém, algumas vezes, é interessante fazer um estudo através de técnicas de tratamentos de dados multidimensionais, e inserir à *Visualização* para facilitar a interpretação dos resultados.

Neste capítulo será visto duas diferentes abordagens, uma para cada estudo realizado nos experimentos, ITAIPU e SIMEPAR. Os resultados obtidos foram gerados através de *softwares* existentes baseados na visualização de dados multidimensionais. Dentre os *softwares* utilizados estão o MDV, XMDVTool, MatLab e Parvis, vistos no capítulo 2. Além disso, uma classe desenvolvida em java, baseados em *Redes Neurais*, foi implementa para acrescentar ao *software* RadVis a funcionalidade de filtragem dos dados.

O objetivo deste capítulo é mostrar como as técnicas de *Visualização* podem auxiliar e facilitar na extração de conhecimento sem muitos cálculos adicionais.

6.2 Primeiro Experimento: ITAIPU

6.2.1 Introdução à ITAIPU

A ITAIPU Binacional, maior hidrelétrica em produção de energia do mundo, teve o início da sua construção em 1973 num trecho do Rio Paraná conhecido por ITAIPU, que, em tupi, quer dizer “a pedra que canta”, localizado no coração da América do Sul na divisa entre o Paraguai e o Brasil. Entre 1975 e 1978, um desvio do rio Paraná de cerca de 2 Km foi escavado para alterar o curso do rio. O novo canal permitiu que o trecho do leito principal do rio fôsse secado, para ali ser construída a barragem principal, em concreto. No dia 14 de novembro de 1978, foi

realizada a concretagem de 7.207 metros cúbicos, equivalente a uma construção de 24 edifícios de dez andares no mesmo dia, um recorde sul-americano na Engenharia Civil. Em 1981, com as obras quase concluídas, dá-se o início a montagem das unidades geradoras com instalação de turbinas. Em outubro de 1982, chegam ao fim as obras da barragem. Com o fechamento das comportas do canal de desvio, para a formação do reservatório da usina, dá-se o início da operação *Mymba Kuera* (que em tupi-guarani quer dizer “pega-bicho”). A operação salva a vida de 36.450 animais que viviam na área a ser inundada pelo lago, cerca de 1350 Km^2. Em 5 de novembro de 1982, com o reservatório já formado, os presidentes do Brasil, João Figueiredo, e do Paraguai, Alfredo Stroessner, acionam o mecanismo que levanta automaticamente as 14 comportas do vertedouro, liberando a água represada do Rio Paraná e, assim, inauguram oficialmente a maior hidrelétrica do mundo (ITAIPU, 2008).

Atualmente a ITAIPU possui 20 unidades geradoras de 700 MW (*megawatts*) cada, gerando uma potência total instalada de 14.000 MW. No ano 2000, a ITAIPU Binacional bateu seu recorde em geração de energia, cerca de 93,4 bilhões de quilowatts-hora (KWh) foram gerados naquele ano.

A ITAIPU Binacional é responsável pelo abastecimento de 95% da energia elétrica consumida no Paraguai e 24% de toda a demanda do mercado brasileiro.

A barragem da ITAIPU tem 7.919 metros de extensão e altura máxima de 196 metros, o equivalente a um prédio de 65 andares. Consumiu 12,3 milhões de metros cúbicos de concreto, enquanto o ferro e o aço utilizados permitiriam a construção de 380 Torres Eiffel, dimensões que transformaram a usina em referência nos estudos de concreto e na segurança de barragens. A figura 65 mostra a estrutura geral da barragem de ITAIPU, e a tabela 7 mostra a as principais características dos trechos da barragem.

Figura 65 - Estrutura geral do complexo ITAIPU (ITAIPU, 2008)

Tabela 7 - Características dos trechos da Barragem do ITAIPU

Trecho		Estrutura	Comprimento (m)	Altura Máxima (m)
1 (L)	Barragem Auxiliar	Terra	2294	30
2 (K)	Barragem Auxiliar	Enrocamento	1984	70
3 (E e I) / 7 (D)	Barragens Laterais	Contraforte	1438	81
4 (H)	Estrutura de Desvio	Concreto Maciço	170	162
5 (F)	Barragem Principal	Gravidade Aliviada	612	196
9 (Q)	Barragem Auxiliar	Terra	872	25
8 (A)	Vertedouro	350 m de Largura		
6 (U)	Casa de Força	20 Unidades Geradoras		

6.2.2 Monitoramento e Instrumentação Estrutural

A Barragem de ITAIPU é composta por dois trechos de barragens de terra, um trecho de barragem de enrocamento e um trecho de concreto. Em toda sua extensão, para acompanhar o desempenho das estruturas de concretos e fundação, são encontrados 2218 instrumentos (1362 no concreto e 856 nas fundações e aterros) sendo 270 automatizados, e 5239 drenos (949 no concreto e 4290 nas fundações), cujas leituras ocorrem em diferentes freqüências, podendo ser, por exemplo, diária, semanal, quinzenal, mensal, dependendo do tipo de instrumento. Conta-se também com a monitoração dos dados hidro-meteorológicos, realizada através de algumas estações que são da própria

ITAIPU e de outras entidades, como a Companhia Paranaense de Energia (COPEL), Agência Nacional de Águas (ANA) e Operador Nacional do Sistema (ONS). A ITAIPU também utiliza dados de sistemas meteorológicos, como imagens de satélite, imagens de radar e localização de descargas elétricas, por meio de convênios com o Sistema Meteorológico do Paraná (SIMEPAR) e da paraguaia Dirección Nacional de Aeronáutica Civil (DINAC) (ITAIPU, 2008).

A tabela 8 mostra resumidamente as funcionalidades dos instrumentos encontrados ao logo da barragem diferenciando em dois tipos, concreto e fundação.

Tabela 8 - Funcionalidades dos instrumentos encontrados na barragem de ITAIPU no concreto e na fundação (ITAIPU, 2008)

Tipo	Instrumento	Funcionalidade
Concreto	Caixa seletora	Reúne os cabos elétricos de vários instrumentos em uma caixa central que, ao ser conectada ao aparelho de leitura, fornece dados destes instrumentos.
	Pêndulo direto	Mede os deslocamentos horizontais de pontos dos blocos instrumentados da barragem em determinadas cotas, em relação à fundação da estrutura.
	Pêndulo invertido	Mede os deslocamentos da fundação da barragem em relação a um ponto da fundação suficientemente profundo para ser considerado fixo.
	Medidor elétrico de junta	Mede os deslocamentos de abertura e fechamento de determinadas juntas de contração de estruturas de concreto.
	Base de alongâmetro	Mede abertura, fechamento, recalque e deslizamento entre blocos ou juntas de monolitos.
	Deformímetro de armadura	Mede as tensões em barras de armadura, no interior de estruturas de concreto.
	Deformímetro de concreto	Mede a deformação do concreto e, por esta deformação, obtém-se a tensão que está atuando na estrutura.
	Termômetro de resistência	Mede a temperatura no interior da estrutura de concreto.
Fundação	Medidor de vazão	Mede as vazões de percolação através das estruturas e fundações das obras de terra e concreto.
	Extensômetro múltiplo de haste	Mede as deformações da fundação com relação ao ponto de ancoragem de sua haste.
	Piezômetro Standpipe	Permite conhecer a subpressão atuante no local da sua instalação.
	Piezômetro elétrico	Permite conhecer a subpressão atuante no local da sua instalação.

Tipo	Instrumento	Funcionalidade
	Medidor de assentamento IPT (medidor de recalque)	Mede deformações verticais ocorridas nas barragens de terra.
	Célula de pressão total	Mede as pressões totais atuantes na zona de contato solo-concreto.
	Medidor triorgogonal	Mede os deslocamentos entre juntas de concreto e zonas fraturadas nos maciços rochosos.
	Medidor de nível d'água	Mede o nível da água presente no subsolo (lençol freático).

Embora os nove trechos da barragem de ITAIPU sejam instrumentados e monitorados, o trecho Barragem Principal (trecho F) merece destaque e um estudo mais aprofundado. O trecho F é responsável pela movimentação das turbinas para geração de energia elétrica, além de ser o trecho de maior altura em coluna de água, tornando-o um dos mais críticos.

O trecho F é constituído de vários blocos, sendo que cada um deles possui instrumentos que fornecem dados a respeito de seu comportamento físico, tanto na estrutura de concreto como na sua fundação. Nas Tabelas 9 e 10, podem-se observar os tipos e quantidades de instrumentos instalados no concreto e na fundação dos blocos do trecho F.

Tabela 9 - Quantidades e tipos de instrumentos no concreto encontrados nos blocos do trecho F da barragem de ITAIPU (ITAIPU, 2008)

Instrumentos no Concreto							
Instrumento	Sigla	Blocos do Trecho F					Total por Instrumento
		5/6	13/14	15/16	19/20	35/36	
Rosetas Deformímetros	RD	4	-	-	11	-	15
Tensômetro	TN	1	-	-	4	-	5
Rosetas de Tensômetros	RT	2	-	-	6	-	8
Medidor de Junta Interna	JM	-	-	-	7	-	7
Pêndulo Direto	PD	5	6	-	6	4	21
Pêndulo Invertido	PI	3	1	1	1	-	6
Termômetro na Massa	TM	3	-	-	17	3	23
Termômetro na Superfície	TS	2	-	-	6	2	10
Total por Bloco		20	7	1	58	9	95

Tabela 10- Quantidades e tipos de instrumentos na fundação encontrados nos blocos do trecho F da barragem de ITAIPU (ITAIPU, 2008)

Instrumentos na Fundação																		
Instrumento	**Sigla**	**Blocos do Trecho F**																**Total**
		1/2	3/4	5/6	7/8	9/10	11/12	13/14	15/16	17/18	19/20	21/22	23/24	27/28	29/30	31/32	35/36	
Piezômetro Standpipe	PS	-	4	6	5	-	6	7	3	6	8	-	4	10	-	6	9	74
Piezômetro Geomor	PG	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Extensômetro de Haste	EM	4	-	1	-	-	-	3	5	4	3	1	-	4	-	-	4	29
Medidor de Aterro	MA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Medidor Triortogonal	MT	-	1	-	-	1	4	1	1	-	-	1	1	-	-	1	-	11
Célula de Pressão Total	CL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Medidor de Vazão	MV	-	1	-	-	-	2	-	2	-	-	2	1	-	1	-	-	9
Drenos	DR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
Medidor de Nível d'Água	PZ	-	-	-	-	-	-	1	1	1	1	-	-	-	-	-	-	4
Total por Bloco		4	6	7	5	1	12	12	12	11	12	4	6	14	1	7	13	127

6.2.3 Organização dos Dados

Os dados das instrumentações da barragem da ITAIPU, originalmente, encontravam-se organizados em arquivos em formato texto. De acordo com Andraos (2006) e Sanchez (2006), estes dados foram reorganizados em um banco de dados em planilhas do Excel, onde se procurou relacionar as informações de projeto com as constantes nos arquivos texto das leituras dos instrumentos.

Buzzi (2007), baseado neste banco de dados, desenvolveu um programa em MATLAB capaz de unir instrumentos diferentes cujas leituras tinham ocorrências numa mesma data. Para alguns instrumentos, o número de leituras ocorridas em mesma data foi baixo. Com isso, certas análises poderiam gerar resultados duvidosos. Para estes casos, Buzzi (2007) propôs adotar tolerâncias de defasagem entre leituras de diferentes instrumentos de um até três dias.

Com base na organização inicial dos dados da ITAIPU, realizados pelos trabalhos de Andraos (2006) e Sanchez (2006), e através do programa elaborado por Buzzi (2007), neste trabalho, os dados foram organizados em tabelas onde as colunas representam as variáveis envolvidas, e as linhas, as leituras nas

diferentes datas. A figura 66 ilustra a organização dos dados em tabela de parte das leituras dos instrumentos do tipo extensômetro.

ANO	MÊS	EMF 21_h1	EMF 21_h2	EMF 22_h1	EMF 22_h2	EMF 22_h3	EMF 23_h1	EMF 23_h2	EMF 23_h3	EMF 24_h1	EMF 24_h2	EMF 24_h3
1996	1	4,81	5,56	-11,05	-8,79	-6,79	-8,06	-6,77	-2,52	-9,17	-6,92	-1,52
1996	2	4,8	5,58	-11,04	-8,79	-6,78	-8,06	-6,76	-2,5	-9,15	-6,96	-1,51
1996	3	4,86	5,47	-11,08	-8,88	-6,81	-8,07	-6,79	-2,52	-9,17	-6,93	-1,53
1996	4	4,73	5,37	-11,12	-8,82	-6,84	-8,09	-6,8	-2,55	-9,18	-6,95	-1,54
1996	5	4,73	5,42	-11,12	-8,83	-6,81	-8,1	-6,81	-2,53	-9,12	-6,98	-1,53
1996	6	4,74	5,42	-11,1	-8,79	-6,77	-8,09	-6,81	-2,53	-9,14	-6,98	-1,52
1996	7	4,72	5,47	-11,07	-8,84	-6,77	-8,1	-6,77	-2,55	-9,18	-7,01	-1,53
1996	8	4,78	5,59	-11,03	-8,82	-6,73	-8,1	-6,78	-2,54	-9,18	-7,02	-1,53
1996	9	4,82	5,65	-11,05	-8,81	-6,73	-8,09	-6,78	-2,54	-9,19	-7,04	-1,52
1996	10	4,79	5,62	-11,08	-8,84	-6,75	-8,12	-6,8	-2,55	-9,23	-7,05	-1,56
1996	11	4,83	5,67	-11,04	-8,79	-6,74	-8,08	-6,76	-2,53	-9,18	-7,01	-1,51
1996	12	4,8	5,61	-11,08	-8,83	-6,78	-8,09	-6,79	-2,57	-9,2	-7,04	-1,54
1997	1	4,79	5,6	-11,1	-8,82	-6,79	-8,09	-6,79	-2,52	-9,18	-7,04	-1,52
1997	2	4,81	5,54	-11,11	-8,83	-6,81	-8,11	-6,81	-2,53	-9,2	-7,03	-1,53
1997	3	4,76	5,5	-11,15	-8,85	-6,84	-8,12	-6,82	-2,54	-9,21	-7,04	-1,53
1997	4	4,79	5,52	-11,15	-8,81	-6,79	-8,07	-6,77	-2,52	-9,14	-6,98	-1,5
1997	5	4,75	5,49	-11,19	-8,86	-6,86	-8,14	-6,85	-2,57	-9,22	-7,07	-1,59
1997	6	4,75	5,5	-11,2	-8,88	-6,831	-8,12	-6,84	-2,56	-9,2	-7,05	-1,56
1997	7	4,81	5,58	-11,19	-8,87	-6,84	-8,14	-6,81	-2,56	-9,21	-7,07	-1,59
1997	8	4,77	5,58	-11,13	-8,84	-6,8	-8,07	-6,81	-2,54	-9,19	-7,04	-1,53
1997	9	4,78	5,6	-11,12	-8,82	-6,79	-8,1	-6,81	-2,54	-9,2	-7,05	-1,55
1997	10	4,73	5,55	-11,13	-8,82	-6,8	-8,14	-6,85	-2,56	-9,18	-7,05	-1,53
1997	11	4,77	5,61	-11,17	-8,86	-6,83	-8,12	-6,82	-2,53	-9,22	-7,08	-1,56
1997	12	4,8	5,6	-11,15	-8,85	-6,81	-8,11	-6,82	-2,53	-9,2	-7,07	-1,53
1998	1	4,9	5,63	-11,08	-8,76	-6,74	-8,02	-6,72	-2,43	-9,11	-6,97	-1,43
1998	2	4,82	5,45	-11,21	-8,88	-6,88	-8,14	-6,84	-2,56	-9,23	-7,09	-1,56

Figura 66 - Representação de parte dos instrumentos do tipo extensômetros

6.2.4 Técnicas Visuais Aplicadas aos Dados de ITAIPU

Conforme observado no capítulo 5, diferentes técnicas de visualização da informação podem ser usadas na análise de dados multidimensionais. A escolha de uma delas não é uma tarefa fácil. Cada técnica possui vantagens e desvantagens em relação às demais.

Nesta seção, diferentes técnicas de *Visualização da Informação* foram aplicadas aos dados de monitoramento da Barragem da ITAIPU com o objetivo principal de analisar as relações existentes entre as variáveis. Em algumas visualizações é necessário usar recursos estatísticos a fim de facilitar a interpretação dos resultados.

Uma enorme quantidade de simulações poderia ser feita com os dados, e aqui optou-se por selecionar somente um caso para mostrar como se pode

realizar a análise visual dos dados através das técnicas de *Mineração Visual de Dados*.

Neste primeiro experimento, selecionou-se do banco de dados aqueles cujas variáveis pertenciam ao grupo dos extensômetros. Esta seleção foi realizada, pois se espera que haja relacionamentos consideráveis entre estes tipos de variáveis. Além destas aplicações aqui mostradas, ainda poderiam ser obtidas novas visualizações conforme necessidade do usuário.

Os extensômetros são instrumentos que medem as deformações da fundação e são formados por várias hastes. No bloco 19/20 do trecho F da Barragem da ITAIPU, são encontrados quatro instrumentos deste tipo, assim denominados: EMF21, EMF22, EMF23 e EMF24. O instrumento EMF21 é constituído de duas hastes enquanto que os demais possuem três hastes. Os dados aqui abordados foram selecionados no período de janeiro de 1996 a janeiro de 2006. Este histórico de dez anos foi suficiente para analisar o comportamento entre os instrumentos que compõem um conjunto de 110 leituras de 13 variáveis (incluindo Ano e Mês).

Conforme mostrado na figura 66, estes dados foram organizados em tabelas facilitando as entradas dos *softwares* utilizados para gerar as visualizações.

A figura 67 ilustra o uso das *Coordenadas Paralelas* para visualizar os dados dos instrumentos do tipo extensômetros. Nesta imagem, pôde se observar claramente a relação existente entre vários pares de variáveis. A correlação entre variáveis, conforme visto anteriormente pode ser analisada quando existem poucos cruzamentos entre as linhas que saem de um eixo a outro vizinho, aproximando-se de retas paralelas. Observe-se que nesta imagem somente os pares ANO x MÊS, MÊS x EMF21_h1 não possuem bons relacionamentos, como era de se esperar.

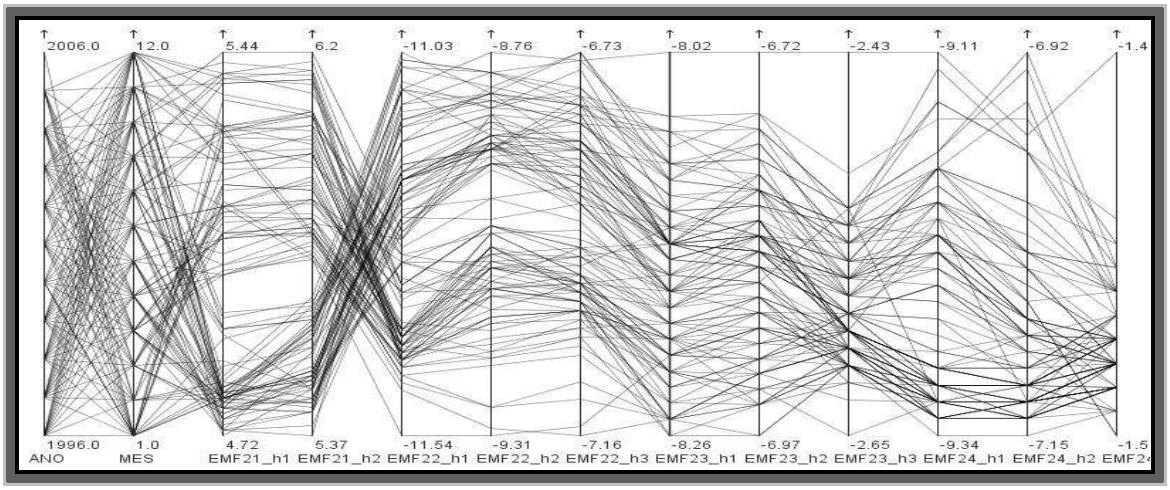


Figura 67 - Análise visual das relações existentes entre pares de variáveis do instrumento do tipo extensômetro, utilizando Coordenadas Paralelas (imagem gerada pelo *software* ParVis)

Olhando para os eixos EMF21_h2 e EMF22_h1, uma análise curiosa pode ser realizada. Observe com o auxílio da figura 68 que para valores baixos da variável EMF21_h2 os valores da variável EMF22_h1 são altos, e vice-versa, isso induz a concluir que estas variáveis possuem comportamento de relacionamento inverso (coeficiente de correlação negativo). Os demais pares, com exceção daqueles mencionados anteriormente, possuem um relacionamento positivo.

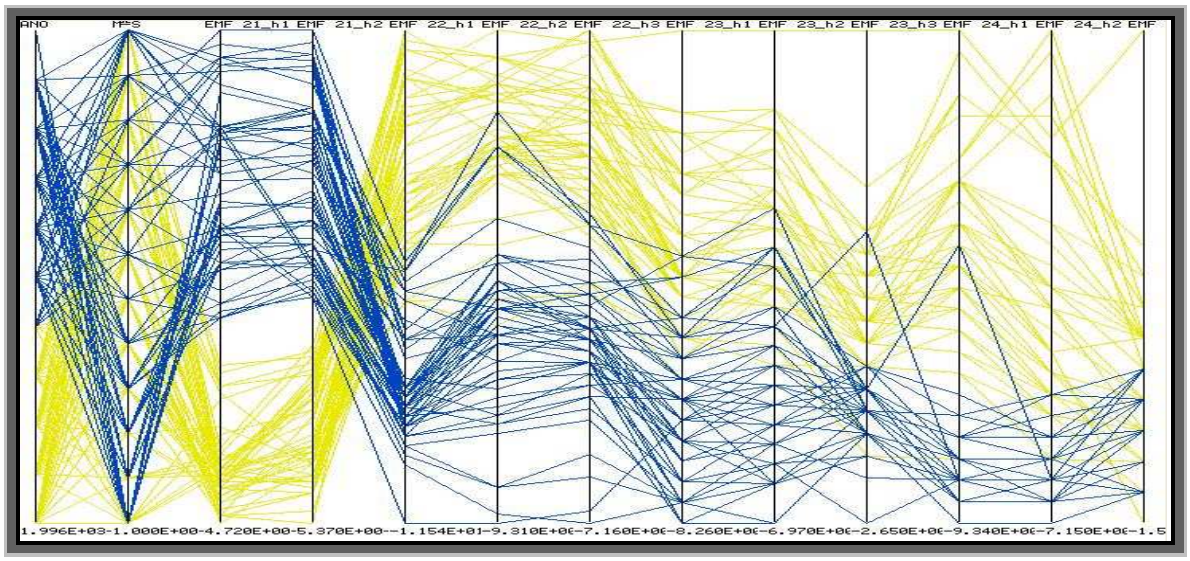


Figura 68 - Ilustração por *Coordenadas Paralelas* do comportamento das variáveis EMF21_h2 e EMF22_h1 (imagem gerada pelo software MDV)

A principal desvantagem deste método está no fato de não poder analisar a relação entre variáveis cujos eixos não são vizinhos. Desta forma, não se pode, de forma natural analisar, por exemplo, se as variáveis EMF22_h1 e EMF22_h3 são bem relacionáveis.

Carvalho (2001), em seu trabalho, propõe um algoritmo para ordenação dos eixos conforme os valores das correlações entre variáveis. Esta *Mineração Visual de Dados,* integrando *Análise de Correlações* com *Visualização da Informação,* permitiu unir os eixos cujas correlações são mais altas. A figura 69 exemplifica o uso deste algoritmo, onde é possível fazer novas observações. Esta nova imagem permitiu observar que as variáveis EMF22_h1 e EMF22_h3, já mencionadas anteriormente, possuem um valor alto de relacionamento. Já as variáveis EMF24_h3 e EMF21_h1, devido a grande quantidade de cruzamentos das linhas, não possuem uma boa relação. A relação entre outras variáveis também podem ser observadas.

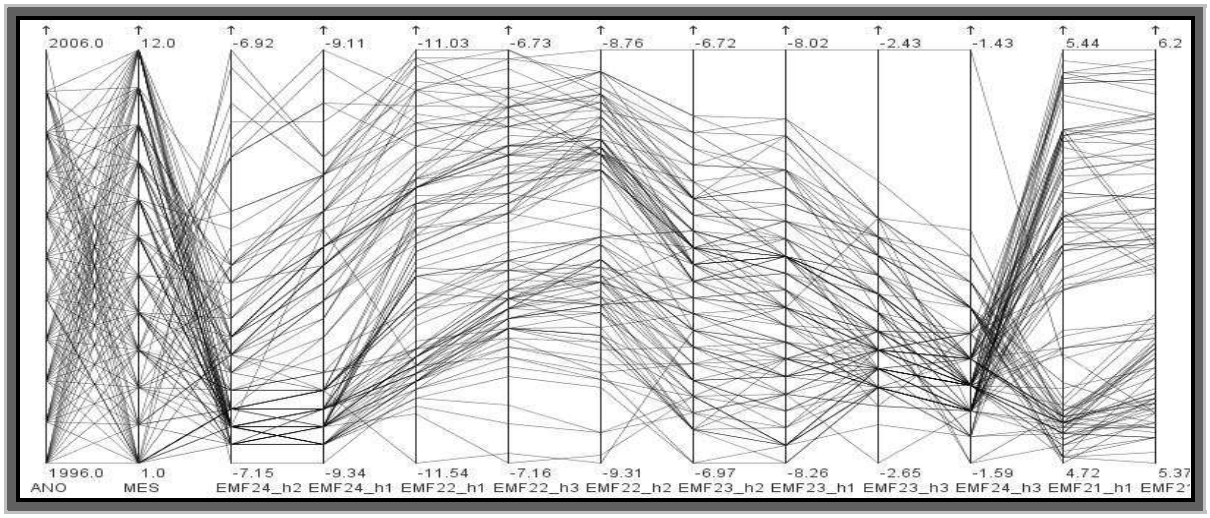


Figura 69 - Técnica *Coordenadas Paralelas* aplicada a visualização dos dados dos instrumentos do tipo extensômetro ordenados pelos valores de suas correlações (imagem gerada pelo *software* ParVis)

As ordenações dos eixos também podem ser feitas de forma interativa, onde intuitivamente o usuário aproxima aqueles eixos que deseja analisar. O *software* ParVis permite este tipo de interatividade.

Outra técnica bastante utilizada para analisar relações entre variáveis são a *Scatterplots* (gráficos de dispersão) e a *Orientada a Pixels.* Estas técnicas mapeiam as variáveis numa única tela permitindo compará-las par a par.

A figura 70 ilustra o uso da técnica *Scatterplots* na visualização das variáveis do instrumento extensômetro, lembrando que variáveis cujos dados estão dispersos aproximando-se do comportamento de uma reta indicam variáveis bem relacionadas.

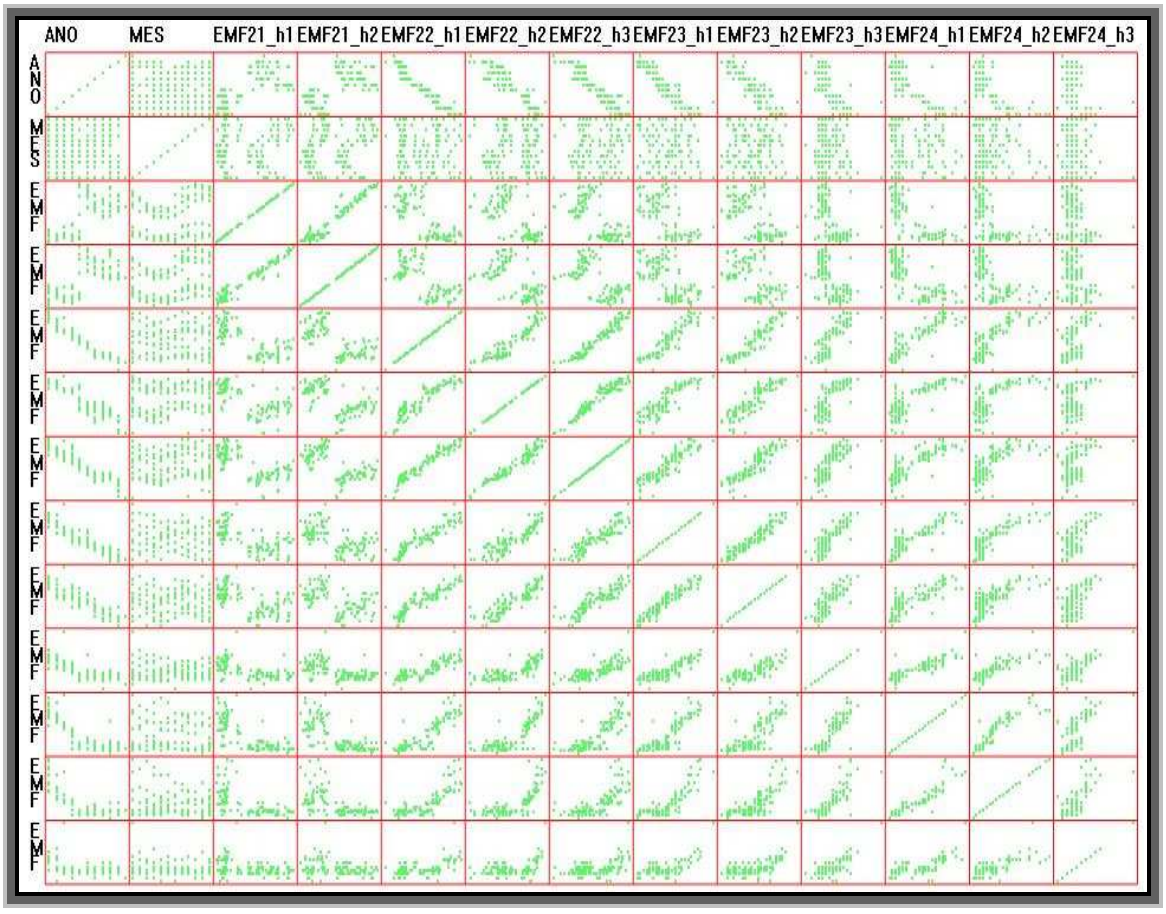


Figura 70 – Relação entre as variáveis do instrumento do tipo extensômetro mostradas pela técnica *Scatterplots* (imagem gerada pelo *software* XmdvTool)

Uma análise na imagem gerada por esta técnica permitiu observar que olhando para a linha 'ANO' e percorrendo pelas colunas, com exceção das variáveis EMF21_h1, EMF21_h2 e EMF24_h3, as demais possuem um

relacionamento inverso ao ano, ou seja, conforme o passar dos anos, os valores destas variáveis foram diminuindo.

A figura 71 utiliza a técnica *Orientada a Pixel* para visualizar os dados do extensômetro, onde imagens visualmente parecidas indicam um bom relacionamento entre as variáveis. Desta forma, uma análise da imagem, permite observar que as variáveis que formam os seguintes grupos: {EMF22_h1, EMF22_h2, EMF22_h3, EMF23_h1 e EMF23_h2}, { EMF21_h1, EMF21_h2}, { EMF24_h1, EMF24_h2} e { EMF23_h3, EMF24_h3} possuem um bom relacionamento entre si.

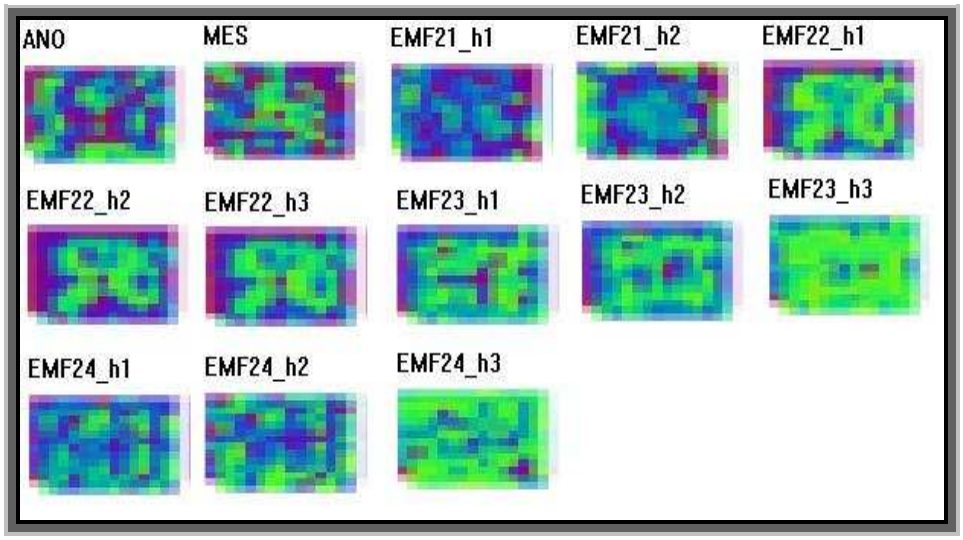


Figura 71 - Uso da técnica *Orientada a Pixel* para representar os dados de extensômetro (imagem gerada pelo *software* XmdvTool)

As técnicas *Star Glyphs* e *Faces Chernoff* também podem ser utilizadas com a finalidade de analisar as correlações entre as variáveis. A figura 72 ilustra o uso destas técnicas, onde *stars* ou *faces* semelhantes indicam um bom relacionamento entre as variáveis. Assim, os seguintes grupos, {EMF21_h1, EMF21_h2}, {EMF22_h1, EMF22_h2, EMF22_h3, EMF21_h1, EMF23_h1, EMF23_h2, EMF24_h1, EMF24_h2} e {EMF23_h3, EMF24_h3} possuem um bom relacionamento de suas variáveis.

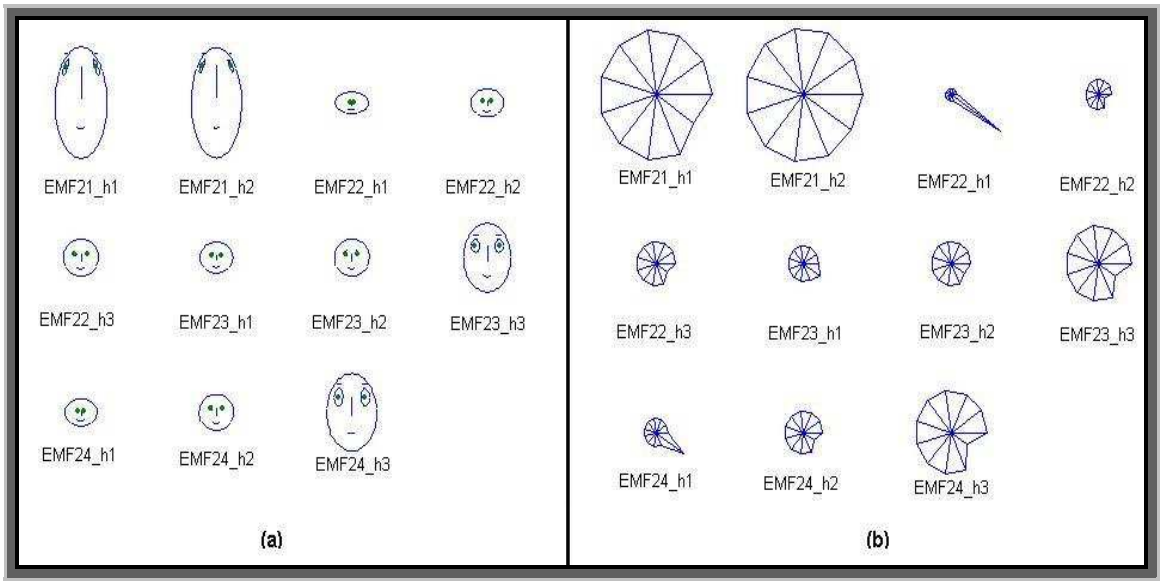


Figura 72 - Relacionamento das variáveis através das técnicas (a) *Star Glyphs* e (b) *Faces de Chernoff* (imagem gerada pelo *software* MATLAB)

Além de analisar as relações existentes entre variáveis, as técnicas de *Visualização da Informação* podem ser usadas na análise e formação de agrupamentos (*clusters*). Estas técnicas permitem, sem a necessidade de ter conhecimentos em áreas como estatísticas ou matemática, analisar visualmente a imagem formada pelas variáveis e nela, intuitivamente, formar os grupos.

Para os dados em análise, a formação dos *clusters* pela variável *'Ano'* induz a interpretar a variação dos valores das demais variáveis. As imagens da figura 73 usam a técnica *Coordenadas Paralelas* para destacar o comportamento destas variáveis ao longo dos anos. Observe que assim como na análise feita pela técnica *Scatterplots*, é possível observar o decrescimento daquelas variáveis ao passar dos anos.

6 MÉTODOS DE PESQUISAS E EXPERIMENTOS

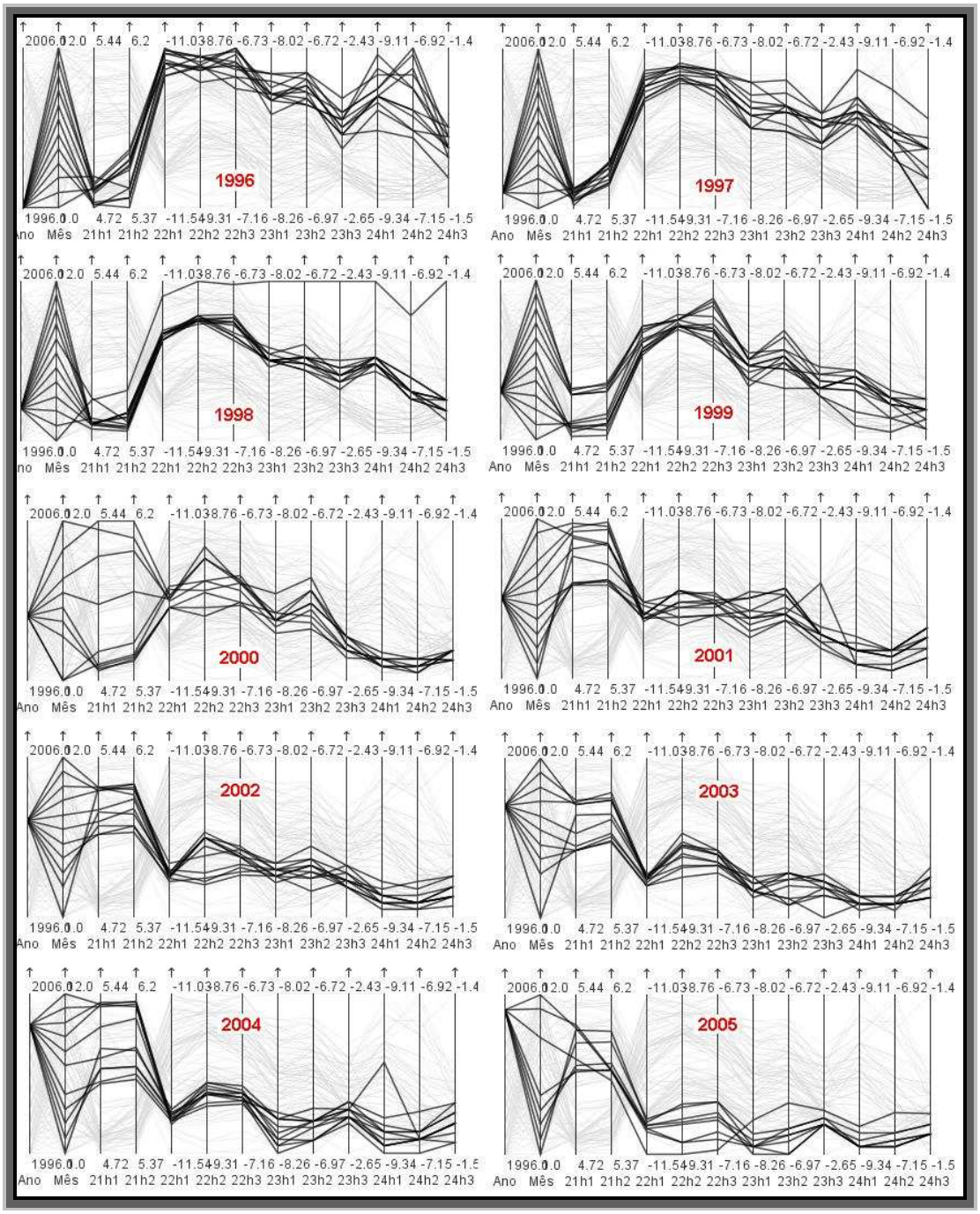


Figura 73 - Uso das técnicas *Coordenadas Paralelas* no agrupamento por ano das variáveis dos extensômetros (imagem gerada pelo *software* ParVis)

Vale observar também que outros agrupamentos poderiam ser realizados. Ao invés de agrupar por ano, poder-se-ia agrupar pelos meses, por exemplo, e então se teria o comportamento mensal das variáveis.

A técnica *RadVis* também permite a análise de agrupamentos. A figura 74 ilustra o uso desta técnica no agrupamento destes mesmos dados conforme a variável ‘ANO’, mostrando a distribuição das variáveis.

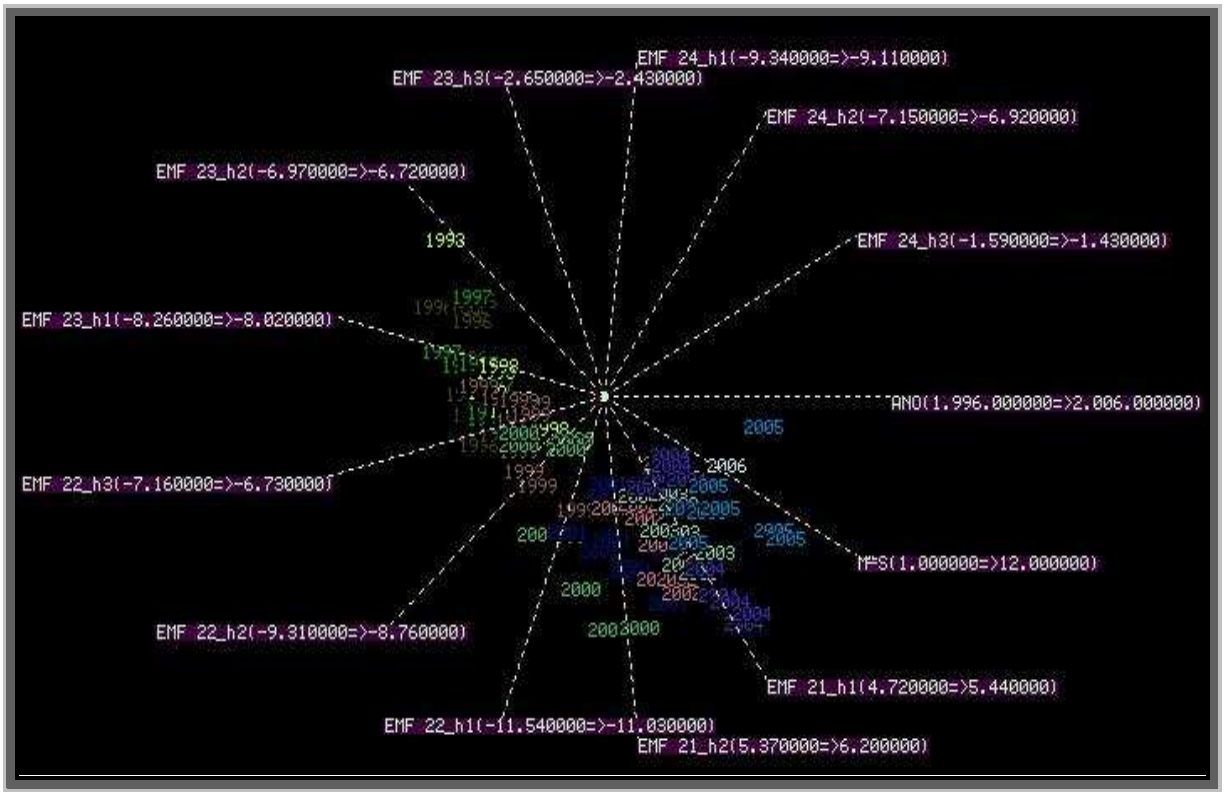


Figura 74 - Técnica *RadVis* aplicada aos dados de extensômetros no agrupamento por ano (imagem gerada pelo *software* MDV)

Estas técnicas, embora possuam limitações, podem ser úteis na extração do conhecimento em banco de dados. Em geral, uma única técnica de *Visualização da Informação* não permite a extração de todo o conhecimento, porém se estas técnicas forem utilizadas em conjunto e ainda unidas com métodos numéricos para tratamentos de dados (técnicas de *Mineração de Dados*), a extração do conhecimento se torna mais fácil e rápida.

6.3 Segundo Experimento: SIMEPAR

6.3.1 Introdução ao SIMEPAR

O Instituto Tecnológico SIMEPAR, empresa de direito privado e interesse público, foi instituído em março de 1993 na forma de um convênio entre o Instituto Agronômico do Paraná (IAPAR) e a Companhia Paranaense de Energia (COPEL). Junto ao Laboratório de Estudos em Monitoramento e Modelagem Ambiental (LEMMA) da UFPR, o SIMEPAR organiza seu processo buscando implementar sistemas e desenvolver soluções tecnológicas nas áreas de Meteorologia, Hidrologia e Meio Ambiente, de forma a obter maior provimento dos dados e das previsões de tais áreas.

O uso das tecnologias e das informações ambientais disponíveis no SIMEPAR auxiliam na tomada de decisões de profissionais de diferentes ramos de atuação. Na agricultura, por exemplo, as informações geradas pelo SIMEPAR contribuem para a determinação da época ideal de plantio e colheita, bem como para indicação do melhor momento para aplicação de adubos e defensivos. Além disso, o SIMEPAR faz previsões de geada, granizo, chuva, etc.

No transporte, uma previsão confiável do tempo garante menos atrasos e maior segurança nos vôos e nas operações portuárias. O SIMEPAR fornece também informações sobre condições do tempo nas estradas.

Com as informações disponíveis é possível, também, programar atividades de lazer e turismo com antecedência, além de poderem ser usadas no ramo de energia, contribuindo para as empresas de geração, transmissão e distribuição, na redução de riscos na operação hidro-energética, diminuindo custos de manutenção de linhas de transmissão e distribuição, agregando segurança em novos projetos, fornecendo informações para avaliação da viabilidade de

exploração de fontes alternativas de energia elétrica, reduzindo a freqüência e a duração das interrupções de fornecimento de energia elétrica e permitindo o controle mais adequado dos níveis dos reservatórios das usinas. O Sistema também proporciona economia de energia e maior segurança para pessoas e propriedades.

Desde a sua instituição, o SIMEPAR já dispunha de toda a infraestrutura de equipamentos para monitoramento e previsão hidrometeorológica. No Estado do Paraná, estão espalhados entre as cidades, 39 estações meteorológicas, 36 estações hidrológicas, seis estações remotas de recepção de descargas atmosféricas e um radar meteorológico.

As estações meteorológicas são compostas de antenas e de sensores responsáveis pela coleta de dados de temperatura, direção e velocidade dos ventos, umidade relativa, chuva, pressão atmosférica e radiação solar, que são transmitidos via satélite para a sede do SIMEPAR, em Curitiba.

A figura 75 mostra a distribuição da temperatura mínima em todo estado do Paraná, gerada com base nos dados das estações meteorológicas dos SIMEPAR do dia 08 de janeiro de 2008 através do *software* GraDS.

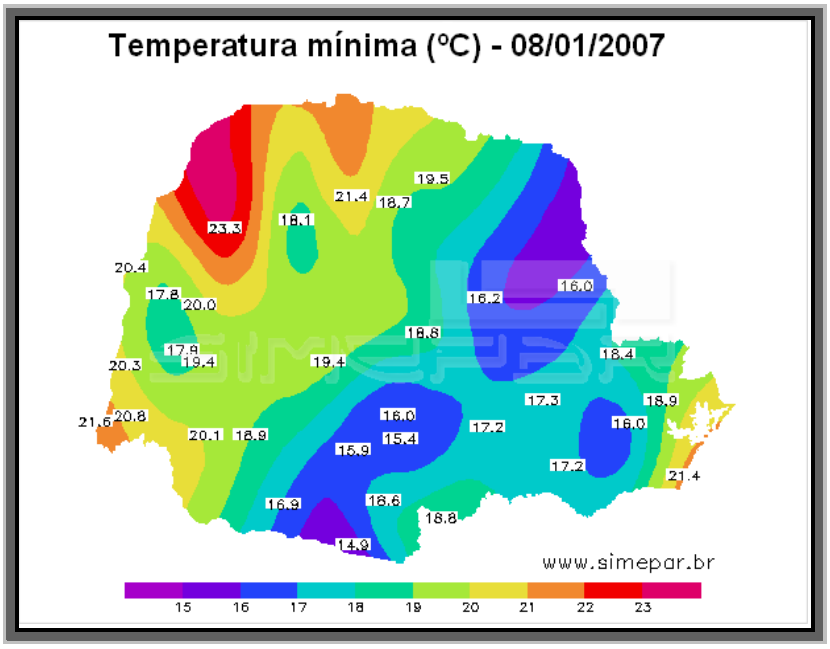


Figura 75 - Distribuição da temperatura mínima no Paraná (SIMEPAR, 2008)

As estações hidrológicas fornecem dados de chuvas e níveis dos rios, a partir dos quais se estima as suas vazões. As estações hidrológicas também fornecem dados de muita importância para operação de reservatórios de usinas hidrelétricas, como a de Foz do Areia, que opera com um medidor de nível a base de sonar.

O Sistema de Detecção e Localização de Descargas Atmosféricas gera pesquisa científica e produtos destinados a aplicações na previsão de tempo, na análise e manutenção de sistemas elétricos de transmissão, de distribuição e na emissão de laudos de análise de eventos severos para seguradoras e empresas de engenharia.

Beneti *et al* (2000), num convênio de cooperação técnico-científica firmado entre a Companhia Paranaense de Energia (COPEL) através do SIMEPAR, a CEMIG e FURNAS, tornou possível a integração dos sistemas de detecção de descargas atmosféricas operados por estas empresas formando a Rede Integrada de Detecção de Descargas Atmosféricas no Brasil (RIDAT), cujo objetivo é desenvolver um intercâmbio das informações técnico-científicas, e dos sinais obtidos pelos sensores das redes de detecção, além de integrar os procedimentos de análise, manutenção e operação conjuntas.

Ao todo, o RIDAT possui 22 sensores de descargas elétricas espalhados pelo Brasil, sendo, 7 do CEMIG, 6 do SIMEPAR, 8 de FURNAS e 1 do INPE, que possibilitam a criação de diversos sistemas de visualização, dentre os quais: (BENETI *et al*, 2000)

- Localização geográfica e temporal de descargas atmosféricas nuvem-terra;
- Localização de temporais
- Determinação de características de descargas como: valor estimado do pico da corrente de retorno, polaridade e número de componentes (multiplicidade) se a descarga for de natureza múltipla.

A figura 76, gerada pelo *software* SisRaios, mostra a disposição de descargas atmosféricas espalhadas em grande parte de Minas Gerais, Goiás,

Espírito Santo, São Paulo, Mato Grosso do Sul e Paraguai, num total de 534 descargas no período de 15 minutos a partir das 18:07 horas do dia 08 de janeiro de 2008. Pontos azuis mais escuros representam descargas mais recentes.

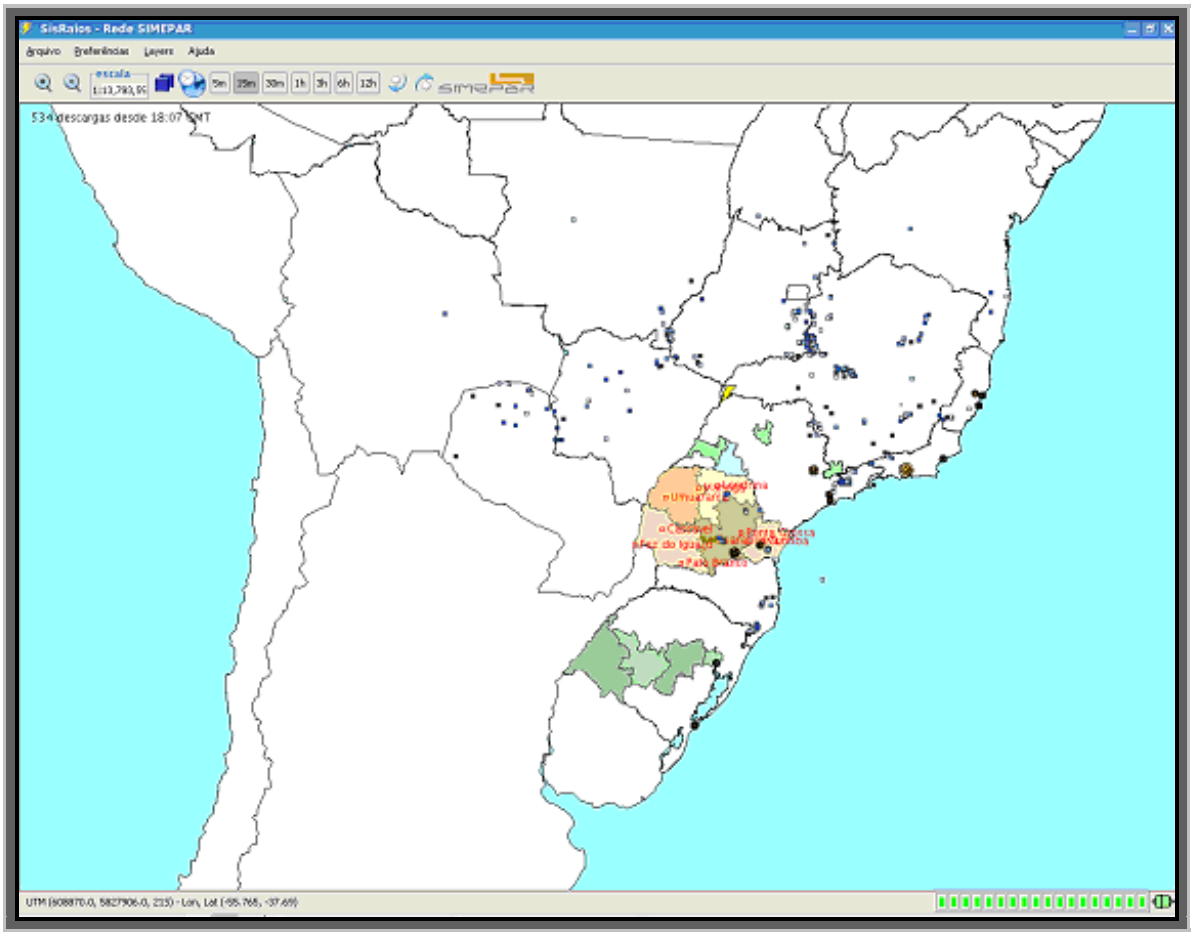


Figura 76 - Detecção de descargas atmosféricas no Brasil (fonte: SIMEPAR)

O Radar Meteorológico *Doppler* (nome dado para a técnica de captura de informação pelo radar) do SIMEPAR está em operação de monitoramento e vigilância ambiental desde outubro de 1998, sendo o primeiro do estado do Paraná. As medições do Radar são realizadas numa área de até 480 km de raio, cobrindo a área do estado do Paraná, Santa Catarina, parte do centro-sul de São Paulo e norte do Rio Grande do Sul. Os dados do Radar do SIMEPAR são obtidos em tempo real (a cada 10 minutos e em dias chuvosos a cada 5 minutos)

para monitoramento e previsão de tempo e armazenados para serem utilizados em pesquisa e desenvolvimento de produtos meteorológicos.

Em geral, os dados provindos do Radar Meteorológico Doppler do SIMEPAR são utilizados no monitoramento em curtíssimo prazo (0 a 3 horas) da precipitação, vento e granizo em eventos de tempo severo (tempestades, chuvas intensas, ventos fortes, ocorrência de granizo, descargas atmosféricas), além de permitir estimar a intensidade da chuva com grande resolução espacial e temporal (CALVETTI *et al*, 2003).

A figura 77 ilustra o uso do *software* RadVis, onde é possível observar a intensa chuva ocorrente naquela data (24/07/2007 às 23:41). A visualização é baseada nas técnicas de *Visualização da Informação*, vistos com mais detalhes no capítulo 5, onde os pixels são mapeados conforme valores da matriz de dados.

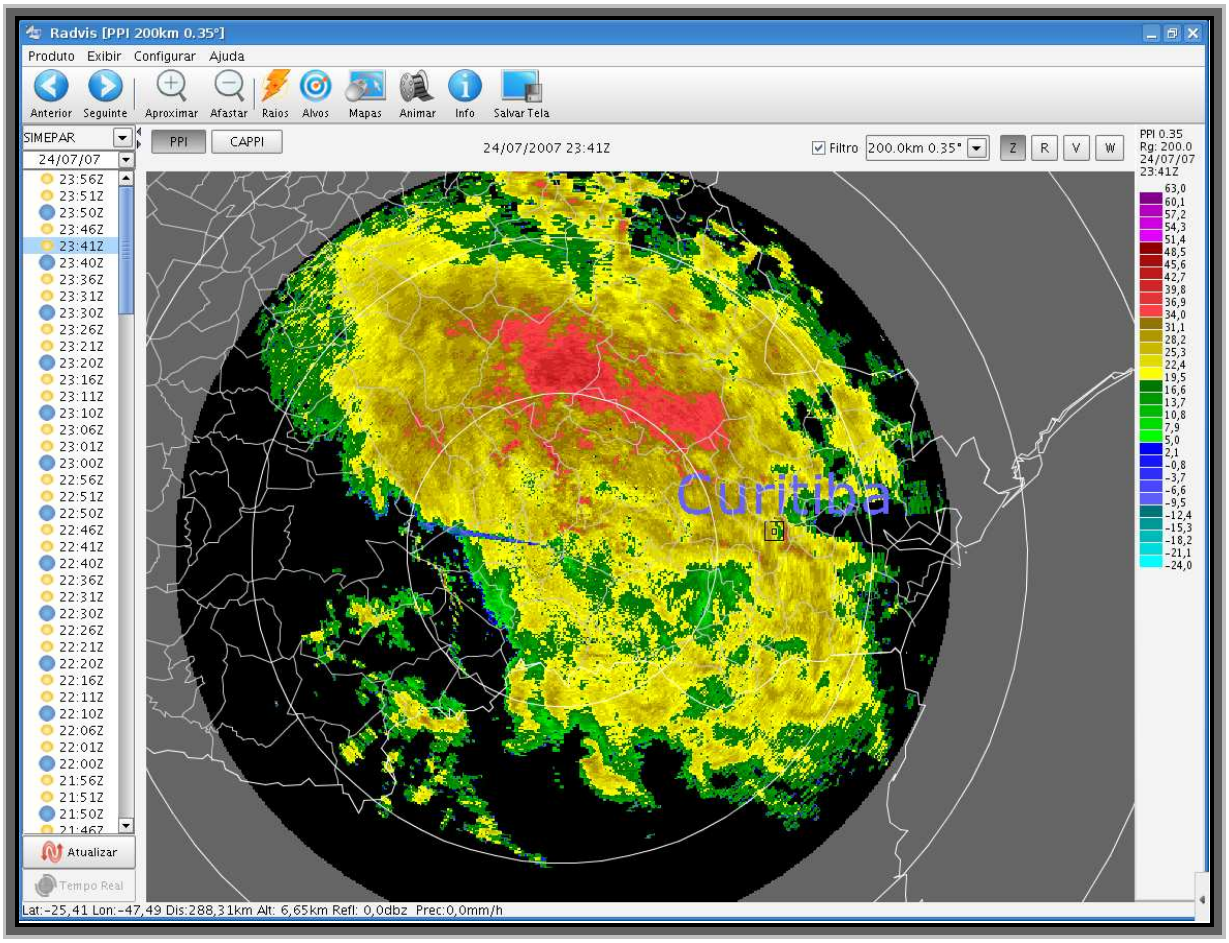


Figura 77 - Visualização de dados de radar através do RadVis (fonte: SIMEPAR)

Na próxima seção será estudado mais detalhadamente o funcionamento dos radares e a manipulação dos dados, pois estes serão posteriormente utilizados na aplicação de predição de chuvas.

Outra fonte de dados disponível no SIMEPAR é a recepção e o processamento de imagens de satélites, responsável pela recepção em tempo real de dados de alta resolução dos satélites meteorológicos das séries GOES e NOAA, bem como pelo processamento e visualização de dados e produtos de satélites.

As informações geradas pelos sistemas são usadas por meteorologistas no monitoramento e previsão do tempo, como também para atividades de pesquisa em modelagem numérica da atmosfera.

O processamento das imagens de satélite disponibiliza diversos produtos, assim classificados:

- Produtos GOES: imagens de infra-vermelho, vapor d' água e visível a cada 3 horas (mínimo), algoritmos de detecção de nevoeiros e queimadas, estimativa de precipitação por satélite;
- Produtos NOAA: índices vegetativos, temperatura da superfície do mar, temperatura do solo e perfil vertical de temperatura e umidade, entre outros.

As informações dos satélites são armazenadas no ambiente de banco de dados do SIMEPAR e estão disponíveis para todos os usuários da sua *homepage* (SIMEPAR, 2008).

As figuras 78 e 79 mostram o uso do *software* SatVis na visualização de dados de satélite na geração de imagens do canal infra-vermelho (IR4) do dia 08 de janeiro de 2008 às 05:45 horas. A primeira imagem numa escala preta e branca e na segunda numa escala colorida. Assim como o RadVis, o SatVis também mapeia os pixels conforme os valores da matriz de dados.

6 MÉTODOS DE PESQUISAS E EXPERIMENTOS

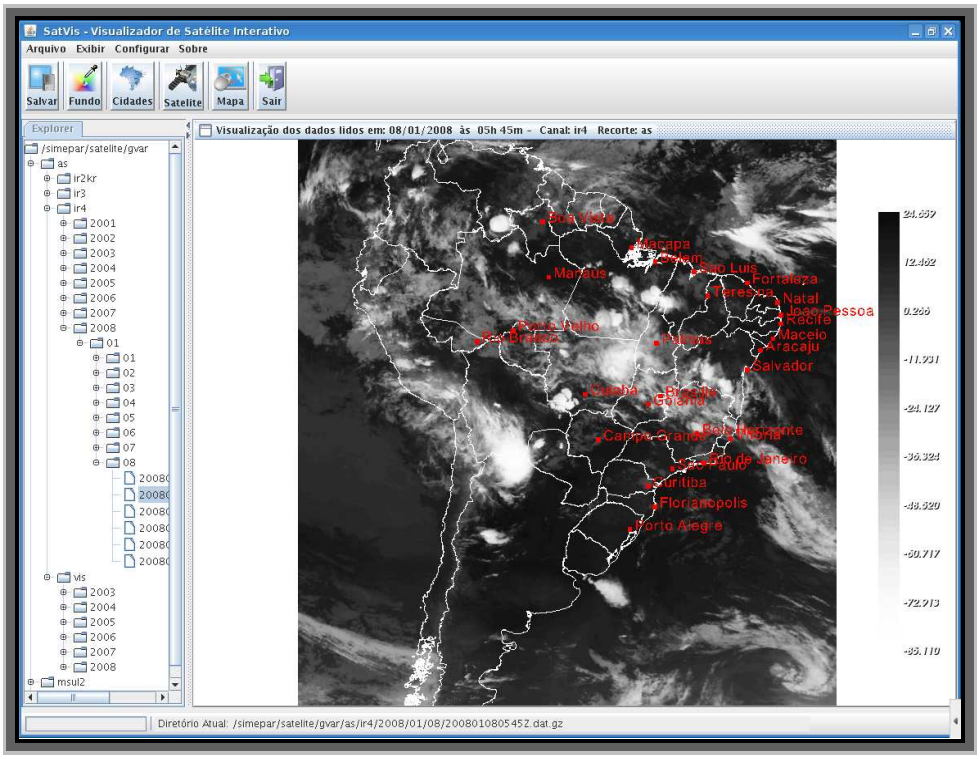


Figura 78 - Visualização da imagem de satélite da América do Sul pelo SatVis usando uma escala preto e branco (fonte: SIMEPAR)

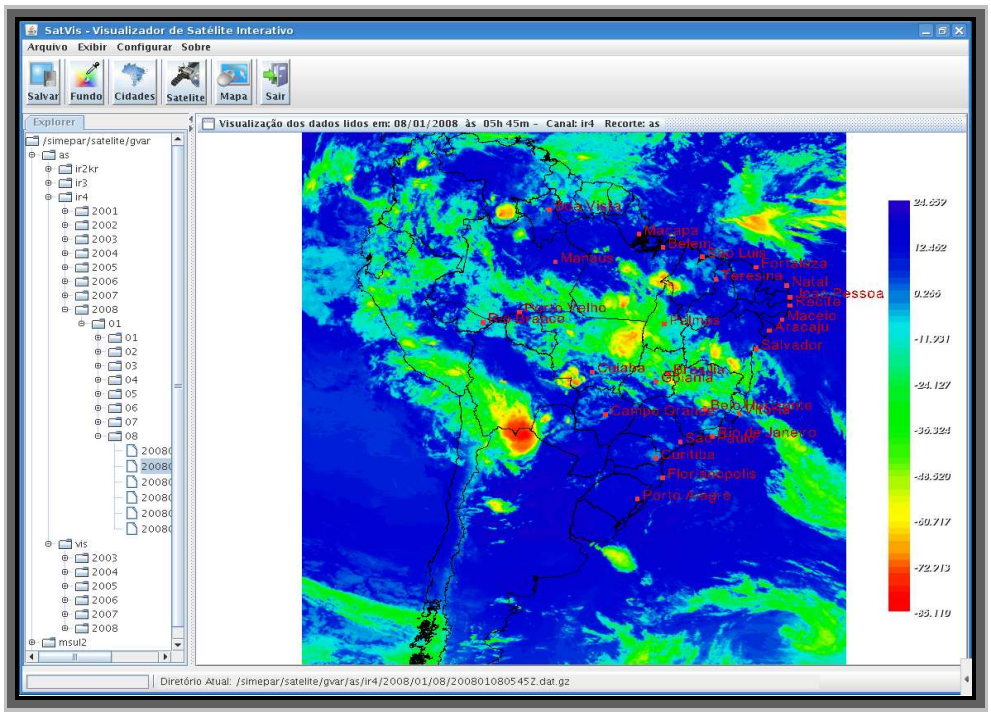


Figura 79 - Visualização da imagem de satélite da América do Sul pelo SatVis usando uma escala colorida (fonte: SIMEPAR)

6.3.2 O Radar Meteorológico

O princípio de funcionamento do radar meteorológico é análogo ao sistema de navegação de um morcego. O morcego emite sons em alta freqüência que ao ser interceptados por obstáculos retornam ao ouvido do morcego. O tempo levado para o som retornar é informação necessária para que o morcego saiba quão distante ele está do obstáculo.

No radar meteorológico, ao invés de sons, são empregadas ondas eletromagnéticas de alta energia capazes de atingir grandes distâncias. Estas ondas causam uma ressonância na freqüência da onda incidente em cada gota de chuva, de forma a irradiar ondas eletromagnéticas em todas as direções. Parte destas ondas são então retornadas ao radar, e através de informações conhecidas como o momento que o feixe de onda foi emitido pelo radar e quanto tempo depois o sinal retornou, é possível determinar a distância do alvo ao radar. Além disso, com informações de elevação da antena e o azimute correspondente, pode-se determinar com precisão a região espacial onde está chovendo (ver figura 80).

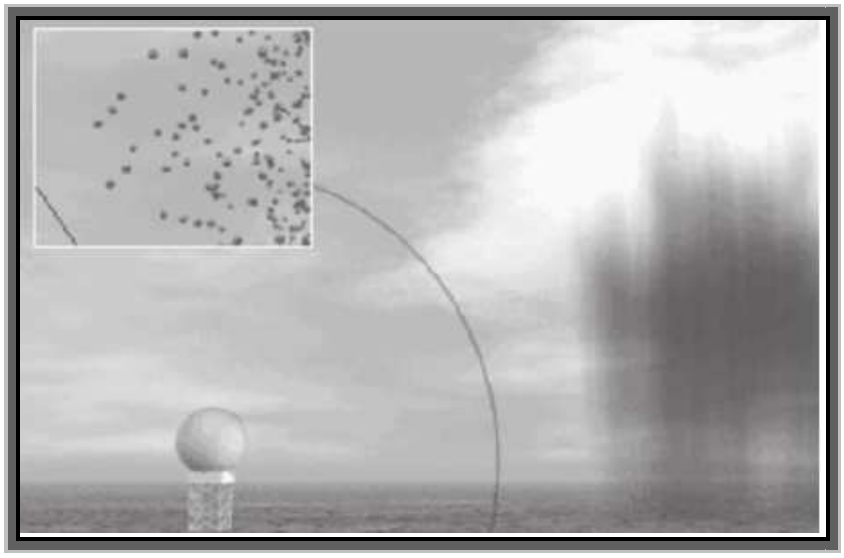


Figura 80 - Funcionamento do Radar (PINHEIRO; VAZ; MARTINHAGO, 2005)

6 MÉTODOS DE PESQUISAS E EXPERIMENTOS

O radar meteorológico do SIMEPAR é um radar Banda S *Doppler* (nome dado para a técnica de captura de informação pelo radar), modelo DWSR-95S, da empresa EEC *Corporation* e está em operação na região central do estado, no município de Teixeira Soares, nas coordenadas espaciais (-25.505313, -50.361330). A antena do radar DWSR-95S, de 8.2m de diâmetro, gera um feixe de ondas eletromagnéticas com aproximadamente 0.9º de abertura e monitora continuamente a atmosfera com uma seqüência pré-programada de varreduras azimutais em 360º. Um volume de varredura corresponde a uma seqüência completa de várias varreduras azimutais com diferentes elevações da antena.

O sistema de aquisição dos dados está configurado de tal forma a permitir uma avaliação de alta resolução espacial (área mínima de $4km^2$) e temporal (10 minutos) das informações de precipitação e vento. O radar é programado para fazer uma varredura num raio de 480 km para uma elevação de 0 grau e varredura com raio de 200 km para cada uma das elevações. O radar do SIMEPAR está calibrado para detectar gotículas de chuvas muito pequenas (garoas) até gotículas muito volumosas (presença de tempestades). As características técnicas do radar e do sistema de medição e processamento dos dados podem ser obtidas em Beneti, Nozu e Saraiva (1998).

O radar permite a medição de três variáveis. A *Refletividade* (Z) é o fator de refletividade entre a irradiação emitida pelo radar e a recebida por ele depois de espalhadas pelas gotas de chuvas presentes na atmosfera.
A unidade utilizada é o dBZ, que é uma escala logarítmica da refletividade. Os valores variam de zero e 60 dBZ e quanto maior forem esses valores, maior serão os diâmetros das gotas presentes no volume medido e, conseqüentemente, maior será a intensidade de precipitação (SIMEPAR, 2008).

A *Velocidade Radial* (V) mede a velocidade de aproximação ou afastamento dos alvos (gotas de chuvas) em relação ao radar na direção do feixe. O vento radial é mostrado em m/s com valores positivos para os alvos que se afastam e valores negativos para os alvos que se aproximam do radar.

A *Largura Espectral* (W) mede o desvio padrão das medidas obtidas em cada amostra (pixel) e, meteorologicamente, indica a turbulência nesse volume conforme a variância dos alvos presentes é medida em m/s.

A figura 81 ilustra as imagens geradas pelo RadVis no dia 24 de julho de 2007 às 23:41 horas para cada uma destas variáveis, em (a) a *refletividade*, em (b) a *velocidade radial* e em (c) a *largura espectral*.

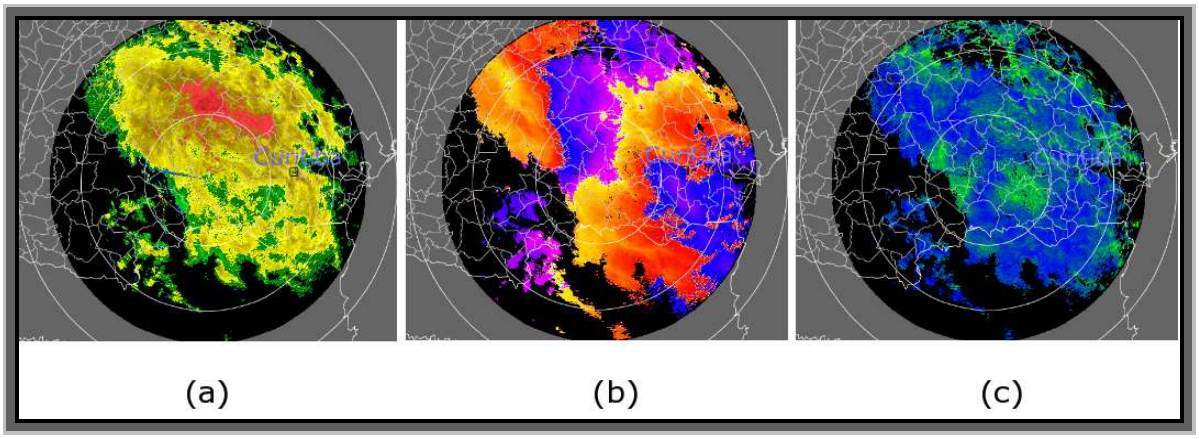


Figura 81 - Ilustração das imagens de radar para as variáveis (a) *refletividade*, (b) *velocidade radial* e (c) *largura espectral* (fonte: SIMEPAR)

Através dos valores (V, Z, W) é possível a visualização das imagens de radar em diferentes modos, dos quais os mais utilizados são o modo PPI, o CAPPI e o RHI (SIMEPAR, 2008; PINHEIRO; VAZ; MARTINHAGO, 2005).

O modo PPI (*Plan Position Indicator*) fornece a projeção num plano horizontal dos dados de *refletividade*, taxa de precipitação, *velocidade radial* média ou *largura espectral*, obtido através de uma varredura em azimute em coordenadas polares, para um ângulo de elevação determinado.

O CAPPI (*Constant Plan Position Indicator*) é o produto que fornece a projeção em um plano horizontal constante pré-definido, dos dados de *refletividade*, taxa de precipitação, *velocidade radial* média ou *largura espectral*, obtidos através de uma varredura volumétrica.

O RHI (*Range Height Indicator*) é o produto que fornece a projeção num plano vertical que passa pelo centro do radar dos dados de *refletividade*, taxa de

precipitação, *velocidade radial* média ou *largura espectral*, obtida através de uma varredura em elevação em coordenadas polares, para um ângulo de azimute determinado.

Os dados que são enviados pelo radar estão em formato numérico e depois são transformados em informações tais como: data, hora, localização, volume de chuva e a altura (em graus) da elevação da antena e uma matriz de dados de dimensão 512 x 512, onde estão armazenados os valores lidos pelas ondas eletromagnéticas do radar. As coordenadas que representam a localização da nuvem podem estar em coordenadas geográficas (latitude, longitude) ou em coordenadas UTM (*Universal Transverse Mecarcator*).

Para uma melhor percepção humana, os dados numéricos do radar são então convertidos em imagens, através do uso da técnica orientada a pixel da visualização da informação, onde os pixels são mapeados conforme os valores da matriz de dados.

6.3.3 Mineração Visual de Dados Aplicada às Imagens do Radar Meteorológico do SIMEPAR

Em algumas imagens geradas através dos dados de radar é possível observar que alguns pontos não correspondem à presença de chuva. Estes pontos são gerados na presença de algum alvo, que não as gotas d'água, dos quais as ondas eletromagnéticas do radar interpretam-os de forma errônea como sendo um dado de chuva, que geralmente ocorrem na presença de ruídos, ecos de terrenos, presença de nuvens de insetos, neves, raios de sol (no por ou no nascer do sol), através da propagação anômala das ondas eletromagnéticas.

Dentre estas possíveis causas, no radar meteorológico do SIMEPAR, estão presentes os ecos de terrenos e os ruídos, conforme podem ser vistos na figura 82. São estes pontos que se pretende eliminar através do treinamento de uma *Rede Neural* baseada em informações históricas conhecidas.

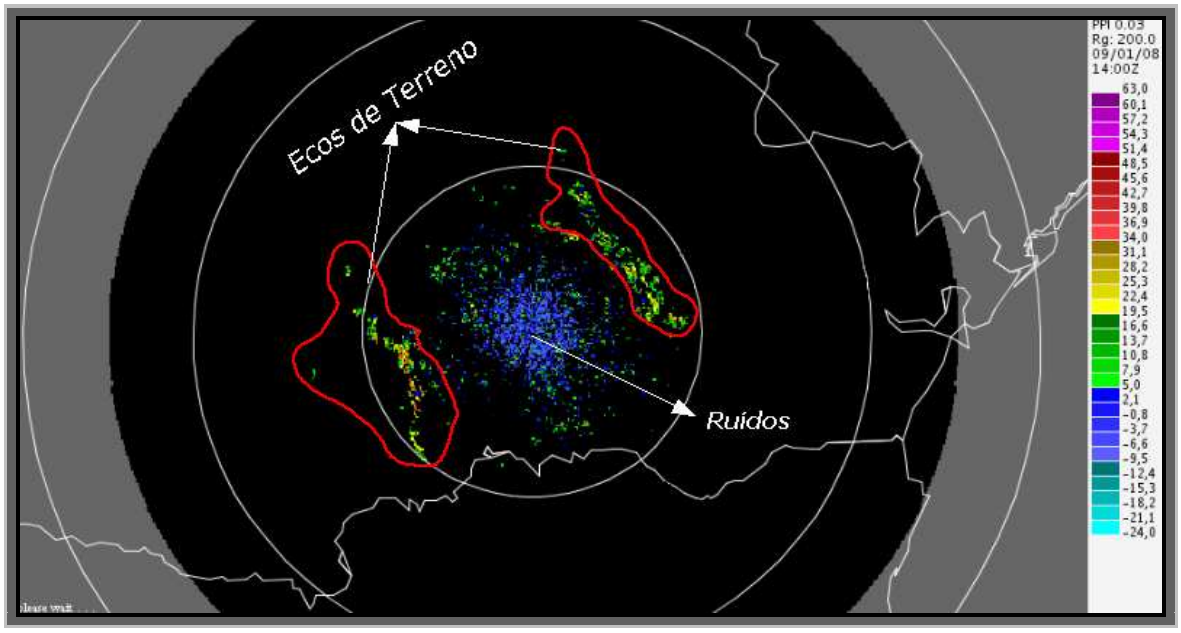


Figura 82 - Tipos de informações que não representam chuvas encontrados nas imagens do radar meteorológico do SIMEPAR (fonte: SIMEPAR)

O treinamento de uma *Rede Neural* baseada nas imagens de radar da base de dados do SIMEPAR possibilita o aprendizado da rede na busca do conhecimento. Neste caso, espera-se que a rede seja capaz de dizer com precisão se os dados mostrados nas imagens (pixel a pixel) são dados de chuva ou não. Esta avaliação permitirá a exclusão destes pontos das imagens, facilitando na interpretação dos meteorologistas.

Lakshmanan *et al* (2006) propõem o uso de *Redes Neurais* na filtragem dos dados estranhos das imagens de radares meteorológicos. O algoritmo proposto pelos autores utiliza como entrada para a *Rede Neural* 27 características extraídas dos dados, dentre as quais estão propriedades do terreno, e análise estatísticas da vizinhança dos pixels, permitindo de uma forma genérica excluir os pontos que não representam chuva oriundos das diferentes fontes.

No SIMEPAR, os dados de radar, que originalmente são recebidos em coordenadas polares e de forma volumétrica, são convertidos para coordenadas cartesianas e calculada as elevações separadamente. Sendo assim, uma implementação simplificada no algoritmo proposto por Lakshmanan *et al* (2006) foi realizada.

A figura 83 mostra o processo executado até a obtenção da imagem final, onde os pixels da imagem serão reclassificados, eliminando aqueles que não representam chuva.

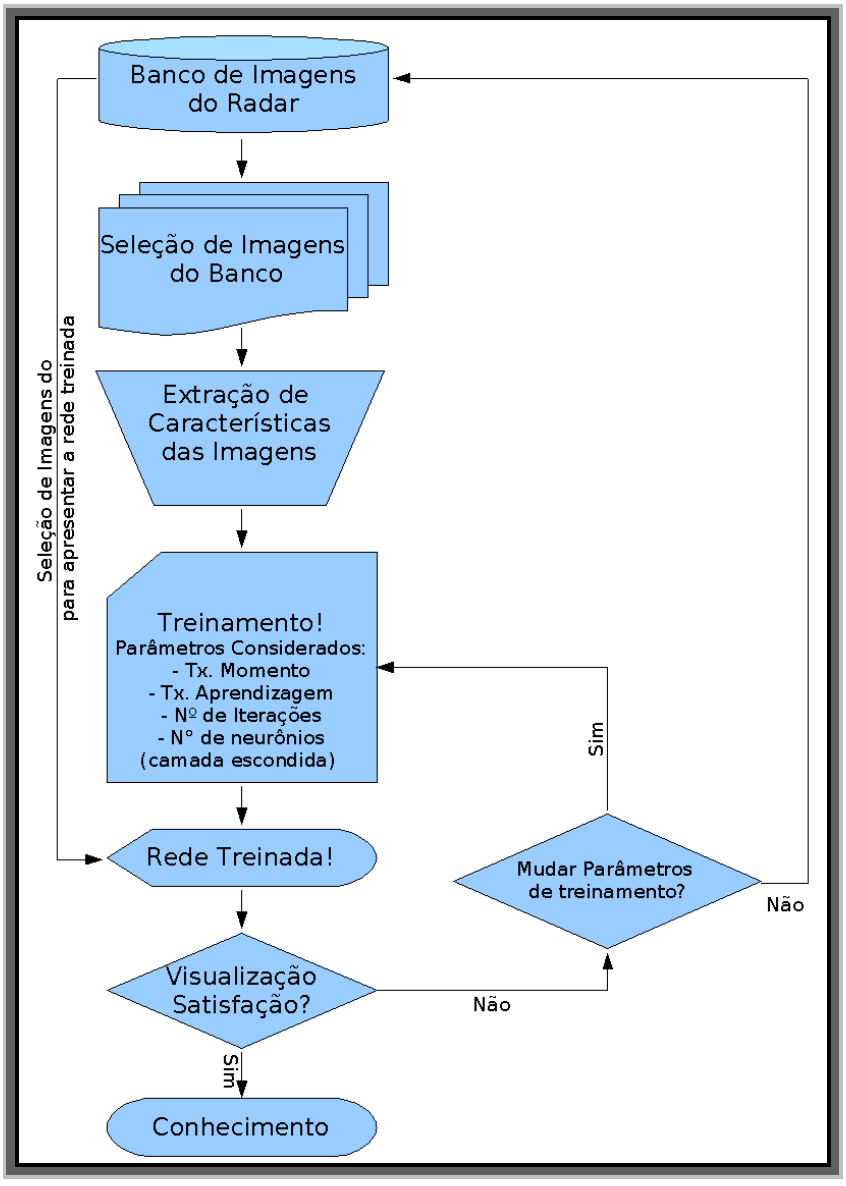


Figura 83 - Mineração Visual de Dados: Algoritmo de mineração de dados com a inserção da visualização em busca da filtragem das imagens de radar

A base de dados do radar meteorológico do SIMEPAR possui informações armazenadas desde a sua implementação, em 1998. Em curtíssimo prazo, cerca de 20 minutos, uma nova imagem pode ser visualizada. Numa primeira tentativa

de filtrar os dados de chuva, uma seleção manual de dez imagens deste banco de dados foram usada para serem apresentadas à rede. Nesta seleção, três imagens possuíam dados de chuva e ruídos, duas com predominância de chuva e cinco imagens de ruídos (sem chuva). Baseada nas experiências dos meteorologistas do SIMEPAR, os pixels pintados de branco foram classificados como sendo dados ruins (não representam chuva). As imagens da figura 84 mostram um exemplo dos tipos de imagens selecionadas e sua respectiva classificação realizada pelos meteorologistas.

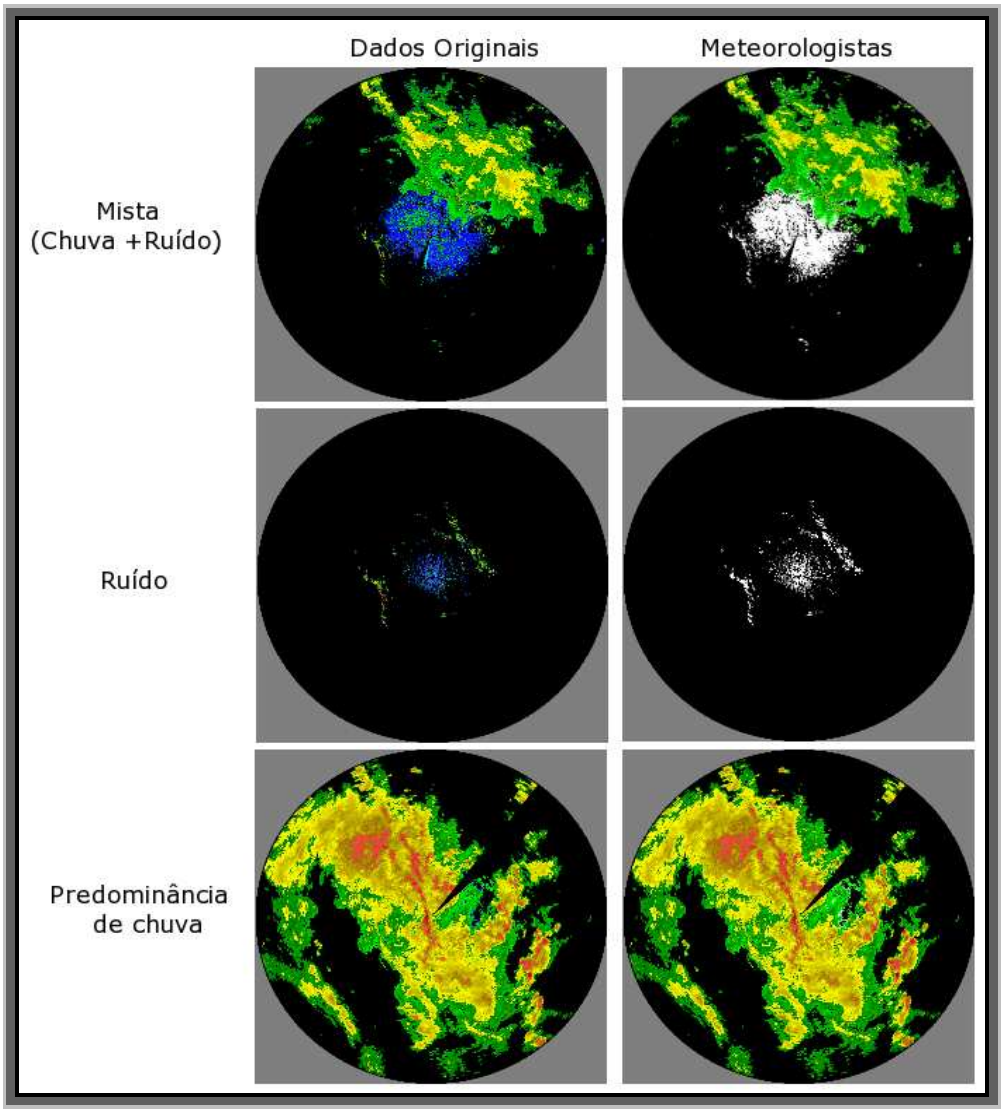


Figura 84 - Classificação dos pixels das imagens como sendo de ruído (branco) (Fonte: SIMEPAR)

Para cada imagem selecionada uma série de informações baseadas na vizinhança dos pixels são calculados. A vizinhança de um pixel é formada pelos pixels vizinhos a este, conforme mostra a figura 85.

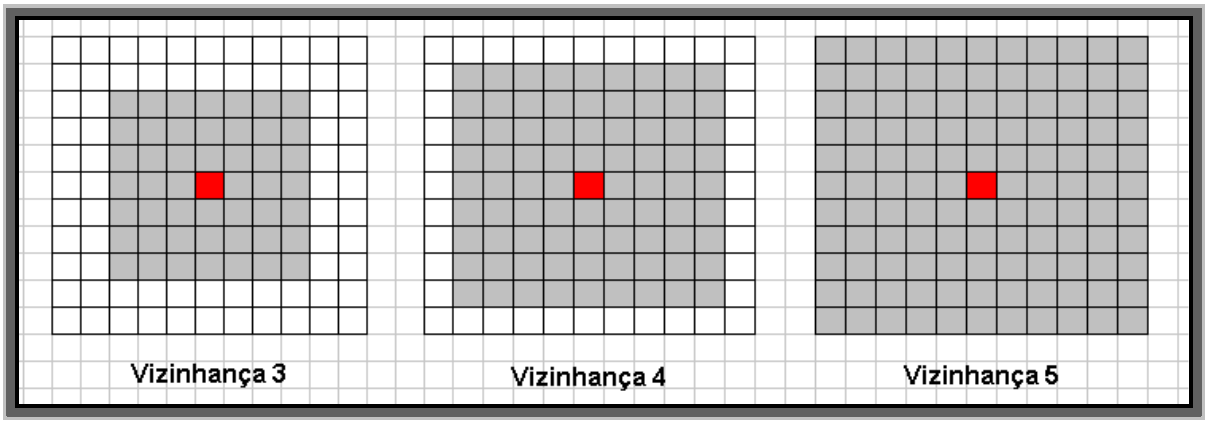


Figura 85 - Vizinhança de um pixel

Neste trabalho, para cada pixel de cada imagem (no total 512 x 512) são analisadas as informações de sua vizinhança de raio de valor igual a “1” e delas são extraídas informações de média, desvio padrão e mediana. Cada pixel agora possui seis características (média, mediana, desvio padrão, valor do pixel, posição *i* na matriz de dados e posição *j* da matriz).

Estas características são apresentadas à rede (pixel a pixel), cuja saída é conhecida (“0” representa chuva e “1”, caso contrário). Baseadas na teoria de *Redes Neurais* vistas no capítulo 4 e com base nos resultados obtidos num comparativo àqueles definidos pelos meteorologistas e após diversos testes, a melhor topologia de rede encontrada nos testes realizados é de seis neurônios na camada de entrada, dez neurônios na camada escondida e um na camada de saída que classifica se o dado é chuva (próximos de “0”) ou dado estranho (próximos de “1”). Nas aplicações que seguem, foi utilizada taxa de aprendizado igual a “0.7” e taxa de momento igual a “0.8”. Notou-se também que após “500” iterações, não houve grande melhoria no treinamento da rede e, por conseguinte, foi utilizado este número de iterações nos treinamentos.

Para novos dados apresentados à rede, consideraram-se saídas menores que “0.5” como sendo dados bons (que representam chuva) e maior ou igual a “0.5” como sendo dados que não representam chuva. Conforme utilizado em Lakshmanan *et al* (2006), neste trabalho também se optou pelo uso da função de ativação tangente hiperbólica na camada escondida e a sigmóide na camada de saída. Os pesos iniciais são gerados diretamente pela biblioteca *Joone*[10] usada na programação do método. A figura 86 mostra a topologia da rede usada nas aplicações.

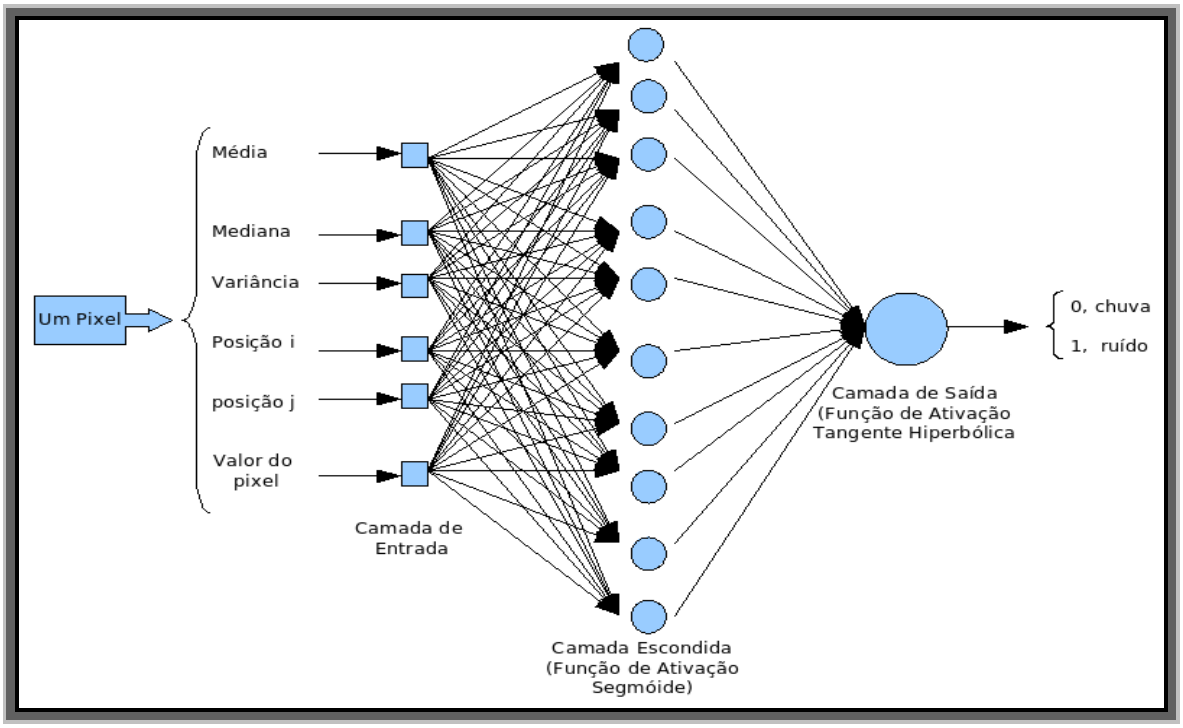


Figura 86 - Topologia de rede neural usada nas aplicações

Após a rede ter sido treinada, novas imagens podem ser apresentadas à rede e então classificadas por ela. Uma análise visual permite a extração do conhecimento. A figura 87 mostra alguns resultados obtidos.

[10] **Joone**: Biblioteca para desenvolvimento de aplicações baseadas em Redes Neurais desenvolvida em linguagem JAVA. Disponível em: http://www.jooneworld.com/

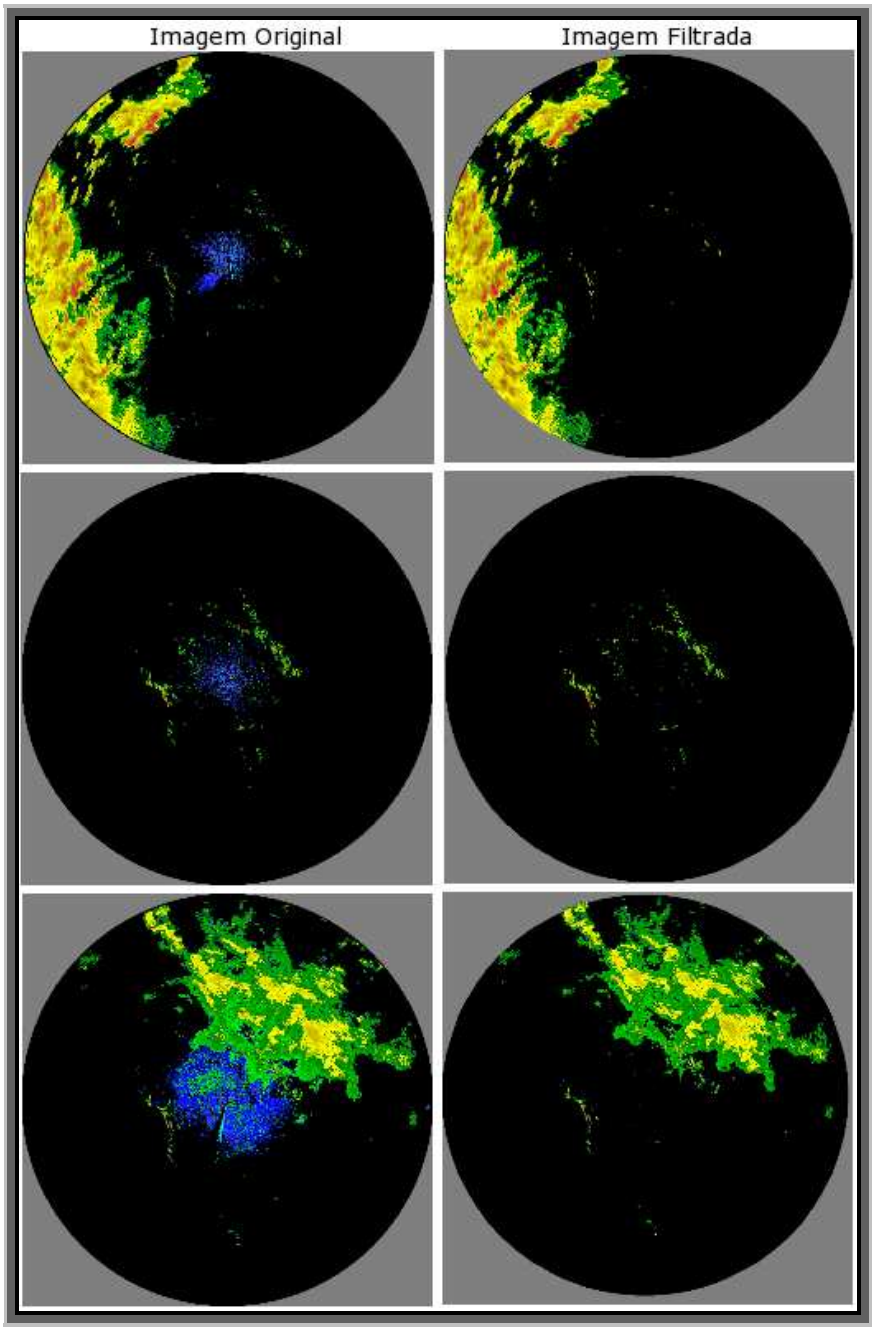


Figura 87 – Imagens filtradas após apresentação à Rede Neural

Uma análise nestes resultados permitiu observar que os dados de ruídos das imagens do radar foram eliminados, porém com os ecos de terrenos, não se teve o mesmo sucesso. Este fato provavelmente ocorreu, pois estes pixels possuem características semelhantes aos de chuva. Os dados de entrada da rede

(média, mediana, desvio padrão, valor do pixel e posição *i* e *j*) não foram suficientes para detectar os ecos de terreno.

Para tanto, optou-se pelo treinamento de duas redes separadamente, uma considerando somente ecos de terrenos e outra considerando os ruídos, cujo conhecimento já foi alcançado pelo treinamento da rede anterior.

Este tipo de topologia de rede é conhecido na literatra, segundo Oh e Suen (2002), por Classe-Modular (Class-Modular), um exemplo de arquitetura do tipo paralela utilizando a técnica de voto para combinar as saídas dos agentes. Os autores propõem o conceito de modularidade em classes utilizando redes neurais no reconhecimento de escritas.

Nesta topologia da rede baseada em Classe-Modular, foi considerado uma vizinhança de raio cinco a fim de capturar informações mais precisas do comportamento dos pixels vizinhos. Além disso, para a rede de treinamento dos ecos de terreno, uma nova característica baseada em duas informações de grande importância para o treinamento foi adicionada.

A primeira, denominada por região do pixel, indica se o pixel em análise pertence à região onde costuma existir ecos de terrenos. A figura 88 mostra a região de maior incidência de ecos de terrenos, obtida pela sobreposição de mais de 200 imagens.

Figura 88 - Região de costume de ecos de terreno

A segunda, numa adaptação de uma das características usadas por Lakshmanan *et al* (2006), o *Spin,* foi utilizado para analisar a diferença entre pixels adjacentes cuja diferença de refletividade era superior a 2 dBZ dividido pelo maior número possível nesta vizinhança (STEINER; SMITH, 2002).

Sendo assim, essa nova entrada, possui valor “1” para pixels que se encontram na região de maior incidência de ecos de terreno e parâmetro *spin* maior que “0.5”, e “0” caso contrario.

A figura 89 mostra dois resultados obtidos pelo treinamento destas redes. A primeira coluna mostra as imagens vindas do radar; a segunda, um exemplo da seleção dos ruídos; a terceira, a seleção dos ecos de terreno e a última coluna, a filtragem das imagens iniciais que foram apresentadas à rede, eliminando os pontos que não representam chuva.

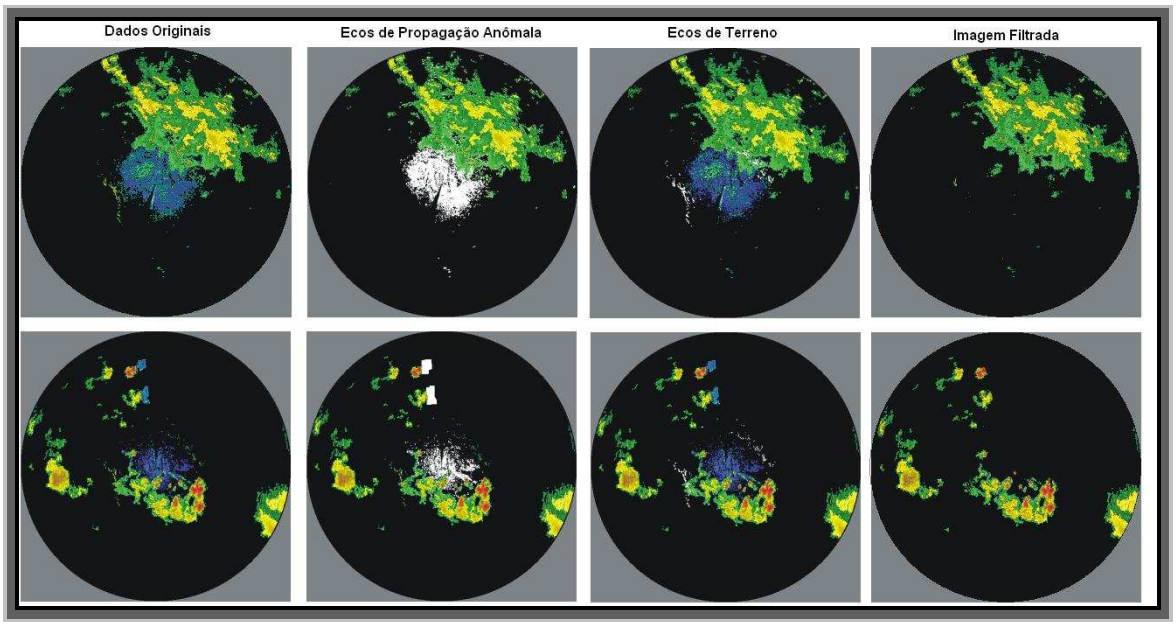


Figura 89 - Resultado obtido pelo treinamento de duas redes, uma para eliminar os ruídos e a outra para eliminar os ecos de terreno

Observe-se que os resultados esperados foram obtidos, ou seja, o treinamento de duas redes separadamente, uma para tratar os ruídos e outra para o tratamento dos ecos de terreno, obtiveram sucesso na detecção dos dados de

chuva. O uso das imagens selecionadas de radar geradas a partir de *Técnicas Orientadas a Pixel* integradas ao uso de *Redes Neurais* para predição de dados de chuvas e finalmente a apresentação final da imagem filtrada, fizeram a inserção da *Visualização* no *Processo KDD*, sendo esta uma aplicação de *Mineração Visual de Dados* do tipo fortemente integrado e de grande importância nos estudos realizados.

6.4 Considerações Finais

Os estudos de casos aqui apresentados foram importantes, do ponto de vista científico, para verificar a adequação de técnicas da MVD e efetivamente analisar visualmente os resultados gerados pelas aplicações abordadas. As técnicas de *Visualização da Informação* foram de grande valia para estes estudos, facilitando a interpretação dos resultados e na extração do conhecimento.

Em alguns casos, o uso de algoritmos de tratamento de dados (*Mineração de Dados*) como os mencionados no capítulo 4, usados juntamente com técnicas de *Visualização*, facilita a análise dos resultados, permitindo obter informações tais como o reconhecimento de padrões, análise de agrupamentos e relacionamento entre variáveis mais rápidos e facilmente.

Assim, no primeiro estudo de caso, diferentes técnicas de *Visualização da Informação* (VI) foram aplicadas aos dados da Barragem de ITAIPU, responsáveis pelo monitoramento de suas estruturas. Estas técnicas permitiram analisar de forma visual o relacionamento entre diferentes variáveis, de certo modo sem necessidade de conhecimentos aprofundados em áreas como estatística e matemática.

Devido às vantagens e desvantagens que uma técnica possui sobre a outra, notou-se a necessidade de utilizar várias técnicas de VI, e num comparativo entre elas pôde ser extraido o conhecimento desejado.

6 MÉTODOS DE PESQUISAS E EXPERIMENTOS

A *Mineração Visual de Dados,* neste estudo de caso, foi usada na ordenação dos eixos (variáveis) da técnica *Coordenadas Paralelas* conforme os valores das correlações existentes entre as variáveis envolvidas. Isso permitiu que variáveis cujo relacionamento fosse maior se encontrassem em eixos vizinhos de forma a facilitar a interpretação dos resultados.

Já no segundo estudo de caso, a Mineração Visual de Dados foi usada na filtragem de dados estranhos das imagens do Radar Meteorológico do SIMEPAR. Para tanto foram utilizadas duas redes neurais com topologias diferentes, uma para cada tipo, ruídos e ecos de terreno, baseada nas características dos pixels vizinhos.

Neste caso, um estudo mais aprofundado deve se realizado na tentativa de suprir as falhas na extração de ruídos ou ecos de terrenos. Já que, embora os resultados obtidos tenham sido satisfatórios, as informações usadas para o treinamento da rede não foram suficientes para filtrar com perfeição os dados, que em alguns casos houve a eliminação de pequenas quantidades chuvas.

Em ambos os estudos de casos, os resultados atingiram as expectativas esperadas e as técnicas de *Visualização da Informação* integradas às de *Mineração de Dados* mostraram ser eficientes no ponto de vista computacional e facilidade de uso.

No entanto, devido à grande quantidade de técnicas de *Visualização* que podem ser usadas com este propósito, a escolha daquelas que melhor se ajustam aos dados em análise não é uma tarefa fácil e devem ser escolhidas com base no que se está buscando. Diferentes técnicas levam a diferentes resultados, e uma comparação entre elas podem ser fundamentais para extração do maior número possível de informações.

A alta dimensionalidade e o alto número de registros contidos nos dados podem ser um problema crucial na aplicação de técnicas visuais, visto que a maioria destas, mapeiam os dados na tela do computador, de forma que a visualização fica limitada pela resolução do monitor. Em muitos casos, devido esta grande quantidade de dados a ser apresentados, as imagens ficam pequenas, dificultando a sua interpretação.

7 CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS

O tratamento de dados multidimensionais requer do usuário conhecimento em várias áreas científicas, dentre elas estão a *Estatística*, a *Inteligência Artificial*, a *Visualização*. Diversas técnicas oriundas destas diferentes áreas podem ser usadas para extração do conhecimento que na literatura ficou conhecida por *Processo KDD*, descoberta do conhecimento em bases de dados. Este processo passa por diversas etapas, incluindo a seleção e tratamento dos dados antes de se aplicar algoritmos específicos para um determinado propósito, como análise de agrupamento e análise de relacionamentos entre variáveis. Na última etapa, o conhecimento pode ser obtido pela análise dos resultados gerados.

A aplicação da *Visualização* no contexto de descoberta do conhecimento em base de dados (*Processo KDD*) é de extrema importância na interpretação dos resultados. Isso, devido ao grande poder de processamento de dados dos computadores atuais e a facilidade dos seres humanos no reconhecimento de padrões visuais.

A *Visualização* é o termo genérico para visualizar qualquer tipo de dado e pode ser dividida em duas grandes áreas. Quando aplicadas a dados com características espaciais, geralmente dados de natureza física (temperatura, velocidade, tempo), então a esta área se dá o nome de *Visualização Científica* (VC). Já quando os dados são de natureza abstrata, ou seja, não se conhece o comportamento no espaço destas informações, então esta área é a *Visualização da Informação* (VI).

7 CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS

A vertente *Visualização da Informação*, por ser uma área relativamente recente, tem seu processo de classificação em formalização. Diferentes autores definem e classificam uma determinada técnica seguindo diferentes critérios. Neste trabalho, procurou-se apresentar as diferentes técnicas de VI já consolidadas e que trazem distinções conceituais claras e intuitivas. Já para as técnicas pouco citadas na literatura, buscou-se uma classificação conforme características dos dados em estudo. Para cada técnica apresentada aqui, foram mostradas as suas vantagens e desvantagens, em relação à interação, processamento e capacidade de apresentar informações das diversas variáveis ao mesmo tempo. Esta categorização sistemática das técnicas de *Visualização da Informação*, mostrando suas vantagens e desvantagens, é uma contribuição deste trabalho.

Como visto, outra área, ainda mais recente, é a Mineração Visual de Dados, como uma tentativa de integrar a Visualização à Mineração de Dados (etapa do *Processo KDD*, responsável por gerar o conhecimento). Esta área merece destaque devido aos experimentos e resultados obtidos na integração daquelas duas outras áreas, gerando bibliografia especializada (SOUKUP; DAVIDSON, 2002) e motivação para novos estudos que conseqüentemente aceleram o processo de definição da Mineração Visual de Dados. Desta forma, o levantamento do referencial bibliográfico e a forma de como estas áreas, *Visualização* e *Mineração de Dados*, podem ser integradas seriam a segunda contribução aqui deixada.

Como terceira e quarta contribuições estão as aplicações das técnicas de *Visualização da Informação* e *Mineração de Dados* aos Experimentos aqui abordados.

No primeiro experimento, ITAIPU, diferentes técnicas de *Visualização da Informação* foram utilizadas com o propósito de encontrar relacionamentos entre variáveis responsáveis pelo monitoramento da barragem. Neste caso, observou-se a necessidade de comparar várias destas técnicas para extração do maior número possível de informações. Isso ocorre devido às limitações que umas técnicas têm sobre as outras.

A técnica *Coordenadas Paralelas*, por exemplo, permite a análise de correlação só para os eixos (variáveis) vizinhos. Formas interativas de troca dos

eixos podem ser utilizadas para analisar par a par, porém quando o número de variáveis é muito grande, esta troca de eixos se torna inviável devido ao comportamento combinatorial. Porém, quando se deseja saber quais são as variáveis mais (ou menos) correlacionáveis, um pré-processamento ordenando os eixos através de uma análise estatística das correlações entre as variáveis envolvidas, pode ser fundamental para facilitar a análise dos resultados.

O segundo experimento foi aplicado na filtragem dos dados do radar meteorológico do SIMEPAR, responsável pelo monitoramento de chuvas. Nas imagens geradas pela coloração dos pixels conforme os valores da matriz de dados é possível observar que alguns pontos não correspondem como sendo dados de chuva. Estes dados estranhos, no caso do radar do SIMEPAR, ocorrem devido à presença de ruídos e ecos de terreno. Para tanto, uma rede neural baseada em imagens cujos dados que não representam chuva foram pintados de branco, seguindo orientações dos meteorologistas do SIMEPAR, foi treinada para cada um dos tipos de dados estranhos encontrados nas imagens (ruídos e ecos de terrenos). Informações estatísticas dos pixels vizinhos para cada pixel da matriz foi utilizada como entrada para rede. Diversas topologias de rede foram utilizadas até se obter aquela cuja imagem gerada após o treinamento era mais satisfatória.

No caso do radar, o uso das imagens selecionadas geradas a partir de *Técnicas Orientadas a Pixel* integradas ao uso de *Redes Neurais* para predição de dados de chuvas e finalmente a apresentação final da imagem filtrada permitiram que se demonstrasse a inserção da *Visualização* no *Processo KDD*. Assim, esta pode ser entendida como uma aplicação de *Mineração Visual de Dados* do tipo fortemente integrado.

Em ambos os estudos, os objetivos foram alcançados, mostrando que as técnicas de *Visualização* contribuíram para a análise visual dos resultados. Em geral, o resultado foi muito mais satisfatório quando a *Visualização* foi integrada à *Mineração de Dados*. No primeiro caso, ITAIPU, a *Mineração Visual de Dados*, foi aplicada de modo integrado com a *Visualização da Informação,* com a ordenação dos eixos pela técnica de Coordenadas Paralelas através dos valores das correlações entre as variáveis. Já no segundo caso, SIMEPAR, a utilização das imagens classificadas pelos meteorologistas para o treinamento de duas redes

neurais, uma para tratar ruídos e outra para tratar os ecos de terreno, e finalmente a apresentação à rede treinada de uma nova imagem, foi à aplicação de *Mineração Visual de Dados* na filtragem dos dados que não representam chuvas.

Para trabalhos futuros, são sugeridas a implementação de um *software*, capaz de ler um arquivo de dados e numa única tela mostrar várias das técnicas de *Visualização da Informação*, facilitando a comparação entre elas. Este recurso é fundamental na análise dos dados e não foi encontrado em nenhum dos *softwares* pesquisados. Recursos adicionais, como seleções de regiões, agrupamento por cores e formas novas de interações podem ser utilizados.

Além disso, no caso do experimento da ITAIPU, um estudo mais aprofundado poderá ser realizado. A idéia é fazer uma análise visual dos demais tipos de intrumentos existentes no bloco F19/20 e dos demais blocos da barragem, misturando numa mesma visualização tipos diferentes de instrumentos. Isso permitirá observar a relação existente entre instrumentos diferentes e em diferentes posições da barragem.

Já no caso do experimento do SIMEPAR, observou-se que em alguns casos, quando a chuva estava presente em posições onde possuíam ecos de terrenos, a rede optava por classificar estes pontos como sendo pontos ruins, deixando buracos no interior da chuva. Este problema poderá ser resolvido, em trabalhos futuros, por exemplo, através de uma interpolação dos pixels vizinhos preenchendo estes espaços sem valores, ou adicionando novas infrmações na entrada para o treinamento da rede, como informações topográficas da região do radar.

Vale salientar que a avaliação do desempenho de classificação da rede foi feito através da visualização, onde após o treinamento, foi posível observal visualmente a filtragem dos dados que não representariam chuvas. Desta forma, sugere para trabalhos futuros desenvolver o treinamente e fazer uma Validação Cruzada (*Cross-Validation*) para avaliar o desempenho da rede.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. **Fast Algorithms for Mining Association Rules**. Proc. 20th Int. Conf. Very Large Data Bases, VLDB, 1994.

AGRAWAL, R.; SRIKANT, R. **Mining Sequential Patterns**. In Proc. of 1995 Int. Conf. on Data Engineering, Taipei, Taiwan, março 1995.

AGRAWAL, R.; SRIKANT, R. **Mining Sequential Patterns: Generalizations and Performance Improvements**. Proc. 5th EDBT, 3-17, 1996.

ANDRAOS, N. C. **Organização, Análise e Mapeamento de Variáveis Relacionadas à Instrumentação da Barragem de ITAIPU.** In: 14° Evento de Iniciação Científica – EVINCE – UFPR, 2006, Curitiba. 14o. Evento de Iniciação Científica - EVINCI UFPR, 2006. v. 14. p. 460-460.

ANDREWS, K.; HEIDEGGER, H. **Information Slices: Visualization and Exploring Large Hierarchies using Cascading, Semi-Circular Discs**. Proceedings IEEE Symposium on InformationVisualization (InfoVis'98), outubro 1998.

ANKERST, M. **Visual Data Mining and Exploration of Large Databases.** Tutorial at PKDD'2001, Freiburg, Alemanha, 2001.

ANKERST, M. **Visual Data Mining.** Tese de Doutorado, Faculty of Mathematics and Computer Science, University of Munich, Munique, 2000.

ANKERST, M.; KEIM, D.A.; KRIEGEL, H.P. **Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets**. Proc. Visualization ‘96, San Francisco, Ca, 1996.

REFERÊNCIAS

ARTERO, A. O. **Estratégias para apoiar a detecção de estruturas em visualizações multidimensionais percentualmente sobrecarregadas.** Tese de doutorado do Instituto de Ciências Matemáticas e de Computação ICMC-USP, São Carlos, Brasil, 2005.

BATISTA, G. E. A. **Pré-processamento de Dados.** Em: Aprendizado de Máquina Supervisionado, São Paulo, 2003.

BENETI, C. A. A.; LEITE E. A.; GARCIA S. A. M.; ASSUNÇÃO L. A. R.; CAZETA FILHO A.; REIS, R. J. **RIDAT - Rede Integrada de Detecção de Descargas Atmosféricas no Brasil: situação atual, aplicações e perspectivas.** In: CONGRESSO BRASILEIRO DE METEOROLOGIA, 11, Rio de Janeiro, RJ. 2000

BENETI, C. A. A.; NOZU, I.; SARAIVA, E. A. **Monitoramento da precipitação e de eventos de tempo severo com radar meteorológico no estado do Paraná**. In: CONGRESSO BRASILEIRO DE METEOROLOGIA, 10., 1998, Brasília. CD-ROM.

BERTIN, J. **Graphics and Graphic Information Processing.** De Gruyter Publishers, 1981.
BESHERS, C.; FEINER, S. **Autovisual: Rule-Based Design of Interactive Multivariate Visualizations.** IEEE Computer Graphics & Applications, p. 41-49, julho, 1993.

BRANCO, V. M. A. **Visualização como Suporte à Exploração de uma Base de Dados Pluviométricos**. Dissertação de mestrado do Instituto de Ciências Matemáticas e de Computação ICMC-USP, São Carlos, Brasil, 2003.

BREJOVA, B.; DIMARCO, C.; VINAR, T.; HIDALGO, S.R.; HOLGUIN, G.; PATTEN, C. **Finding Patterns in Biological Sequences**. Project Report, Department of Biology, University of Waterloo, 2000.

BRODLIE, K. et al. **Scientific Visualization, techniques and applications**. Springer-Verlag, 1992.

BURIOL, T. M. **Processamento e Visualização de Campos em Ambientes Virtuais e Sistemas CAD 3D Aplicados a Projetos de Iluminação em Subestações.** Dissertação de Mestrado em Métodos Numéricos em Engenharia, Universidade Federal do Paraná, Curitiba, Brasil, 2006.

BURIOL, T. M.; SILVA NETO, M. A.; SCHEER, S.; GODOI, W. C. **Modelagem em VRML para visualização científica utilizando o Visualization Toolkit**. In: IX Symposium on Virtual and

REFERÊNCIAS

Augmented Reality, 2007, Petrópolis. Proceedings SVR2007. Petrópolis : LNCC e SBC, 2007. v. 1. p. 1-3.

BUZZI, M. F. **Avaliação das Correlações de Séries Temporais de Leituras de Instrumentos de Monitoração Geotécnico-Estrutural e Variáveis Ambientais em Barragens - Estudo de Caso de ITAIPU**. Dissertação de Mestrado do Programa de Pós Graduação em Métodos Numéricos em Engenharias (PPGMNE), UFPR, Curitiba, Brasil, 2007.

CALVETTI, L.; BENETI, C.; PEREIRA FILHO, A. J. **Integração do radar meteorológico dopller do Simepar e uma rede pluviométrica para a estimativa da precipitação.** In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 11. 2003, Belo Horizonte. CD-ROM.

CARD, S.K.; MACKINLAY, J.D.; SHNEIDERMAN, B. **Information Visualization**. In: Readings in Information Visualization - Using Visualization to Think. San Francisco, Morgan Kaufmann Publ., 1999. Card, S.K.; Mackinlay, J.D.; Shneiderman, B. (eds.). p. 1-34.

CARVALHO, J. G. **Coordenadas Paralelas: Uma Metodologia para Visualização em 3D**. Dissertação de Mestrado, Programa de Pós Graduação em Ciência da Computação, Pontifícia Universidade Católica do Rio Grande do Sul (PUC-RS), Porto Alegre, Brasil, 2001.

CHERNOFF, H. **The use of Faces to Represent Points in K-Dimensional Space Graphically**. Journal of American Statistical Association, vol. 68, p. 361-368, 1973.

CHOU, S. Y.; LIN, S. W.; YEH, C. S. **Cluster Identification with Parallel Coordinates**, Patterns Recognition Letters, vol. 20, p. 565-572, 1999.

CHUAH, M. C.; ROTH, S. F.; MATTIS, J.; KOLOJEJCHICK, J. **SDM: Selective dynamic manipulation of visualizations**. In Proceedings of the ACM Symposium on User Interface Software and Technology, 3D User Interfaces, pages 61–70. 1995.

CLEVELAND, W. S. **Visualizing Data.** Hobart Press, Summit, 1993.

COELHO, C. J. **Seminário de Análise Multivariada.** Disponível em:<http://agata.ucg.br/formularios/NPI/clarimar/>. Acesso em: 17/12/2007.

COHEN, M.; MANSSOUR, I. H. **OpenGL – Uma Abordagem Prática e Objetiva**. Ed. Novatec, São Paulo, SP, Brasil, 2006.

REFERÊNCIAS

DAVIS, J.C. **Statistics and Data Analysis in Geology** 2th ed., John Wiley and Sons, Inc.1986.

DYAS, N.; RAGAN, S. **Clusteing Algorithm. The Maryland Virtual High School of Science And Mathematics**, 1995. Disponível em: <http://mvhs1.mbhs.edu/mvhsproj/clustering/cluster.html>. Acesso em: 17/12/2007.

ESTER, M.; KRIEGEL, H-P; SANDER, J.; XU, X. **A density-based Algorithm for Discovering clusters in Large Spatial Databases with Noise**. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, KDD 1996.

EVERITT, B. **Cluster Analyses.** 2a. Ed., Gower Publishing Co., 1980.

FALOUTSOS, C.; LIN, K., **FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets.** In ACM SIGMOD, Zurich, Suiça, 1995, p. 163-174.

FANEA, E.; CARPENDALE, S.; ISENBERG, T. **An Interactive 3D Integration of Parallel Coordinates and Star Glyphs**. In Proceedings of the IEEE Symposium on Information Visualization (InfoVis 2005, October 23--25, 2005, Minneapolis, Minnesota, USA). Los Alamitos, CA. (John Stasko and Matt Ward, Eds.) IEEE Computer Society, p. 149-156, 2005.

FAYYAD, U.; SHAPIRO, P. G.; SMYTH, P. **From Data Mining to Knowledge Discovery: An Overview**. In Fayyad, U., Piatetsky-Shapiro, G., Amith, Smyth, P., and Uthurusamy, R. (eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, p. 1-36, Cambridge, 1996.

FOLEY, J., RIBARSKY, B. **Next-generation Data Visualization Tools, in Scientific Visualization.** Advances and Chalenges, Ed: L. Rosenblum, R.A. Earnshaw, J.Encarnacao, H. Hagen, A. Kaufman, S. Klimenko, G. Nielson, F. Post, D. Thalmann, Academic Press. 1994.

FREITAS, C. M. D. S.; CHUBACHI, O. M.; LUZZARDI, P. R. G.; CAVA, R. A. **Introdução à Visualização de Informações.** Revista de Informática Teórica e Aplicada. Porto Alegre, RS, v. 8, n. 2, p. 143-158, 2001.

FREITAS, C.M.D.S.; WAGNER, F.R. **Ferramentas de Suporte às Tarefas da Análise Exploratória Visual**. Revista de Informática Teórica e Aplicada, v.2, n.1, p.5-36, jan. 1995.

REFERÊNCIAS

FRIENDLY, M. **A Brief History of Data Visualization.** Handbook of Computational Statistics: Data Visualization, Vol III, Springer, 2007. Disponível em: <http://www.math.yorku.ca/SCS/Gallery/milestone/>. Acesso em: 11/12/2007.

FRIENDLY, M. **Milestones in the history of data visualization: A case study in statistical historiography**. In C. Weihs and W. Gaul, eds., Classification: The Ubiquitous Challenge, 2005, (p. 34-52). New York: Springer. 25

FRIENDLY, M. **Visions and re-visions of Charles Joseph Minard**. Journal of Educational and Behavioral Statistics, 2002, 27(1), 31-52. 14

FUNKHOUSER, H. G. (1936). **A note on a tenth century graph**. Osiris, 1, 260-262. 2, 3, 4

FUNKHOUSER, H. G. (1937). **Historical development of the graphical representation of statistical data**. Osiris, 3(1), 269-405. Reprinted Brugge, Belgium: St. Catherine Press, 1937. 2, 12, 14, 18, 26

FURNAS, G. W. **Generalized fisheye views**. In Proceedings of ACM CHI'86 Conference on Human Factors in Computing Systems, Visualizing Complex Information Spaces, pages 16–23. 1986.

FURNAS, G.; JUL, S. **Navigation in electronic worlds**. In Proceedings of ACM CHI 97. Conference on Human Factors in Computing Systems, volume 2 of Workshop 9, page 230. 1997.

GANESH, M.; HAN, E-H.; KUMAR, V.; SHEKHAR, S.; SRIVASTAVA, J. ? **Visual Data Mining: Framework and Algorithm Development.** Technical Report TR-96-021, Department of Computer Science, University of Minnesota, Minneapolis, 1996.

GERSHON, N. **From Perception to Visualization, in Scientific Visualization**. Advances and Chalenges, Ed: L. Rosenblum, R.A. Earnshaw, J. Encarnacao, H. Hagen, A. Kaufman, S. Klimenko, G. Nielson, F. Post, D. Thalmann, Academic Press. 1994.

GILBERT, E. W. **Pioneer maps of health and disease in England**. Geographical Journal, 1958, 124, 172-183. 11, 12

GIMENES, E. **"Data Mining – Data Warehouse" A Imortância da Mineração de Dados em Tomadas de Decisões**. Centro Estadual de Educação Tecnológica Paula Souza. Faculdade de

REFERÊNCIAS

Tecnologia de Taguaritinga. Monografia de conclusão para Tecnólogo em Processamento de Dados. Taquaritinga, 2000.

GIMENES, E. **A importância da Mineração de Dados em Tomadas de Decisão**. Disponível em: <http://br.geocities.com/dugimenes/>. Acesso em: 13/12/2007.

GORDON, A. D. **Classification.** Chapman and Hall, 1981.

GORNI, A. A. **Redes Neurais Artificiais – Uma abordagem revolucionária em Inteligência Artificial**. Micro Sistemas, São Paulo, 1993.

GREIGH-SMITH, P. **Quantitative Plant Ecology.** University of California Press, Berkeley, 1983.
GUHA, S.; RASTOGI, R.; SHIM, K. **CURE: An Eficient Clustering Algorithm for Large databases**. ACM/SIGMOD 1998.

HALLEY, E. (1686). **On the height of the mercury in the barometer at different elevations above the surface of the earth, and on the rising and falling of the mercury on the change of weather**. Philosophical Transactions, (p. 104-115). 6

HALLEY, E. (1701). **The description and uses of a new, and correct sea-chart of the whole world, shewing variations of the compass**. London. 7, 14

HALLEY, E. **From Wikipedia, the free encyclopedia**. Disponível em: <http://pt.wikipedia.org/wiki/Edmond_Halley>. Acesso em: 11/12/2007.

HAYKIN, S. **Neural Networks – A Comprehensive Foundation.** Macmillian College Publishing, inc., 1994.

HEICHT-NIELSEN, R. **Neurocomputing**. Addison-Wesley Publishing Company, New York, 1991.

HINNEBURG, A; KEIM, D.A; WAWRYNIUK, M. **HD-Eye: Visual Mining of High-Dimensional Data.** IEEE Computer Graphics and Applications, v.19, n.5, p.22-31, set./out. 1999.

HOFFMAN, P. E. Table Visualization: A formal Model and Its applications. Doctoral Diss, Computer Science Department, University Of Massachusetts, Lowell, Ma,1999.

REFERÊNCIAS

HOFFMAN, P.; GRINSTEIN, G. **A Survey of Visualizations for High-Dimensional Data Mining**. In: FAYYAD, U.; GRINSTEIN, G.G.; WIERSE, A. **A Information Visualization in Data Mining and Knowledge Discovery**. San Francisco, Morgan Kaufmann Publishers, 1999. p.47-82.

HÖRNE, K. H.; *et al*. **VOXELman – Simulação Visual de Corpos Humanos.** Universidade de Hamburgo. Disponível em: <http://www.voxel-man.de>. Acesso em: 10/12/2007.

INSELBERG, A. **Don´t Panic ... just do it in Parallel!.** J. of Comp.Stat.14:53 - 77, 1999.

INSELBERG, A.; AVIDAN, T. **The automated multidimensional detective**. InfoVis Conf. .99. Proc. 112-119. IEEE Computer Society.

ITAIPU. **ITAIPU Binacional**. Disponível em: <http://www.itaipu.gov.br>. Acesso em: 07/01/2008.

JEONG, C.; PANG, A. **Reconfigurable disc trees for visualizing large hierarchical information space**. Proceedings of IEEE Information Visualization, Raleigh Durham, North Carolina, Outubro 1998. p. 19-25.

JOHNSON, B.; SHNEIDERMAN, B. **TreeMaps: A space - filling approach to the visualization of hierarchical information structures**. Proceedings of IEEE Visualization. San Diego. 1991. p 284 – 291

JOHNSON, R.A & WICHERN, D.W. **Applied Multivariate Statistical analysis**. 2a ed. New Jersey: Prentice Hall, Inc., 1998.

JOHNSON, S.; EDWARDS, J. **Vis5D+ Project.** Disponível em: <http://www.ssec.wisc.edu/~billh/vis5d.html>. Acesso em: 10/12/2007.

JOHNSTON, W.M.; HANNA, J.R.P.; MILLAR, R.J. ***Advances in dataflow programming languages***. ACM Computing Surveys (CSUR) 36 (1): 1-34. Retrieved on 2007, 03-31.

KACHIGAN, S. K. **Statistical Analysis An Interdisciplinary Introduction to Univariate & Multivariate Methods**. Radius Press, New York, NY. 1986.

KANDOGAN, E. **Visualizing Multi-Dimensional Clusters**, Trends, and Outliers using Star Coordinates, Proc. ACM Int. Conf. Knowledge Discovery and Data Mining, p.107-116, 2001.

REFERÊNCIAS

KEIM, D. A. **Designing Pixel-Oriented Visualization Techniques:Theory And Applications**. IEEE Transactions on Visualization and Computers Graphics, vol. 6, n.1, p.59-78, 2000.

KEIM, D. A. **Information Visualization and Visual Data Mining.** IEEE Transactions on Visualization And Computers Graphics, vol. 8, n.1, p. 1-8, 2002.

KEIM, D. A. **Visual Exploration of Large Data Sets**. Communications of the ACM, v.44, n.8, p.38-44, agosto. 1979.

KEIM, D. A., KRIEGEL, H. P. **VisDB: Database Exploration using Multidimensional Visualization.** IEEE Computer Graphics and Applications, vol. 14, n. 5, setembro, p. 40-49,1994.

KEIM, D. A., KRIEGEL, H. P. **Visualization Techniques for Mining Large Databases: A Comparison**. IEEE Trans. Knowledge & Data Engineering, vol. 8, n. 6, p. 923-936, 1996.

KNORR, E.M.; NG, R.T. **Algorithms for Mining Distance-Based Outliers in Large Datasets**. Proceedings of the 24th International Conference on Very Large Data Bases, VLDB 1998.

KRÖSE, B. J. A.; VAN DER SMAGT, P. P. **An Introduction to Neural Networks.** Amsterdam, University of Amsterdam, 1993.

LAMPING J.; RAO, R.; PIROLLI, P. **The hyperbolic browser: a focus+context technique for visualizing large hierarchies.** Journal of Visual Languages and Computing,7(1):33-55, março 1996.

LANDIM, P. M. B. **Análise estatística multivariada de dados geológicos.** Disponível em: <http://www.rc.unesp.br/igce/geologia/GAA02144/aulas.html> Acesso em: 17/12/2007.

LEBLANC, J.; WARD, M. O.; WITTELS, N. **Exploring N-Dimensional Databases**. Proc. IEEE Visualization'90, IEEE CS Press, p. 230-237, 1990.

LOURDES, H. L. **Data Mining – Teoria e Prática**. Instituto de Informática, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brasil. Disponível em: <www.inf.ufrgs.br/~clesio/cmp151/cmp15120021/artigo_lourdes.pdf>. Acesso em: 13/12/2007.

LU, H.; SETIONO, R.; LIU, H. **Neurorule: A connectionist approach to data mining**. In Proc. 1995 Int. Conf. Very Large Data Bases (VLDB'95), 478-489, Zurich, Switzerland, Sept. 1995.

REFERÊNCIAS

LUO, X. -L.; *et al.* **Mathematical and Information Sciences - Computational Modelling.** Disponível em: <http://www.cmis.csiro.au/cfd/index.htm>. Acesso em: 10/12/2007.

LUO, X. -L.; STOKES, A. N.; BARTON, N.G. **Turbulent flow around a car body - Report of Fastflo solutions**, WUA-CFD Freiburg (1996).

MACKINLAY, J. D.; ROBERTSON, G. G.; CARD, S. K. **The perspective wall: Detail and context smoothly integrated**. In Proceedings of ACM CHI?91 Conference on Human Factors in Computing Systems, Information Visualization, p. 173-179. 1991.

MÁSSON, E. e WANG, Y. **Introduction to computation and learning in artificial neural networks.** European Journal of Operational Research. v. 47, p. 1-28, 1990.

MCCORMICK, B. H.; DEFANTI, T. A.; BROWN, M. D. **Visualization in Scientific Computing**. Computer Graphics (special issue), vol. 21, no. 6, Nov. 1987.

MCULLOCH, W. S.; PITTS, W. **A Logical Cauculos of the Ideas Immanent in Nervus Activity**. Bulletin of Mathematical Biophysics, vol 5, p. 115-133 - 1943;

MEHTA, M.; AGRAWAL, R.; RISSANEN, J. **SLIQ: A Fast Scalable Classifier for Data Mining.** Proc. of the Fifth Int'l Conference on Extending Database Technology, Avignon, France, março 1996.

MINARD, C. J. **Infografia**. From Wikipedia, the free encyclopedia. Disponível em: <http://pt.wikipedia.org/wiki/Infogr%C3%A1fico> Acesso em: 11/12/2007.

MINGHIM, R., LEVKOWITZ, H. **Laboratório de Computação de Alto Desempenho - Visualização Computacional**. USP, 2006. Disponível em: <http://www.lcad.icmc.usp.br/~rosane/Vis.html>. Acesso em: 27/11/2006.

MOITA NETO, J. M. **Estatística Multivariada.** Disponível em: <http://criticanarede.com/cien_estatistica.html>. Acesso em: 17/12/2007.

MUKHERJEA, S.; FOLEY, J. D.; HUDSON, S. **Visualizing complex hypermedia networks through multiple hierarchical views.** In Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, volume 1 of Papers: Creating Visualizations, pages 331–337. 1995.

REFERÊNCIAS

NG, R.T.; HAN, J. **Eficient and Effective Clustering Methods for Spatial Data Mining**. In Proceedings of the 1994 International Conference Very Large Data Bases, Santiago, Chile, p. 144–155. Morgan Kaufmann, San Francisco, CA, setembro, 1994

PALSKY, G. **Des Chiffres et des Cartes: Naissance et développement de la Cartographie Quantitative Francais au XIX^{e} siècle**. Paris: Comité des Travaux Historiques et Scientifiques (CTHS). 1996; 2, 7, 10, 14, 15, 18

PARSAYE, K.; CHIGNELL, M. **Intelligent Database Tools and Applications**. John Wiley & Sons, 1993.

PEARSON, K., **On Lines and Planes of Closest Fit to System of Points in Space**, Philosophy Magazine, vol. 6, p. 559-572, 1901.

PICKETT, R.M.; GRINSTEIN, G.G. **Iconographic Displays for Visualizing Multidimensional Data**. Proceedings of IEEE Conference on Systems, Man and Cybernetics'88, Piscataway, NJ, 1988, p.361-370.

PIELOU, E. C. **The Interpretation of Ecological Data.** Wiley-Interscience, 1984.

PINHEIRO, L. C.; VAZ, M. S. M. G.; MARTINHAGO, A. Z. **Proposta de uma Extenção do Padrão FGDC/CSGDM para Dados de Radar Meteorológicos.** Revista Publicatio UEPG, ciências exatas e da terra, ciências agrárias e engenharias, ed. 03-2005, ano 11.

PLAYFAIR, W. H. **William Playfair.** From Wikipedia, the free encyclopedia. Disponível em: <http://en.wikipedia.org/wiki/William_Playfair>. Acesso em: 10/12/2007.

PRENTICE, I. C. **Multidimensional scaling as a research tool in Quarternary palybology: A review of theory and methods.** Review of Paleobotany & Palynology, 71 – 104, 1980.

RENDERIZAÇÃO. Wikipédia – a enciclopédia livre. Disponível em: <http://pt.wikipedia.org/wiki/Renderiza%C3%A7%C3%A3o> Acesso em 11/02/2008.

REZENDE, S.O.; PUGLIESI, J.B.; MELANDA, E.A.; DE PAULA, M.F. **Mineração de Dados**. In S. O. Rezende (Ed.), Sistemas Inteligentes – Fundamentos e Aplicac¸ ˜oes, p. 307–335. Editora Manole, 2003.

REZENDE, S.O. Sistemas Inteligentes: fundamentos e aplicações. Barueri, SP. Manole, 2003.

REFERÊNCIAS

RHYNE, T. M. **Does the Difference between Information and Scientific Visualization Really Matter??**, IEEE Computer Graphics and Applications, maio/junho, 2003, p. 6-8.

RIBARSKY, W.; KATZ, J.; JIANG, F.; HOLLAND, A. **Discovery Visualization Using Fast Clustering.** IEEE Computer Graphics and Applications, v.19, n.5, p.32-39, setembro/outubro. 1999.

RICH, E.; KNIGHT, K. **Inteligência Artificial**. Makron Books. 2ª. Edição. São Paulo, 1994.mins 722p.

ROBERTSON, G. G.; MACKINLAY, J. D.; CARD, S. K. **Cone trees: Animated 3D visualizations of hierarchical information**. In Robertson, S. P., Olson, G. M., and Olson, J. S., editors, Proc. ACM Conf. Human Factors in Computing Systems*,* CHI, pages 189–194. ACM Press. 1991.

ROHRER, R.M.; SIBERT, J.L.; EBERT, D.S. **A Shape-based Visual Interface for Text Retrieval**. IEEE Computer Graphics and Applications, v.19, n.5, p.40-46,setembrp/outubro. 1999.

SANCHEZ, P. F. **Análise e Previsão de Séries Temporais de Alguns Instrumentos de Auscultação da Barragem de ITAIPU**. In: 14°Evento de Iniciação Científica – EVINCI – UFPR, 2006, Curitiba. Anais 14°EVINCI. Ed. UFPR, 2006.

SANTOS, B. S. **Introdução à Visualização de Dados e Informação**. Disponível em: <http://www.ieeta.pt/~bss/disciplinas/ADVI/ADVI.htm>. Acesso em: 11/12/2007.

SANTOS, C. R.; GROS, P.; ABEL, P. **Visualização Tridimensional de Grandes Volumes de Informações.** CLME'99, Congresso Luso-Moçambicano de Engenharia, 14-16, 1999, Maputo, Mozambique, Proceedings Volume 2.

SARKAR, M.; BROWN, M. H. **Graphical fisheye views of graphs**. In Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems, Visualizing Objects, Graphs, and Video, p. 83–91. 1992.

SARKAR, M.; SNIBBE, S.; TVERSKY, O. J.; REISS, S. P. **Stretching the rubber sheet: A metaphor for viewing large layouts on small screens**. Technical Report CS-93-39, Department of Computer Science, Brown University. 1993.

SHNEIDERMA, B.; *et. al*. **Treemap. Human-Computer Interection Lab. University of Maryland**. Disponível em: <http://www.cs.umd.edu/hcil/treemap/>. Acesso em: 27/12/2007.

REFERÊNCIAS

SHNEIDERMAN, B. **The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations**. Proceedings of IEEE Symposium on Visual Languages, Boulder, CO, 1996. p. 336- 343.

SILVA NETO, M. A.; BURIOL, T. M.; SCHEER, S. **Um visualizador interativo para exploração de dados volumétricos obtidos em análise pelo método dos elementos finitos**. In: XXVIII Congresso Ibero Latino Americano de Métodos Computacionais em Engenharia, 2007, Porto. CMNE / CILAMCE 2007. Porto : Faculdade de Engenharia - Universidade do Porto, 2007. v. 1. p. 1-17.

SILVER, D.L. **Knowledge Discovery and Data Mining**. Technical Report MBA6522 CogNova Technologies London Health Science Center, 1996.

SIMEPAR. **Institudo Tecnológico SIMEPAR**. Disponível em: <www.simepar.br>. Acesso em: 08/01/2008.

SNEATH, D. H.; SOKAL, R. R. **Numerical Taxonomy.** W. H. Freeman & Co., 1973.

SOBOL, M. G.; KLEIN, G. **New Graphics As Computerized Displays For Human Information Processing**. IEEE Transactions on Systems, Man, and Cybernetics, vol. 19, n. 4, 1989.

SOUKUP, T.; DAVIDSON, I., **Visual Datamining - Techniques and Tools for Data Visualization and Mining**, Ed. Wiley Publishing, Inc. 2002.

SPENCE, R.; APPERLEY, M.D. **Data Base Navigation: An Office Environment for the Professional**. Behaviour and Information Technology,1(1): 43-54, 1982.

STASKO, J.; ZHANG, E. **Focus+Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visualizations**. Proceedings of IEEE Information Visualization, San Francisco, California, October 2000. p. 57-65.

STEINER, M.; SMITH, J. **Use of three-dimensional reflectivity structure for automated detection and removal of non-precipitating echoes in radar data.** J. Atmos. Ocea. Tech., 2002, 19, 673-686.

TOLEDO, G. L.; OVALLE, I. I. **Estatística Básica**. Editora Atlas, São Paulo, SP. 1995.

REFERÊNCIAS

TRIOLA, M. F. **Introdução à Estatística**. LTC Livros Técnicos e Científicos Editora S.A., Rio de Janeiro, RJ, 1999.

TUFTE, E.R. – **Envisioning Information.** Graphics Press, USA, 1990

TUFTE, E.R. **The Visual Display of Quantitative Information.** Graphics Press, USA, 1983

VAN WIJK, J.J.; VAN DE WETERING, H. **Cushion Treemaps: Visualization of Hierarchical Information.** Proceedings of IEEE Information Visualization, Outubro 1999. p. 73-78.

WALKER, G. **Challenges of information visualization**. British Telecommunications Engineering

WALTON, J. **Get the picture: a new direction in data visualization**. In: *Earnshaw, R. A.; Watson, D. (Eds.) Animation and Scientific Visualization: tools & applications.* Academic Press, 1993, P. 29-36.

WARD, M. O.; RUNDENSTEINER, E. A.; CUI, Q.; XIE, Z.; YANG, D.; WAD, C.; NGUYEN, D. Q. **Xmdv Tool Release – The Multivariate Data Visualization Tool.** Disponível em: <http://davis.wpi.edu/~xmdv/>. Acesso em: 26/12/2007.

WARD, M.O. **XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data.** Proceedings IEEE Visualization '94, Washington, DC, 1994, p.326-33.

WONG, P. C.; BERGERON, R. D. **30 Years of Multidimensional MultivariateVisualization**. In NIELSON, G. M., HAGEN, H., MÜLLER, H. Scientific visualization: overviews, methodologies, and techniques, Los Alamitos, California, 1997, 400p.

WONG, P.C. **Visual Data Mining.** IEEE Computer Graphics and Applications, v.19, n.5, p.20-21, set./out. 1999.