SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito: 19.05.2004

Assinatura:_____

Visualizações Temporais em uma Plataforma de Software Extensível e Adaptável

Milton Hirokazu Shimabukuro

Orientadora: *Profa. Dra. Maria Cristina Ferreira de Oliveira*

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional.

**USP – São Carlos
Maio de 2004**

Dedico este trabalho
Aos meus filhos Victor Eiji e Nicole Meyi
À minha esposa Cristina Satiko
Às minhas irmãs Edina Harumi e Adriana Ruriko
Aos meus pais Hiromoti e Mineko
À minha avó Matsu
À memória de meu irmão Gilson Hideki
À memória de meu avô Hosei
pelos Valores que todos representam

AGRADECIMENTOS

À professora Maria Cristina Ferreira de Oliveira, orientadora, pelo incentivo, apoio, compreensão, parceria e direção na realização deste projeto.

Às professoras Maria Carolina Monard e Agma Juci Machado Traina, do ICMC-USP-São Carlos, e ao professor Marcelo Knörich Zuffo, da Escola Politécnica-USP-São Paulo, membros da Comissão Examinadora do Exame de Qualificação, pela contribuição.

À professora Rosane Minghim e ao professor Luis Gustavo Nonato, pela amizade e valorização.

Aos amigos Silvio Tadao Fujisaki e Rogério Eduardo Garcia, com os quais compartilhei alegrias e superei adversidades durante esta caminhada, pelo companheirismo.

Aos colegas Ricardo Barbosa e Almir Artero, pela cooperação.

Aos professores Neri Alves e Messias Menguette Jr., da FCT-Unesp-Presidente Prudente, que muito colaboraram para a concretização desta jornada, pela oportunidade.

Aos amigos professores da FCT-Unesp-Presidente Prudente, pelo incentivo, valorização e colaboração.

Aos funcionários do ICMC-USP-São Carlos e da FCT-Unesp-Presidente Prudente, sempre dispostos na prestação de seus serviços, pela colaboração.

Ao amigo professor Edílson Ferreira Flores, pela disposição em sempre colaborar e paciência.

Aos senhores Domingos Hiroto e Antonio Augusto Vicente, da indústria Bebidas Wilson de Presidente Prudente, pela oportunidade e colaboração.

Aos amigos Paulo Justiniano e Silvia Emiko, pela cooperação, apoio e incentivo.

À minha esposa Cristina e aos meus filhos Eiji e Meyi, pela cumplicidade, paciência, incentivo e valorização.

À minha família, pelo incondicional apoio que suporta minhas caminhadas.

À USP e à Unesp, pela oportunidade.

ÍNDICE

1. INTRODUÇÃO	1
2. VISUALIZAÇÃO INTERATIVA E DESCOBERTA DE CONHECIMENTO	5
2.1 Descoberta de Conhecimento e Mineração de Dados	7
2.2 Visualização e Descoberta de Conhecimento	11
2.2.1 Técnicas Orientadas a Pixel	15
2.2.2 Técnicas de Projeção Geométrica	18
2.2.3 Técnicas Iconográficas	20
2.2.4 Técnicas Hierárquicas	22
2.2.5 Visualização como Apoio à Mineração de Dados	23
2.2.6 Integração de dois Aplicativos Distintos	27
2.2.7 Múltiplas Visualizações Integradas e Coordenadas	28
2.3 Soluções com Coordenação Flexível entre Visualizações	30
2.3.1 Snap-Together Visualization	30
2.3.2 Projeto SPIN!	37
2.3.3 GeoVISTA Studio	40
2.3.4 Considerações Gerais sobre as Soluções	42
2.4 Considerações Finais	43
3. DEFINIÇÃO DE REPRESENTAÇÕES VISUAIS TEMPORAIS E CONCEPÇÃO DO MODELO <i>AdaptaVis</i>	45
3.1 Representações Visuais para Atributos Temporais	47
3.1.1 Variação Temporal Uni-escala	49
3.1.2 Variação Temporal Multi-escala	50
3.2 O Modelo <i>AdaptaVis</i>	53
3.3 Considerações Finais	59

4. PLATAFORMA <i>InfoVis</i> : UMA INSTANCIACÃO DOS NÚCLEOS BÁSICO E FUNCIONAL DO <i>AdaptaVis</i>	60
4.1 Estrutura Geral	61
4.2 Núcleo Básico	62
4.3 Núcleo Funcional	66
4.3.1 Técnicas de Visualização Implementadas	73
4.4 Ferramenta de Desenvolvimento	77
4.5 Considerações Finais	79
5. VALIDAÇÃO DO MODELO E DAS REPRESENTAÇÕES VISUAIS TEMPORAIS ...	81
5.1 Exploração de Dados de Precipitação Pluviométrica	82
5.1.1. Definição de um Conjunto de Postos para Processamento	87
5.1.2. Preenchimento de Falhas	94
5.1.3. Determinação de Anos Representativos	99
5.2. Exploração de Dados de Transações Comerciais	103
5.3. Considerações Finais	107
6. CONCLUSÕES	110
REFERÊNCIAS BIBLIOGRÁFICAS	117
APÊNDICE A	A.1
APÊNDICE B	B.1

LISTA DE FIGURAS

2.1. Etapas do processo de KDD	6
2.2. Etapas agrupadas do processo de KDD	9
2.3. Múltiplas janelas para um caso de 6 atributos em técnicas orientadas a pixel	15
2.4. Possíveis esquemas para arranjo de pixels em técnicas orientadas a pixel	17
2.5. Formato circular para apresentação dos pixels em Segmentos de Círculo	18
2.6. Esquema de arranjo dos pixels em Gráficos de Barras de Pixels	18
2.7. Correlação e distribuição de valores em Coordenadas Paralelas	19
2.8. Sobreposição de linhas com apresentação de 473 itens	19
2.9. Correlação linear em uma matriz de Gráficos de Dispersão	20
2.10. Configurações para Figura de Arestas usando cinco arestas	21
2.11. Diferentes texturas representadas por Figura de Arestas	21
2.12. Resultado do empilhamento dimensional. Quatro atributos são aninhados aos pares	23
2.13. Passos da Classificação Visual	25
2.14. Interface da ferramenta de Classificação Visual	26
2.15. Um estágio da árvore de decisão gerada pela ferramenta de Classificação Visual.	26
2.16. Uso conjunto de XGobi e ArcView	27
2.17. Formas de representação visual na integração GVis/KDD	29
2.18. Coordenação no <i>Snap</i> pelo esquema relacional	31
2.19. Passos no <i>Snap</i> para coordenação de visualizações	32
2.20. Funcionamento de uma coordenação no <i>Snap</i>	33
2.21. Interface web de acesso ao <i>Snap Visualization Server</i>	35
2.22. Camadas da arquitetura do <i>Snap Visualization Server</i>	36
2.23. Funcionamento do Descartes	38
2.24. Elementos da interface do SPIN!	38
2.25. Visualizações resultantes de uma configuração de espaço de trabalho no SPIN! .	39
2.26. Elementos da arquitetura do SPIN!	39
2.27. Arquitetura e programação visual no <i>Studio</i>	41

2.28. Organização de componentes no painel principal do <i>Studio</i>	41
2.29. Criação visual de um projeto no <i>Studio</i>	42
3.1. Objeto espacial com várias medidas de um atributo ao longo do tempo	48
3.2. Construção da representação visual ‘Variação Temporal Uni-escala’	49
3.3. Elementos de construção de uma representação visual ‘Variação Temporal Multi-escala’	51
3.4. Divisão do retângulo para representação da escala diária	52
3.5. Opções de organização da representação visual ‘Variação Temporal Multi-escala’	52
3.6. Módulos que integram o modelo <i>AdaptaVis</i>	56
3.7. Composição dos núcleos Básico e Funcional no <i>AdaptaVis</i>	58
4.1. Configurações local e remota de plataforma/ambiente no <i>InfoVis</i>	61
4.2. Estrutura de diretórios da plataforma <i>InfoVis</i>	62
4.3. Passos para a ativação da interface principal da plataforma <i>InfoVis</i>	63
4.4. Formato do arquivo de configuração <i>menu.cfg</i> e o resultado no <i>InfoVis</i>	64
4.5. Interface Principal do <i>InfoVis</i> e as áreas dos Gerenciadores	65
4.6. Registro de ações coordenadas e histórico	66
4.7. Configuração da máquina virtual <i>HotSpot</i> que permite a escolha de um perfil para melhor adequação à aplicação	79
5.1. Mapa do aplicativo BcDAEE impresso, com delimitação manual de região de interesse, oeste do Estado de SP	86
5.2. Modelo de tabela sintética sobre quantidade e qualidade de dados	86
5.3. Configuração da plataforma <i>InfoVis</i> para as sessões de uso com dados de chuva .	87
5.4. Ambiente configurado para seleção visual de postos	89
5.5. Interface da plataforma <i>InfoVis</i> com ambiente para seleção visual de postos	89
5.6. Componentes funcionais para a seleção visual de postos	89
5.7. Influência da escala de cores na percepção da informação	91
5.8. Componentes para suporte ao passo de descarte de postos	93
5.9. Componente funcional ‘Tabela de Dados’ após acréscimo de funcionalidades	94
5.10. Avaliação da chuva em postos para preenchimento de falha	96
5.11. Disposição alternativa para avaliação do comportamento de chuva	97
5.12. Comparação do comportamento global de chuva dos postos	99

5.13. ‘Distribuição Espacial’ e ‘Intervalo/Categoria Generalizado’ na escolha de anos representativos	101
5.14. Ganho de detalhe controlado pela escala de cor	101
5.15. Múltiplas visualizações como suporte ao processo cognitivo	102
5.16. Variação diária de transações de produtos	104
5.17. Transações de produtos em diversas cidades em um determinado ano	106
 6.1. Enfoque inter e multidisciplinar do modelo <i>AdaptaVis</i>	 114
 A.1. Cartograma com uma classificação climática	 A.3
A.2. Árvore de ligação com agrupamentos para anos representativos	A.4
A.3. Gráfico de Análise Rítmica	A.5
A.4. Carta sinótica	A.5
A.5. Tela inicial do aplicativo BcDAEE	A.6
A.6. <i>Zoom</i> em torno de uma região no aplicativo BcDAEE	A.7
A.7. Dados do posto D8-003 no aplicativo BcDAEE	A.7
A.8. Exemplo de pluviograma no aplicativo BcDAEE	A.8
 B.1. Elementos da interface para os estudos iniciais	 B.2
B.2. Visualização da opção “Anos início-fim”	B.3
B.3. Totais mensais a partir de 1930	B.3
B.4. Totais mensais a partir de 1968	B.4
B.5. Visualização de dados de um único posto em escala diária	B.4
B.6. Múltiplas visualizações em escala diária	B.5

LISTA DE TABELAS

3.1. Descrição funcional do protocolo de serviços para a comunicação entre os componentes	58
4.1. Métodos a serem implementados por um componente funcional	66
4.2. Mensagens trocadas entre os componentes em diversas atividades na plataforma <i>InfoVis</i>	69
4.3. Características dos componentes funcionais implementados	73
4.4. Tempos, em segundos, para a execução de tarefas por códigos gerados por um compilador C++, Java com compilador JIT e Java com interpretação de <i>bytecodes</i>	78

RESUMO

Repositórios com volumes de dados cada vez maiores foram viabilizados pelo desenvolvimento tecnológico, criando importantes fontes de informação em diversas áreas da atividade humana. Esses repositórios freqüentemente incluem informação sobre o comportamento temporal e o posicionamento espacial dos itens neles representados, os quais são extremamente relevantes para a análise dos dados. O processo de descoberta de conhecimento a partir de grandes volumes de dados tem sido objeto de estudo em diversas disciplinas, dentre elas a Visualização de Informação, cujas técnicas podem apoiar diversas etapas desse processo. Esta tese versa sobre o uso da Visualização Exploratória em conjuntos de dados com atributos temporais e espaciais, empregando a estratégia de múltiplas visualizações coordenadas para apoiar o tratamento de dados em estágios iniciais de processos de descoberta de conhecimento. São propostas duas novas representações visuais temporais – denominadas ‘Variação Temporal Uni-escala’ e ‘Variação Temporal Multi-escala’ – para apoiar a análise exploratória de dados temporais. Adicionalmente, é proposto um modelo de arquitetura de software – *AdaptaVis*, que permite a integração dessas e outras representações visuais em uma plataforma de visualização de informação flexível, extensível e adaptável às necessidades de diferentes usuários, tarefas e domínios de aplicação – a plataforma *InfoVis*. Sessões de uso realizadas com dados e usuários reais dos domínios de Climatologia e Negócios permitiram validar empiricamente as representações visuais e o modelo. O modelo *AdaptaVis* e a plataforma *InfoVis* estabelecem bases para a continuidade de diversas pesquisas em Visualização de Informação, particularmente o estudo de aspectos relacionados ao uso coordenado de múltiplas visualizações, à modelagem do processo de coordenação, e à integração entre múltiplas técnicas visuais e analíticas.

ABSTRACT

Data repositories with ever increasing volumes have been made possible by the evolution in data collection technologies, creating important sources of information in several fields of human activity. Such data repositories often include information about both the temporal behavior and the spatial positioning of data items that will be relevant in future data analysis tasks. The process of discovering knowledge embedded in great volumes of data is a topic of study in several disciplines, including Information Visualization, which offers a range of techniques to support different stages of a discovery process. This thesis addresses the application of Exploratory Visualization techniques on datasets with temporal and spatial attributes, using the strategy of coordinating multiple data views, to assist data treatment on early stages of knowledge discovery processes. Two temporal visual representations are proposed – ‘Uni-scale Temporal Behavior’ and ‘Multi-scale Temporal Behavior’ – that support the exploratory analysis of temporal data. Moreover, a software architecture model is introduced – *AdaptaVis*, that allows the integration of these and other visualization techniques into a flexible, extensible and adaptable information visualization platform – called *InfoVis* – that may be tailored to meet the requirements of different users, tasks and application domains. Sessions conducted with real data and users from the Climatology and Business application domains allowed an empirical validation of both the visual representations and the model. The *AdaptaVis* model and the *InfoVis* platform establish the basis for further research on issues related to the coordinated use of multiple data views, the modeling of the coordination process and the integration amongst multiple visual and analytical techniques.

INTRODUÇÃO

Em várias áreas da atividade humana a capacidade de geração, coleta e armazenamento de dados é cada vez maior [Eic 2000][Goe 1999][Kei 1995]. Keim, Kriegel e Seidl, em 1994 [Kei 1994], mencionavam uma estimativa de que a quantidade de informação disponível no mundo duplicaria a cada 20 meses, afirmando que provavelmente o tamanho e o número de bases de dados tendia a crescer ainda mais rapidamente. Tais dados são potencialmente inúteis, um simples depósito volumoso, se não houver formas adequadas de processamento dos mesmos. Explorar e analisar o crescente volume de dados é tarefa difícil, excedendo, inclusive, a capacidade de análise do ser humano, especialmente no caso de dados definidos em espaços de alta dimensionalidade. Suporte para descobrir, de forma automática ou semi-automática, informações de valor ocultas em bases de dados enormes é fornecido por técnicas de ‘Mineração de Dados’ (DM, *Data Mining*) [Kei 1995][Fay 1998]. O termo ‘Descoberta de Conhecimento em Bases de Dados’ (KDD, *Knowledge Discovery in Databases*) é freqüentemente usado como sinônimo de mineração de dados. Entretanto, KDD é entendido e aceito, também, como um processo mais geral no qual a mineração de dados é uma das etapas, de acordo com a definição de Fayyad, Piatetsky-Shapiro e Smyth [Fay 1998][Fay 1996].

Visualização é a disciplina que reúne métodos que permitem interagir com representações visuais de dados, exibidas como imagens, para ganhar entendimento e introspecção sobre eles [Rhy 2003][Sch 1998]. O acoplamento de técnicas de visualização e de

mineração e análise de dados fornece um novo paradigma, único e poderoso, para a exploração e descoberta, referenciado comumente na literatura como ‘Mineração Visual de Dados’ (VDM, *Visual Data Mining*). Nas técnicas de VDM, procura-se unir o visual e o analítico, acoplando técnicas da Visualização a técnicas de Mineração de Dados com o objetivo de aprimorar o processo de descoberta de conhecimento [Won 1999][Hof 1997]. A visualização de dados permite explorar o sentido da visão, aproveitando as capacidades únicas de percepção e cognição do ser humano [Won 1999][Roh 1999][Mac 2001a]. É importante incluir o ser humano no processo de descoberta de conhecimento, pois nosso sistema de percepção processa diferentes tipos de dado de forma bastante flexível, analisa eventos complexos com rapidez, reconhece automaticamente propriedades não usuais e, ao mesmo tempo, desconsidera propriedades sem interesse. Adicionalmente, o ser humano é capaz de traçar conclusões a partir de descrições vagas e informação imprecisa utilizando seu conhecimento geral [Kei 1996].

Ankerst [Ank 2000] propõe uma definição mais ampla de VDM, identificando três abordagens que diferem na forma de uso das técnicas de visualização e que têm em comum um ciclo interativo e iterativo para descoberta de conhecimento: (1) resultados de algoritmos de DM são visualizados para apoiar sua interpretação; (2) resultados intermediários de algoritmos são visualizados para acompanhamento e controle, pelo usuário, do progresso do algoritmo; e (3) os dados brutos são diretamente visualizados para exploração pelo usuário. Essas abordagens ilustram que técnicas de visualização são úteis e aplicáveis em diversas etapas do processo de descoberta de conhecimento.

Diferentes visualizações individuais podem ser integradas em um ambiente, e usadas de forma coordenada aproveitando os pontos fortes de cada uma, estabelecendo uma ligação semântica entre as visualizações, de maneira que ações sobre uma se reflitam nas demais. Essa estratégia, denominada ‘Múltiplas Visualizações Coordenadas’ (CMV, *Coordinated Multiple Views*) [Bal 2000][Nor 2000a], permite a manipulação de conjuntos de dados complexos e de alta dimensionalidade por meio de visualizações apropriadamente selecionadas, sendo adotada em diversas soluções [Nor 2000][Con 2002][May 2002][Gah 2000] que buscam oferecer flexibilidade na definição da coordenação entre visualizações.

Retomando a questão da análise de grandes volumes de informação, dados de natureza espacial e temporal representam hoje um desafio particular [Kop 1998][Est 1999]. MacEachren e

Kraak [Mac 2001a] citam estimativas que sugerem que 80% dos dados gerados na época, em formato digital, incluíam referências espaciais, tais como coordenadas geográficas, endereços e código postal. O atributo espacial define o posicionamento relativo entre os objetos representados em um conjunto de dados espaciais, e as respectivas relações de proximidade e densidade devem ser consideradas na análise [Cha 2000][Kop 1996]. A observação repetida, ao longo do tempo, de um mesmo objeto ou fenômeno espacial estabelece uma componente temporal para os dados que representam o objeto ou fenômeno [Mac 1999]. A correta compreensão de fenômenos de natureza espacial e temporal em domínios como meio ambiente, saúde pública, negócios e turismo, é importante para traçar estratégias de ação que beneficiem as comunidades afetadas por esses fenômenos. Nesse sentido, estratégias de análise integrando abordagens visuais podem auxiliar especialistas a entender os resultados da análise, bem como a caracterizar os dados geradores e, com isso, aumentar a qualidade das análises e a confiança nos resultados, levando a previsões e planejamentos mais consistentes.

O objetivo central deste projeto de doutorado consistiu em investigar a aplicabilidade de técnicas da Visualização na etapa de tratamento de dados em processos de descoberta de conhecimento, concentrando-se na visualização exploratória de conjuntos de dados com atributos temporais e espaciais. Para isso, técnicas visuais foram propostas e integradas em uma plataforma que agrega múltiplas visualizações coordenadas, e é flexível e adaptável aos requisitos de usuários, suas tarefas e diferentes domínios de aplicação. As principais contribuições do trabalho são: (1) a definição de um modelo de arquitetura de software – o modelo *AdaptaVis* – que viabiliza a concepção de uma solução computacional genérica e extensível para visualização de informação; (2) a instanciação desse modelo em uma plataforma de visualização de informação, denominada *InfoVis*; (3) a definição e implementação de novas representações visuais interativas para apoiar o tratamento de dados com atributos temporais, as quais foram integradas à plataforma *InfoVis*; e (4) a validação do modelo proposto e sua plataforma associada, bem como das representações visuais temporais, por meio da sua utilização para apoiar tarefas de análise de dados com atributos temporais e espaciais na plataforma *InfoVis*.

O modelo *AdaptaVis* busca identificar os elementos essenciais, em termos de componentes e funcionalidades, para garantir uma plataforma de software adaptável e extensível. Buscou-se, também, manter uma independência de tecnologia de maneira a conferir

liberdade na escolha da ferramenta computacional para a implementação. As sessões de uso realizadas para validar o modelo e verificar a aplicabilidade das representações temporais propostas foram realizadas com dados de diferentes domínios (Climatologia e Negócios), com participação direta de usuários especialistas do domínio. Os conjuntos de dados tratados nas sessões de uso são relativamente grandes (medidas diárias de chuva registradas em centenas de postos de coleta para tratamento considerando um período mínimo de 30 anos; e dados de transações comerciais diárias em dezenas de cidades para tratamento separado por estado), mas não têm o volume típico dos dados de entrada para técnicas de mineração de dados, pois, no escopo deste projeto, a ênfase foi na verificação da aplicabilidade das técnicas de visualização em tarefas que requerem a exploração de dados com atributos temporais e espaciais.

Técnicas apropriadas para a visualização de conjuntos de dados multidimensionais, o relacionamento entre a visualização exploratória, as tarefas de mineração de dados e o processo de descoberta de conhecimento são abordados no Capítulo 2. Ainda, no mesmo capítulo, são descritos alguns ambientes que suportam o uso coordenado de múltiplas visualizações com diferentes graus de flexibilidade, provendo uma visão do contexto e dos pontos de embasamento no desenvolvimento deste projeto. Os fundamentos e as características do modelo *AdaptaVis* e as representações visuais temporais propostas são detalhadas no Capítulo 3. A instanciación do modelo e das representações visuais temporais propostas, bem como de outras representações, gerou a plataforma *InfoVis*, cujos detalhes de implementação são apresentados no Capítulo 4. As sessões de uso conduzidas após a disponibilização da plataforma, seus objetivos e resultados estão descritos no Capítulo 5. Finalmente, no Capítulo 6, são feitas as considerações finais, avaliando os resultados e apresentando as conclusões e possibilidades de trabalhos futuros. No Apêndice A são apresentadas informações complementares relacionadas a procedimentos e conjunto de dados do domínio da Climatologia, e no Apêndice B, os resultados de estudos preliminares com visualizações de dados pluviométricos.

VISUALIZAÇÃO INTERATIVA E DESCOBERTA DE CONHECIMENTO

O desenvolvimento tecnológico, incorporando avanços em produtos e serviços, gerou condições para medições mais precisas, simulações mais velozes e registro de uma gama variada de tipos de dados, produzindo conjuntos de dados cada vez mais volumosos e complexos. O benefício da quantidade esbarra na natural dificuldade do ser humano em analisar tamanho volume de dados. A necessidade de extrair, analisar, descobrir e utilizar as informações embutidas em dados tem despertado o interesse de pesquisadores em áreas como Visualização, Inteligência Artificial, Estatística, Reconhecimento de Padrões e Banco de Dados [Fay 1998][Won 1999][Fay 2002]. Essa característica multidisciplinar dos processos de análise de dados tem potencial para produzir um conjunto de instrumentos complementares que, integrados e coordenados, podem oferecer um ferramental poderoso para uso no contexto descrito.

Em particular, o uso de representações visuais como instrumento de comunicação para transmitir idéias, resultados e conclusões é anterior ao surgimento dos computadores. Entretanto, o advento dos computadores com recursos gráficos propiciou a geração de imagens mais sofisticadas e complexas, e a manipulação direta das mesmas em tempo real [Kei 2000][Goe 1999][Tuf 1983][Tuf 1990][Bro 1995]. Representações visuais manipuladas dinamicamente facilitam a interpretação dos dados pelo usuário, possibilitando a percepção de padrões, tendências, relacionamentos e exceções embutidos nos mesmos. É nesse contexto que se insere a Visualização, disciplina que investiga estratégias capazes de explorar a capacidade humana de

interpretar representações visuais para ganhar entendimento sobre os dados [Won 1999][Sch 1998][Won 1997][Gan 1996][Car 1999].

O termo ‘Descoberta de Conhecimento em Base de Dados’ (KDD, *Knowledge Discovery in Databases*), estabelecido no primeiro *workshop* da área em 1989 ([Pia 1991] apud [Fay 1998]¹), enfatiza que o conhecimento – *knowledge* – é o produto final de um processo de descoberta guiado por dados. No modelo proposto por Fayyad, Piatetsky-Shapiro e Smyth [Fay 1998][Fay 1996], bastante referenciado na literatura, KDD é o processo global de identificar nos dados um padrão, uma estrutura válida, nova, potencialmente útil e interpretável, conforme ilustrado na Figura 2.1.

Figura 2.1. Etapas do processo de KDD. Adaptado de [Fay 1996].

O processo de descoberta de conhecimento é interativo, iterativo e composto por diversas etapas executadas ao longo de várias atividades. Técnicas de Visualização podem contribuir nas diversas etapas do processo, auxiliando desde a exploração inicial dos dados, passando pela determinação de possíveis modelos, guiando a extração de informações e a avaliação do conhecimento descoberto [Fay 1998][Won 1999][Roh 1999][Rib 1999][Ank 2000][Kei 2001]. A Visualização provê ao analista um instrumental que lhe permite utilizar sua intuição e conhecimento em uma análise exploratória, exigindo pouco ou nenhum conhecimento prévio sobre os dados [Goe 1999], auxiliando, inclusive, a tarefa de apontar um modelo correto para descrevê-los [Fay 2002] [Str 2000]. Um ciclo recursivo em que uma visualização direciona outra visualização, e depois outras sucessivamente, permite ao analista ganhar introspecção

¹ [Pia 1991] Piatetsky-Shapiro, G.; Frawley, W. (editors) – **Knowledge Discovery in Databases**, MIT Press, Cambridge, MA, 1991

gradativamente, introduzindo esse novo conhecimento, juntamente com os prévios, em estágios precoces do processo de descoberta de conhecimento. Isso pode reduzir a quantidade de padrões desinteressantes e diminuir a complexidade do processo como um todo [Der 1997][Gan 1996].

O foco principal neste Capítulo é o uso de técnicas de Visualização no processo de descoberta de conhecimento. Inicialmente, são detalhados os passos do processo de acordo com o modelo da Figura 2.1, explicitando a relação do processo geral com a etapa de mineração de dados. Algumas categorias de técnicas empregadas para mineração de dados são descritas e, também, são apresentadas as implicações da presença de atributos espaciais/temporais nos dados. Em seguida, técnicas de Visualização são descritas, e discutidas as suas aplicações no processo de descoberta de conhecimento, apresentando diferentes exemplos de uso dessas técnicas. Ao final, ambientes que permitem o uso simultâneo, de forma coordenada, de diferentes técnicas de visualização são descritas. Devido ao caráter multidisciplinar do tema, são padronizados os termos ‘item de dado’ ou ‘item’ referindo-se a uma instância ou transação de um conjunto de dados ou registro em um banco de dados; e ‘atributo de dado’ ou ‘atributo’, para indicar uma variável, dimensão, ou campo. Banco ou base de dados, conjunto de dados ou dados são empregados como sinônimos.

2.1 DESCOBERTA DE CONHECIMENTO E MINERAÇÃO DE DADOS

KDD é uma área interdisciplinar e avanços significativos requerem o uso de técnicas de diversas áreas e uma forte interação das respectivas comunidades. As áreas já citadas (Visualização, Inteligência Artificial, Banco de Dados, Estatística, Reconhecimento de Padrões) e Computação de Alto Desempenho são exemplos de tais áreas, que compartilham o objetivo de descoberta de conhecimento de alto nível a partir de conjuntos volumosos de dados brutos. Diferentemente de abordagens tradicionais em Estatística, as abordagens em KDD focalizam na extração de modelos dos dados sem necessariamente uma hipótese formulada *a priori*, e em conjuntos de dados mais volumosos e heterogêneos do que os tradicionalmente manipulados por técnicas estatísticas. Conseqüentemente, uma propriedade importante nos algoritmos é a escalabilidade [Fay 1998], i.e., a capacidade de continuar operando correta e adequadamente a medida que o tamanho dos arquivos de dados aumenta. Questões como os dados não estarem contidos todos na memória principal, estarem espalhados em diversos arquivos ou apresentarem uma organização hierárquica, ou estarem contaminados com valores inesperados ou por ausência de informação,

podem inviabilizar a aplicação direta de métodos estatísticos clássicos, e motivam a criação de novos métodos, bem como o desenvolvimento de técnicas integradas para um processo eficiente e eficaz [Che 1996].

Interatividade e iteratividade são características do processo descoberta de conhecimento, que não requer necessariamente uma seqüência pré-definida, sendo composto por iterações e experimentações ao longo de diversas etapas, o que exige a tomada de muitas decisões pelo usuário. Na Figura 2.1, apresentada no início deste Capítulo, são esquematizadas as etapas que compõem o processo de KDD. No percurso da transformação de dados brutos em informação útil, ou conhecimento, diversos tipos de processamento são aplicados:

- **Seleção**. É gerado um conjunto de dados de interesse para a condução do processo de descoberta, a partir da escolha de determinados itens de dados de um conjunto, da focalização em um subconjunto de atributos, ou de um processo de amostragem;
- **Pré-processamento**. São executadas operações para remoção ou tratamento de ruídos (dados estranhos ou distorções incorporadas aos dados), manipulação de informações ausentes (*missing*) e ordenação segundo algum critério;
- **Transformação**. São aplicadas transformações para reduzir a dimensionalidade dos itens de dados, diminuindo o número de atributos a considerar, ou para a discretização de atributos;
- **Mineração**. São aplicados métodos para a detecção de padrões nos dados resultantes das etapas anteriores. A escolha do método a ser aplicado envolve, entre outras, considerações sobre o tipo de informação a ser extraída (classes, agrupamentos, sumários) e o tipo do dado a ser processado (categórico, numérico). Esse tópico é detalhado mais adiante;
- **Interpretação/avaliação**. É realizada uma interpretação dos padrões extraídos, que pode levar à repetição das etapas anteriores. Se aceito, o conhecimento descoberto pode ser incorporado a outros sistemas, documentado para uso posterior, e comparado com outros descobertos anteriormente.

O termo KDD é freqüentemente usado como sinônimo de ‘Mineração de Dados’ (DM, *Data Mining*). Entretanto, segundo o modelo de Fayyad, Piatetsky-Shapiro e Smyth (Figura 2.1), a mineração é uma etapa do processo de descoberta de conhecimento, responsável pelo mapeamento dos dados em algum tipo de informação. O grau de interesse e a validade das muitas informações que podem ser extraídas dos dados são dependentes do domínio de aplicação

e dos objetivos do usuário. A mineração de dados é a componente do processo de KDD relacionada ao uso de algoritmos para a extração e enumeração de padrões a partir dos dados [Fay 1996]. Embora DM tenha tido uma conotação negativa, sendo denominada ‘dragagem de dados’ (*data dredging*) na literatura da Estatística, ela é definida como o processo de análise secundária de grandes bases de dados a fim de encontrar relacionamentos não suspeitos que podem ser de interesse para o usuário, em contraposição à análise primária, na qual os dados são coletados com uma particular questão ou conjuntos de questões em mente [Han 1998].

Considerando o processo completo de KDD, estima-se que a etapa de DM represente entre 15% e 25% do esforço do processo de KDD [And 1999]([Bra 1996] apud [Goe 1999]²). Cabena et al. [Cab 1998] consideram ainda uma etapa anterior, de determinação dos objetivos, que ocupa 20% do tempo; ficando as demais etapas com a distribuição: preparação (60%); mineração (10%); análise dos resultados e assimilação do conhecimento (10%). Observando apenas as três últimas etapas (retângulos com cantos arredondados na Figura 2.2), que somam 80%, verifica-se que a preparação dos dados corresponde a 75% do tempo. Uma enquete em Kdnuggets³, entre 30/09 e 12/10/2003, da qual participaram 187 votantes respondendo à pergunta “Qual o % de tempo gasto em seu projeto de DM com a limpeza e a preparação dos dados?”, produziu o seguinte resultado:

- 20% ou menos do tempo: 8% dos votantes;
- 21%-40%: 4%;
- 41%-60%: 25%;
- 61%-80%: 39%;
- mais de 80%: 25%.

Figura 2.2. Etapas agrupadas do processo de KDD [Oli 2004].

² Brachman, R.; Anand, T. – The Process of Knowledge Discovery in Databases: A Human-Centered Approach, In Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (eds.), *Advances in Knowledge Discovery and Data Mining*, MIT Press, 1996

³ www.kdnuggets.com

Atividades em mineração de dados podem ser ‘preditivas’ – prever o valor futuro de um atributo com base em valores de outros atributos do conjunto de dados, utilizando uma generalização formulada a partir de exemplos, ou ‘descritivas’ – descrever itens de dados, identificando-os com base em padrões ou comportamentos encontrados no conjunto de dados [Rez 2003]. Ainda, técnicas para mineração podem ser categorizadas segundo diversos critérios, tais como o tipo de informação a ser extraída, ou o tipo de dado a ser processado (BD relacional, dados com atributos espaciais e/ou temporais, por exemplo), ou ainda o tipo de método utilizado (estatístico, aprendizagem de máquina, dentre outros). A seguir, são descritas algumas categorias apresentadas por Chen, Han e Yu [Che 1996], que tem como base o tipo de informação extraída dos dados, a qual pode ser avaliada como útil – conhecimento – pelo usuário:

- **Regras de Associação.** O objetivo é obter um conjunto composto de regras de associação na forma $A_1 \wedge \dots \wedge A_m \Rightarrow B_1 \wedge \dots \wedge B_n$, com A_i e B_j representando valores de atributos de um conjunto de dados, isto é, descobrir associações relevantes entre atributos tal que a presença de alguns atributos implica na presença de outros na mesma transação. Duas medidas importantes utilizadas na busca por regras de associação são a *confiança* e o *suporte*. Uma regra $X \Rightarrow Y$ tem confiança c e suporte s , se $c\%$ dos itens de dados que contêm X também contêm Y , e $s\%$ dos itens de dados na base de dados contêm X ou Y , respectivamente. As regras de interesse, denominadas fortes, são aquelas que têm alta confiança e forte suporte;
- **Classificação de Dados.** É feita uma divisão dos dados em classes com base nos valores de alguns de seus atributos, de acordo com um modelo de classificação (um descritor para cada classe) construído a partir de um subconjunto de treinamento;
- **Agrupamento (Clustering) de Dados.** Esse método é denominado classificação não supervisionada. É executado um agrupamento de dados, sem predefinição de descritores de classe, baseado no princípio conceitual de agrupamento: “maximizar a similaridade intraclasse e minimizar a similaridade interclasse”. O objetivo é identificar regiões, no espaço de atributos, densamente povoadas, os agrupamentos (*clusters*), de acordo com alguma medida de distância;
- **Pesquisa por similaridade baseada em padrões.** A busca por padrões similares em bases de dados com atributos espaciais e temporais é importante para descobrir e prever risco, causalidade e tendências associadas a um padrão específico. Operações desse tipo são aplicáveis às áreas financeira, médica e meteorologia, por exemplo. As buscas devem resultar, em geral, em correspondências similares e não em correspondências exatas. Por

exemplo, ações com movimentos de preço similares, ou imagens com padrões climáticos similares. Distância euclidiana e correlação são métricas usuais para medir similaridade.

Dados espaciais trazem embutidas informações relativas à topologia, distância e direção derivadas da relação de vizinhança entre os objetos espaciais que representam, as quais devem ser consideradas em sua análise [Kop 1996], constituindo um caso particular de KDD [And 1999]. O entendimento de tais tipos de dados, descobrindo relacionamentos entre objetos espaciais, considerando, inclusive, atributos não espaciais, constitui o objetivo de técnicas de ‘Mineração de Dados Espaciais’ (SDM, *Spatial Data Mining*) [Kop 1998]. A disseminação de bases de dados espaciais, como as relacionadas com sistemas de informação geográfica e de sensoriamento remoto, provocou um crescente interesse em técnicas de busca por padrões espaciais [Est 1999]. Estimativas sugerem que 80% dos dados em formato digital apresentam referência digital, tais como coordenadas geográficas, endereços e código postal [Mac 2001a]. Ecologia, monitoramento ambiental, saúde pública, logística em negócios, turismo e controle de tráfego são exemplos de domínios para os quais tais técnicas são particularmente relevantes [Cha 2000]. Conjuntos de dados espaciais podem conter descrições de fenômenos ou eventos que se repetem ao longo do tempo, como ocorrências climáticas. A presença de múltiplas instâncias de um mesmo evento caracteriza uma componente temporal no conjunto. Essa componente fornece informações sobre a ocorrência de um evento em relação a outros, e sobre a sua duração, que podem ajudar a descobrir e explicar relacionamentos e padrões nos eventos [Mac 1999][Ram 2000].

Pela ilustração da Figura 2.2, o processo de KDD é composto, resumidamente, pelas etapas de preparação dos dados brutos, extração de informação e descoberta de conhecimento. Representações visuais são úteis em todas as etapas, e a Seção seguinte descreve o relacionamento entre técnicas de Visualização e o processo de KDD.

2.2 VISUALIZAÇÃO E DESCOBERTA DE CONHECIMENTO

Visualização é o processo de obter entendimento e introspecção sobre um fenômeno por meio de uma representação visual criada a partir de dados que o descrevem. Os termos mais específicos, Visualização Científica e Visualização de Informação, são utilizados de acordo com a natureza dos dados sendo manipulados. O primeiro termo, estabelecido no relatório *Visualization in*

Scientific Computing ([McC 1987] apud [Sch 1998]⁴) da NSF (*National Science Foundation*) em 1987, denota a visualização de dados obtidos tipicamente em disciplinas de Ciências e Engenharia. O segundo refere-se à visualização de dados abstratos, i.e., que não têm necessariamente um objeto correspondente no mundo físico, como documentos hipertexto na WWW e estrutura de diretórios em um sistema de arquivos de computador [Sch 1998][Won 1997]. Entretanto, Rhyne [Rhy 2003] questiona a necessidade de diferenciação entre Visualização de Informação e Científica em virtude do uso intercambiado das técnicas de uma e outra, particularmente em domínios emergentes, como Visualização Geográfica e Visualização em Bioinformática.

Na perspectiva de Spence [Spe 2001], a visualização é uma atividade interna, de cognição do ser humano, que conduz à formação de um modelo mental com base em estímulo visual recebido através de uma imagem, facilitada por ferramentas computacionais para visualização. Um ponto complexo no projeto de tais ferramentas é o de unir representações visuais com controles interativos que confirmam um certo grau de liberdade ao usuário em domínios e tarefas diversificadas. Além disso, deve ser considerada a finalidade da visualização: visualização descritiva para apresentar e sumarizar idéias bem definidas, visualização analítica para verificar uma hipótese, ou visualização exploratória para buscar hipóteses [Gri 2002][Ank 2000] que, basicamente, relacionam a visualização com diferentes etapas do processo de KDD. Card, Mackinlay e Shneiderman [Car 1999] também apresentam a Visualização como um instrumento de suporte externo para ampliar a cognição, propiciada por imagens geradas no computador e manipuladas diretamente.

O processo de descoberta de conhecimento pode ser fortalecido pela incorporação de técnicas de Visualização, como um instrumental para estimular a participação mais ativa do usuário na exploração e análise de dados; oferecendo-lhe suporte para monitoramento, avaliação e controle do processo, e elevando o grau de confiança no resultado [Ank 2000][Kei 2002]. Shneiderman [Shn 1996], analisando vários projetos de busca visual de informação, resumiu as principais atividades na seqüência que denominou *Visual Information Seeking (VIS) mantra: Overview first, zoom and filter, then details-on-demand* (Visão panorâmica inicialmente, foco e filtragem, seguida de detalhes sob demanda). A integração das habilidades únicas da mente

⁴ [McC 1987] McCormick, B.H.; DeFanti, T.A.; Brown, M.D. – Visualization in Scientific Computing, **Relatório do NSF Advisory Panel on Graphics, Image Processing and Workstations**, 1987

humana na percepção de padrões, exceções, tendências e relacionamentos, com a capacidade de processamento dos computadores, viabiliza a criação de ambientes poderosos de exploração e mineração de dados e descoberta de conhecimento. A complementação mútua entre recursos de visualização e procedimentos analíticos, desenvolvidos em disciplinas como Visualização, Mineração de Dados, Estatística, Aprendizagem de Máquina e Banco de Dados, com extensões para manipular conjuntos de dados volumosos, multidimensionais e multivariados, oferecem a base para esse propósito [Won 1999][Roh 1999][Rib 1999][Gan 1996][Fay 2002].

O uso integrado de técnicas de DM e Visualização, balanceando entre o processo automático e o interativo, é referenciado na literatura como ‘Mineração Visual de Dados’ (VDM, *Visual Data Mining*) [Won 1999][Gan 1996][Ins 1997][Kei 2002][Ank 2000]. Outros termos aparecem, como ‘Visualização da Descoberta’ (DV, *Discovery Visualization*) [Rib 1999] e Análise Visual [Roh 1999]. Técnicas de Visualização proporcionam vantagens como interação direta e orientação do processo pelo usuário [Kei 2000], reforçando o processo de análise [Roh 1999]. Wong [Won 1999] aponta duas formas de integrar Visualização e DM. Na primeira, denominada acoplamento forte, a Visualização e o processo analítico são integrados em uma única ferramenta, aproveitando os pontos fortes de cada uma das áreas. Tal acoplamento permitiria, por exemplo, que alguns passos executados automaticamente pelo algoritmo que implementa o processo analítico fossem substituídos por decisões humanas guiadas pela representação visual observada. Segundo Hinneburg, Keim e Wawryniuk [Hin 1999], que integraram fortemente a visualização e algoritmos de agrupamento, avaliações experimentais mostram que a combinação de técnicas visuais e automáticas melhora consideravelmente a eficiência do processo de DM e estimula um melhor entendimento dos resultados. Outros trabalhos [Tej 2003][Tra 2002] também empregam a estratégia de acoplamento forte. Por outro lado, em um acoplamento fraco, técnicas das duas áreas são simplesmente intercaladas, possibilitando um aproveitamento parcial do potencial no uso conjunto.

Em geral, técnicas de Visualização são usadas juntamente com técnicas de interação [Kei 2000], uma vez que devem apresentar interfaces eficazes para atender às necessidades de interatividade e iteratividade do processo de descoberta [Won 1999][Rib 1999], permitindo a exploração dinâmica dos dados. Flexibilidade em processos de exploração e análise de dados é alcançada por meio de operações que permitam (1) seleção de itens de dados individuais ou subconjuntos de interesse dos mesmos, (2) focalização em determinados atributos, (3) visão

panorâmica ou detalhada de trechos dos dados, (4) rolagem pelos dados para acesso a diversas partes do conjunto, (5) navegação em estruturas hierárquicas e (6) controle do mapeamento dos atributos nas imagens [Gri 2002][Mac 1999][Nor 1997]. Uma forma alternativa de exploração e análise é fornecida por técnicas de distorção, cujo princípio é prover uma visualização detalhada de uma porção dos dados, com o restante sendo visualizado em um nível de detalhe menor, mas possibilitando ainda a percepção do contexto global [Kei 2002].

Múltiplas visualizações de um mesmo conjunto de dados, ou de conjuntos distintos relacionados, permitem observá-los sob várias perspectivas, bem como explorar os pontos fortes e minimizar o efeito dos pontos fracos das técnicas envolvidas [Kei 1996][Eic 2000a]. Nesse caso, é interessante que as múltiplas visualizações estejam coordenadas de forma que ações sobre uma delas sejam propagadas para as demais [Nor 1999] durante as etapas exploratórias. Nesse contexto, as interações ocorrem em duas modalidades: intravisualização e intervisualizações; no caso dessa última modalidade, as visualizações coordenadas também são ditas “fortemente acopladas”, com significado distinto do caso de integração entre técnicas de Visualização e algoritmos de mineração. Uma lacuna importante a ser preenchida, apontada por Baldonado, Woodruff e Kuchinsky [Bal 2000], é a ausência de diretrizes específicas para o uso e desenvolvimento de sistemas que oferecem múltiplas visualizações coordenadas.

A criação de representações visuais de dados precede o surgimento de computadores, como bem ilustrado por Tufte [Tuf 1990][Tuf 1983]. O emprego do computador para essa finalidade propiciou o desenvolvimento de novas técnicas de visualização e a extensão das existentes para manipular conjuntos de dados maiores e permitir interação. Muitas técnicas para visualização exploratória de grandes bases de dados foram propostas e implementadas, e têm ganho importância [Kei 2000][War 2002]. O termo ‘multidimensional’ é usado para denotar dados com mais de três atributos (os três atributos espaciais x , y e z , e um temporal t , por exemplo).

Keim e Kriegel [Kei 1996] descrevem técnicas de visualização multidimensional agrupando-as nas categorias de ‘técnicas orientadas a pixel’, ‘geométricas’, ‘iconográficas’ e ‘hierárquicas’. Os princípios de cada categoria são apresentados a seguir.

2.2.1 TÉCNICAS ORIENTADAS A PIXEL

Nessa classe de técnicas, a idéia básica consiste em associar um atributo dos dados a uma janela na tela, mapeando a cor de cada pixel na janela segundo o valor do atributo que ele representa. Para conjuntos de dados com m atributos, ou dimensões, a tela é particionada em m janelas, como ilustrado na Figura 2.3. A associação de um único pixel a cada item de dado maximiza a quantidade de dados que pode ser apresentada. Por exemplo, se um único atributo é apresentado em uma tela de computador na resolução de 1280x1024, é possível exibir simultaneamente mais de 1.000.000 de valores. Correlação e dependência funcional entre atributos são relações que podem ser detectadas pela análise de regiões correspondentes nas janelas [Kei 1996][Kei 2000].

Figura 2.3. Múltiplas janelas para um caso de 6 atributos em técnicas orientadas a pixel. Adaptado de [Kei 2000].

Diversos aspectos devem ser considerados para a aplicação bem sucedida desse tipo de técnica [Kei 2000][Kei 1996][Kei 1994][Kei 1994a]:

- **Arranjo dos pixels nas janelas.** A distribuição apropriada dos pixels nas janelas tem implicações na percepção de relações existentes nos dados. Arranjos em espiral simples (Figura 2.4-a) ou que usam as curvas de Peano-Hilbert e Morton (Figura 2.4-b) possibilitam que dados relacionados sejam posicionados e apresentados próximos uns aos outros, diferente de um arranjo no esquema linha-a-linha (esquerda para direita/cima para baixo). Entretanto, com as curvas citadas, o usuário pode ter dificuldade em estabelecer relações entre as múltiplas janelas, em função do traçado mais complexo dessas curvas. Dados com uma ordenação inerente, como séries temporais, podem ser visualizados

ordenadamente diretamente da base de dados, enquanto o arranjo de dados sem uma ordenação natural pode ser ordenado com base no resultado de uma consulta à base de dados. Nesse último caso, além dos itens de dados que satisfazem plenamente a consulta, é importante exibir também aqueles que o fazem parcialmente, proporcionando uma visão mais abrangente do conjunto de dados. Nessa situação, a cor do pixel representa uma medida da distância entre o valor do atributo e o valor fornecido na consulta. A fórmula para o cálculo da distância depende do domínio da aplicação e do tipo de dado (numérico, categórico, data, alfabético). Os atributos referentes a um determinado item de dado ocupam a mesma posição relativa nas diversas janelas, definida por uma distância global calculada pela combinação das distâncias de cada atributo ponderadas por um peso que determina a relevância dada ao atributo. Os itens de dados que satisfazem integralmente a consulta têm distância global igual a zero. Os valores não nulos de distância podem ser positivos ou negativos e, assim, informam a direção de variação (para mais ou para menos) dos valores dos atributos em relação à consulta, a qual pode ser visualizada pelo acréscimo de dois eixos com origens no centro de uma janela, conforme a Figura 2.4-c. Contudo, o uso de eixos pode produzir regiões sem preenchimento, em virtude da probabilidade da distribuição dos valores de distância, entre positivo e negativo, não ser uniforme. Outra consequência é que o número de itens que podem ser apresentados simultaneamente é menor, pois uma janela é dividida em quatro quadrantes;

- **Mapeamento da cor.** O uso de cor permite, em relação à escala de cinzas, um número maior de JNDs (*Just Noticeable Differences*), que são as cores perceptualmente distintas para o ser humano. Um grande desafio é determinar uma escala de cores que, além de maximizar o número de JNDs, também seja natural para o usuário, evidenciando as relações entre os atributos e evitando a introdução de artefatos visuais [Tut 1998]. Para uma distribuição de um único parâmetro, o brilho é o fator mais importante na distinção das cores. Características particulares de um domínio ou aplicação específicos podem sugerir estratégias para o uso adequado de cor e brilho;
- **Formato das janelas.** O formato retangular para as janelas garante um bom uso do espaço na tela do computador; entretanto, a percepção de relacionamento entre atributos pode ser dificultada pela sua distância relativa na tela. Quanto maior o número de atributos (maior número de janelas) visualizados, maior o problema. Um formato alternativo é o circular, adotado na técnica denominada ‘Segmentos de Círculo’ (*Circle Segments*) [Kei 2000][Ank 1999], conforme ilustra a Figura 2.5-a. Cada atributo é visualizado em um

segmento do círculo, com valores posicionados a partir do seu centro e seguindo o caminho dado pelas linhas de desenho, que são ortogonais às linhas divisoras, conforme mostra a Figura 2.5-b. Outra alternativa, embora adote o formato retangular, é o arranjo da técnica ‘Gráfico de Barras de Pixels’ (*pixel bar chart*)[Kei 2001a], que aproveita o interior das barras, colorindo cada pixel interno com base no valor de um atributo representado. Todas as barras têm a mesma altura para usar ao máximo o espaço disponível, e a largura varia de acordo com a quantidade de itens a serem representados em cada barra. Um segundo atributo, denominado atributo de divisão, é escolhido para separar as barras, e outros dois são escolhidos para ordenar os itens de dados na distribuição dentro das barras, conforme ilustra a Figura 2.6-a. O posicionamento interno às barras é determinado por uma heurística que busca manter itens similares próximos e evitar sobreposição entre os pixels. Um exemplo é mostrado na Figura 2.6-b, na qual são visualizadas transações de comércio eletrônico. Cada pixel representa um cliente, o atributo de divisão é o tipo de produto, os atributos de ordenação interna são número de visitas ao *site* e o valor gasto, e a cor de cada pixel mapeia o valor gasto pelo cliente correspondente (cores brilhantes e escuras mapeiam valores maiores e menores, respectivamente).

Figura 2.4. Possíveis esquemas para arranjo de pixels em técnicas orientadas a pixel [Kei 1994][Kei 1996].

Figura 2.5. Formato circular para apresentação dos pixels em Segmentos de Círculo. Adaptado de [Kei 2000].

Figura 2.6. Esquema de arranjo dos pixels em Gráficos de Barras de Pixels. Adaptado de [Kei 2001a].

2.2.2 TÉCNICAS DE PROJEÇÃO GEOMÉTRICA

Técnicas nessa categoria buscam encontrar projeções, bi ou tridimensionais, que auxiliem a revelar informações de interesse em conjuntos de dados multidimensionais. Uma técnica bastante conhecida é denominada ‘Coordenadas Paralelas’ (*Parallel Coordinates*), na qual um espaço de dimensão k é mapeado em um espaço visual bidimensional, usando k eixos eqüidistantes e paralelos a um dos eixos principais (x ou y). Cada eixo está associado a um atributo, e sobre ele é mapeado linearmente o respectivo intervalo de valores dos dados. Cada item de dado é exibido como uma linha poligonal que intercepta cada um dos eixos no ponto correspondente ao valor do atributo associado ao eixo (polilinha vermelha na Figura 2.7). Ao gerar uma representação planar, essa técnica transforma relações multivariadas em padrões bidimensionais [Ins 1997][Weg 1996] que permitem identificar características como diferenças na distribuição dos dados e correlação entre atributos.

Observando a Figura 2.7 é possível identificar uma correlação negativa entre os atributos associados aos eixos 6 e 7, evidenciada pelo cruzamento em X (destacado em verde) das linhas entre os eixos [War 2002][War 1994]. Ainda no penúltimo eixo, associado ao atributo 6, é possível perceber uma concentração de linhas em determinados intervalos de valores (retângulos azuis). A técnica apresenta problemas devido à sobreposição de linhas (ver Figura 2.8) e, em consequência, uma baixa quantidade de itens pode ser apresentada simultaneamente sem a ocorrência de congestionamento visual (alguns poucos milhares) [Kei 1996]. A visualização de grandes volumes de dados requer que a técnica esteja integrada a operações de interação adequadas para seleção e filtragem de itens de interesse.

Figura 2.7. Correlação e distribuição de valores em Coordenadas Paralelas [War 1994].

Figura 2.8. Sobreposição de linhas com apresentação de 473 itens [Ins 1997].

Matriz de ‘Gráficos de Dispersão’ (*Scatterplots*) é uma das técnicas mais antigas e populares para a projeção de dados de alta dimensionalidade em uma representação visual bidimensional. Projeções dos atributos, aos pares, são organizadas em formato de matriz, com

cada célula associada a dois atributos identificados nas linhas e colunas da matriz, permitindo a observação de relacionamentos entre os atributos [Won 1997]. A Figura 2.9 exemplifica um caso no qual a correlação linear entre dois atributos (região delimitada pela elipse) pode ser visualmente observada pela distribuição dos pontos na célula. Como na diagonal ambos os eixos representam o mesmo atributo, é desenhada uma linha de 45°; entretanto, no caso da Figura 2.9, a diagonal foi aproveitada para fornecer uma informação sobre a distribuição de valores do atributo correspondente em um histograma. A facilidade de interpretação é uma vantagem dessa técnica, por outro lado, a quantidade de atributos que podem ser apresentados simultaneamente é bastante limitada [War 1994]. Considerando n atributos, $n(n-1)/2$ gráficos são suficientes para mostrar as combinações possíveis de atributos, pois a parte inferior à diagonal (marca vermelha na Figura 2.9) é equivalente à parte superior e, portanto, pode ser suprimida [Gri 2001] sem perda de informação.

Figura 2.9. Correlação linear em uma matriz de Gráficos de Dispersão [Gri 2001].

2.2.3 TÉCNICAS ICONOGRÁFICAS

Nas técnicas iconográficas, cada item de dado multidimensional é mapeado em um ícone cujos atributos visuais podem ser configurados para exibir atributos dos dados. Um ícone que permite visualizar grandes volumes de dados é a ‘Figura de Arestas’ (*Stick Figure*). As duas dimensões da tela são usadas no mapeamento de dois atributos dos dados, com os demais atributos sendo mapeados para ângulos e/ou comprimentos de segmentos da figura de arestas. Na Figura 2.10 é mostrada uma configuração com cinco arestas e algumas possíveis variações. As orientações das

arestas permitem mapear quatro atributos e a inclinação do corpo mapeia um quinto atributo. Variações de comprimento, espessura e cor das arestas criam outras possibilidades de representação. Um conjunto de figura de arestas gera padrões de textura que podem ser de detecção e interpretação fáceis, entretanto, a configuração escolhida para as arestas influencia no discernimento visual de padrões. Uma outra limitação é o número de dimensões que podem ser apresentadas simultaneamente, da ordem de aproximadamente uma dezena [Kei 1996][Won 1997]. A Figura 2.11 é uma imagem composta de ícones do tipo figura de arestas, gerada a partir de 5 imagens de satélite da região dos Grandes Lagos, na qual diversas texturas são identificadas [Gri 2001].

Figura 2.10. Configurações para Figura de Arestas usando cinco arestas [Kei 1996].

Figura 2.11. Diferentes texturas representadas por Figura de Arestas [Gri 2001].

2.2.4 TÉCNICAS HIERÁRQUICAS

Nas técnicas hierárquicas de visualização, subespaços de um espaço de dados k -dimensional (não necessariamente hierárquico) são organizados e apresentados na forma de hierarquias, projetando ou embutindo dimensões (que representam atributos) dentro de outras dimensões. A técnica de ‘Empilhamento Dimensional’ (*Dimensional Stacking*) pode ser aplicada para visualizar dados categóricos ou que foram agrupados em categorias, e nela o espaço multidimensional é dividido em subespaços bidimensionais. Agrupamentos, pontos discrepantes e padrões podem ser detectados [Kei 1996][Won 1997][War 1994].

A Figura 2.12 mostra o resultado de uma aplicação do empilhamento dimensional em um conjunto de dados contendo as medidas de comprimento e altura das sépalas e pétalas de flores, totalizando quatro medidas que são organizadas em dois níveis para serem exibidas com o empilhamento dimensional. No primeiro nível (mais externo do gráfico) estão as três categorias dos comprimentos das pétalas em x , e da sépala em y , que visualmente formam os quadrados maiores delimitados com linhas mais escuras na figura. No outro nível (mais interno do gráfico), estão as medidas de altura, também categorizadas, das mesmas partes da flor. Cada quadrado menor, que representa um conjunto de quatro medidas, é preenchido com uma ou mais cores associadas ao tipo de flor e, assim, é possível observar se existem agrupamentos que levam a uma caracterização do tipo de flor pelas medidas de suas sépalas e pétalas (os quadrados preenchidos com uma única cor mostram grupos bem distintos para cada tipo de flor) [Hof 1999].

Hoffman [Hof 1999] alerta que cada atributo deve ser categorizado em não mais do que quatro ou cinco categorias e, que a visualização de mais de nove atributos é difícil. Ambas as limitações são decorrentes do espaço disponível na tela do computador, e podem ser aliviadas com operações de interação, tais como ampliação/redução e deslocamento. A escolha do arranjo hierárquico dos atributos, bem como o critério de categorização dos dados, também são parâmetros determinantes da sua eficácia da técnica de empilhamento dimensional como ferramenta de análise.

Figura 2.12. Resultado do empilhamento dimensional. Quatro atributos são aninhados aos pares [Hof 1999].

Essas e outras técnicas para visualização multidimensional são revisadas por Grinstein, Trutschl e Cvek [Gri 2001], e Oliveira e Levkowitz [Oli 2003]. Keim [Keim 2002] apresenta um critério de classificação para técnicas de VDM utilizando a categorização de técnicas de Visualização apresentada, juntamente com técnicas de interação/distorção e tipos de dados (unidimensional, bidimensional, multidimensional, texto/web, hierarquia/grafos e algoritmo/software).

Técnicas de visualização podem ser aplicadas de diferentes maneiras na exploração e análise de dados. A seguir, são apresentados exemplos ilustrativos dessas formas de uso: visualização para realizar uma tarefa de mineração de dados, integração entre um aplicativo de visualização e um sistema de informação geográfica, e o uso de múltiplas visualizações integradas e coordenadas.

2.2.5 VISUALIZAÇÃO COMO APOIO À MINERAÇÃO DE DADOS

Empregando uma técnica de visualização multidimensional, o Segmentos de Círculo, associada a recursos de interação, Ankerst et al. [Ank 1999] apresentam uma estratégia totalmente interativa de construção de árvores de decisão, um classificador bastante popular usado para tarefas de mineração voltado, principalmente, para atributos categóricos (i.e., cujo domínio consiste de

conjuntos de valores discretos). O processo interativo permite retroagir passos na fase de construção, e superar a limitação de divisão binária para atributos numéricos, presente em muitas árvores de decisão. A classificação visual interativa também permite reduzir o tamanho da árvore, além de prover uma acurácia equivalente a da árvore construída algoritmicamente, mesmo quando o usuário não tem conhecimento prévio dos dados. Essa abordagem permite que o conhecimento de um especialista seja aproveitado já durante a fase construção da árvore e, que, ao final, ele tenha um entendimento melhor dos dados e da própria árvore. Em consequência, o conhecimento é transferido em ambas as direções, isto é, o especialista atua em decisões no processo de construção e recebe uma resposta (*feedback*) por meio das representações visuais, o que é viabilizado pela visualização da informação extraída e dos dados brutos.

Os dados de treinamento para o classificador são visualizados com Segmentos de Círculo, e cada atributo é arranjado de acordo com sua ordenação individual, ou seja, não é calculada uma distância global. Cada pixel, que representa o valor de um determinado atributo, recebe a cor da classe à qual pertence o item de dado associado. A escala de cores adotada para rotular classes foi definida experimentalmente, e foi observado que a configuração mais adequada, com o maior número de cores distintas, é aquela que mantém a saturação e a intensidade constantes em todo o intervalo, e o matiz varia linearmente.

Os retângulos do diagrama da Figura 2.13 representam os passos do processo de classificação interativa. Inicialmente, o conjunto de dados de treinamento completo é visualizado na janela esquerda da interface do classificador, mostrada na Figura 2.14, enquanto na janela direita é mostrada uma árvore vazia. Em seguida, passando pelos passos 2 e 3, o usuário seleciona um atributo de divisão e um número arbitrário de pontos de divisão (na Figura 2.14, isso é indicado no segmento do Segmentos de Círculo em destaque), provocando a expansão da árvore de decisão no quadro correspondente. Em seguida, o usuário pode remover um nível da árvore ou selecionar um de seus nós. Ao selecionar um nó, o usuário pode atribuir ao mesmo um rótulo de classe e torná-lo uma folha da árvore, ou requisitar a visualização dos dados de treinamento associados a ele. Nesse último caso, é gerada uma nova visualização com todos os atributos, exceto os usados como critério de divisão no mesmo trecho da árvore.

Figura 2.13. Passos da Classificação Visual [Ank 1999].

Os pontos de divisão são escolhidos interativamente sobre as linhas de divisão, exemplificado na Figura 2.14. Tais linhas são apresentadas em escala maior, juntamente com controles deslizantes, possibilitando uma escolha mais precisa de pontos de divisão (janela inferior com linhas de divisão ampliadas e ponto de divisão exato na figura). Na Figura 2.15 é ilustrado um determinado estágio de construção de uma árvore de decisão. O atributo 7 é a raiz da árvore, com um ponto de divisão no valor 6. O filho à esquerda está rotulado com a cor correspondente à classe, como 'Bypass'. O filho à direita é composto por outros atributos. Os atributos 2 e 1 formam uma sub-árvore ternária.

Figura 2.14. Interface da ferramenta de Classificação Visual [Ank 2000].

Figura 2.15. Um estágio da árvore de decisão gerada pela ferramenta de Classificação Visual [Ank 1999].

2.2.6 INTEGRAÇÃO DE DOIS APLICATIVOS DISTINTOS

O XGobi é um aplicativo que permite visualizar e explorar dados multidimensionais usando, dentre outras formas, coordenadas paralelas e matrizes de gráficos de dispersão. Símbolos gráficos ou cores podem ser usados para rotular os elementos representados. Opções para o tratamento de grandes volumes de dados incluem amostragem aleatória dos itens de dados, manipulação de arquivos em formato binário, que garante rapidez de acesso em relação a arquivos textuais, e representação de atributos por pixels, que contribuem para evitar congestionamento visual. Duas outras características importantes dessa ferramenta são o tratamento de valores ausentes e suporte para comunicação entre processos através de RPC (chamada a procedimento remoto, *remote procedure call*). Essa última característica é utilizada para a integração do XGobi com o ArcView, um sistema de informações geográficas, permitindo a cada ferramenta estender e complementar suas capacidades. A Figura 2.16 exemplifica esse uso conjunto. Os elementos selecionados, dentro do retângulo vermelho, são mostrados também no mapa. Desenvolvido originalmente para ambientes Unix/X-Windows, o XGobi requer uma adequação para ser usado em plataformas MS-Windows [Sym 2002][Sym 2000][Swa 2001]⁵. Ele é a base para o desenvolvimento do GGobi⁶, que apresenta melhorias como gerenciamento de cores mais flexível, melhor portabilidade para o ambiente MS Windows e manipulação de arquivos XML [Swa 2002].

Figura 2.16. Uso conjunto de XGobi e ArcView [Sym 2000].

⁵ www.research.att.com/areas/stat/xgobi/

⁶ www.ggobi.org

2.2.7 MÚLTIPLAS VISUALIZAÇÕES INTEGRADAS E COORDENADAS

MacEachren et al. [Mac 1999] discutem a integração entre ‘Visualização Geográfica’ (GVis, *Geographic Visualization*) e KDD para a construção de conhecimento pela exploração de grandes volumes de dados espaço-temporais derivados de estudos do meio-ambiente. Esses dados, provenientes de sistemas de monitoramento de ecossistemas, por exemplo, têm atributos espaciais por serem referenciados geograficamente, e o atributo temporal resulta das repetidas observações de um fenômeno. GVis foi definido como o uso de representações visuais concretas – em papel, no computador, ou outro meio – para tornar visível o contexto e os problemas espaciais. A diferença entre GVis e KDD reside na ênfase no sistema visual humano, no caso do primeiro, ou em métodos computacionais para processar os dados, no caso do segundo. A proposta de integração é feita porque os autores acreditam que a visualização tem papel importante em todas as etapas do processo de KDD.

Um protótipo integrando GVis e KDD foi implementado utilizando três ferramentas: (1) o *Data Explorer* (DX) da IBM (atualmente, OpenDX, disponível como código aberto) para realizar as análises e as visualizações, (2) *scripts* em Tcl/Tk para implementar a interface gráfica com o usuário e (3) programas na linguagem C para integrar os scripts Tcl/Tk e o DX. Foram utilizadas três formas de representação visual:

- **GeoVisão (Geoviews)**. Janelas tridimensionais, nas quais duas dimensões são usadas para mapear atributos espaciais, e a terceira dimensão para mapear elevação/profundidade ou tempo. Na janela inferior direita da Figura 2.17 está ilustrada uma geovisão, cujo eixo z representa tempo. Os símbolos gráficos (*glyphs*) representam valores de precipitação ao longo do tempo nos locais aonde estão posicionados, identificados no mapa de fundo;
- **Gráficos de dispersão 3D**. Permitem representar relações entre três atributos. Uma característica interessante é a espacialização de atributos não espaciais, o que permite que uma relação entre itens de dados seja revelada pela proximidade relativa dos mesmos. Itens posicionados próximos provavelmente apresentam uma similaridade maior do que os distantes. A janela inferior esquerda da Figura 2.17 exemplifica essa representação;
- **Coordenadas Paralelas**. A janela superior da Figura 2.17 exemplifica o uso dessa representação, discutida na Subseção 2.2.2, apresentando o resultado de uma classificação. Os itens de uma mesma classe são pintados com a cor correspondente, e o exame das outras visualizações permite verificar se o resultado é apropriado.

Figura 2.17. Formas de representação visual na integração GVis/KDD [Mac 1999].

As três formas de representação da Figura 2.17 ilustram o potencial do uso integrado de diversas técnicas de visualização. No caso, foram estabelecidas classes, mapeadas para cor, pela combinação de variáveis climáticas. A perspectiva multidimensional, isto é, a visão dos dados sob vários ângulos, possibilita que os aspectos espacial e temporal, e a influência dos atributos nas classes sejam observados.

Para a manipulação das formas de representação, de forma independente ou conjunta, foram implementadas as seguintes operações de interação:

- **Varredura (Brushing)**. Permite ao usuário selecionar e destacar determinados itens de dados, como agrupamentos de interesse, para uma análise mais detalhada;
- **Focalização (Focusing)**. Possibilita que o usuário restrinja a análise a determinados intervalos de valores de um atributo;
- **Manipulação de mapa de cores (Colormap Manipulation)**. Possibilita que o usuário configure esquemas de mapeamento de cores de acordo com as suas preferências e objetivos da análise.

Nesta Subseção foi exemplificado o potencial de uso de diferentes representações visuais para a exploração e análise de dados. Diversos ambientes genéricos de visualização de informação oferecem múltiplas técnicas e permitem o seu uso coordenado para visualizar um mesmo conjunto de dados [Swa 2001][War 2002]. Entretanto, em geral o conjunto de técnicas e as formas de coordenação disponíveis são fixos, pré-estabelecidos no projeto da ferramenta, o que limita a potencialidade de uso das técnicas em contextos de aplicação diversos. Ambientes que permitem flexibilidade na configuração da coordenação entre múltiplas visualizações têm sido objeto de investigação, e três iniciativas nessa direção são descritas na próxima Seção.

2.3 SOLUÇÕES COM COORDENAÇÃO FLEXÍVEL ENTRE VISUALIZAÇÕES

Já foi discutida a relevância de múltiplas visualizações coordenadas como um instrumento de apoio à exploração e análise de dados. Nesta Seção são descritos ambientes e arquiteturas concebidos e implementados para viabilizar o uso simultâneo de múltiplas visualizações coordenadas, e a configuração flexível das coordenações possíveis. Em virtude desses ambientes terem sido implementados sobre a plataforma Java, que é complexa e dotada de diversos recursos, termos específicos relativos à plataforma serão usados sem, contudo, apresentar detalhes técnicos.

2.3.1 SNAP-TOGETHER VISUALIZATION

Snap-Together Visualization (STV), ou simplesmente *Snap*, é uma solução que viabiliza a coordenação entre múltiplas ferramentas de visualização composta por um modelo de coordenação, uma interface com o usuário e uma arquitetura de sistema. O ambiente implementado viabiliza, de forma genérica, o uso coordenado de múltiplas ferramentas de visualização, permitindo ao usuário criar cenários personalizados integrando ferramentas independentes. A meta é oferecer ao usuário flexibilidade para acesso a diferentes conjuntos de dados, para a escolha de visualizações adequadas aos dados, e para a configuração dinâmica de coordenações, de acordo com as tarefas a serem executadas. Ferramentas de visualização são integradas ao ambiente por uma API, e interagem por meio de mecanismos de comunicação entre processos [Nor 1999]. Os dados visualizados estão armazenados em tabelas ou são provenientes de consultas a bases de dados relacionais.

O modelo conceitual do *Snap* para coordenação de visualizações é baseado no esquema de dados relacional. As visualizações disponibilizadas no STV exibem os atributos associados às tuplas armazenadas em tabelas ou resultantes de consultas a um banco de dados relacional, mapeando cada tabela/consulta para uma representação visual escolhida pelo usuário. A coordenação entre duas representações visuais também é configurada pelo usuário, e utiliza os relacionamentos entre as tabelas, os quais são definidos por chaves primária e estrangeira.

A Figura 2.18 mostra a equivalência entre os elementos do esquema de dados relacional e os elementos do modelo de coordenação do *Snap*, que varia de acordo com a cardinalidade dos relacionamentos entre as tabelas – um-para-um, ou um-para-muitos. A tabela *Diretórios*, que contém informações sobre os diretórios de um sistema de arquivos, tem uma relação de um-para-muitos com a tabela *Arquivos*, que contém os arquivos de cada diretório. Na interface visual, cada tabela é representada por uma visualização, e a coordenação estabelecida permite que, tendo sido selecionado um diretório em uma visualização, seu conteúdo seja carregado e exibido na outra visualização. A coordenação entre duas visualizações é comutativa, e uma sequência de coordenações tem a propriedade transitiva [Nor 2000].

Figura 2.18. Coordenação no *Snap* pelo esquema relacional. Adaptado de [Nor 2000].

A interface do *Snap* dá ao usuário condições de criar uma solução específica que atenda os requisitos de exploração e análise de seus dados, definindo um cenário personalizado de múltiplas visualizações coordenadas, sem necessidade de programação. A Figura 2.19 ilustra o processo de configurar uma coordenação, que envolve múltiplas interfaces. Uma visualização é ativada, entre um conjunto de técnicas disponíveis, escolhendo-se, em uma lista de tabelas/consultas, a entidade a ser visualizada, e arrastando-a para o botão correspondente à representação visual desejada (Figura 2.19-a). A interface de cada visualização possui um botão

“snap”, que é arrastado e solto sobre o botão correspondente da outra visualização para estabelecer uma coordenação (Figura 2.19-b) entre ambas; as ações de interação que podem ser coordenadas são estabelecidas para cada técnica visualização (Figura 2.19-c) [Nor 2000b].

Figura 2.19. Passos no *Snap* para coordenação de visualizações. Adaptado de [Nor 2000].

Na arquitetura do STV, a intermediação entre as ferramentas de visualização, bem como o acesso à base de dados, são feitos por um elemento central (retângulo identificado por ‘*Snap-Together Visualization*’ na Figura 2.20), que torna os dados disponíveis para a representação visual escolhida. A Figura 2.20 também ilustra uma situação na qual três visualizações – *Treemap Visualization* (esquerda), *Outliner Visualization* (centro), e *Grid Visualization* (direita) – são alimentadas a partir de uma base relacional contendo informações sobre uma estrutura de diretórios e arquivos. Uma coordenação entre as três visualizações pode ser definida como ilustrado na figura: a seleção de um particular diretório no *Treemap* provoca a seleção do mesmo em *Outliner*, que por sua vez, provoca uma consulta à base para buscar os detalhes do diretório e apresentar em *Grid* [Nor 1999].

Figura 2.20. Funcionamento de uma coordenação no *Snap* [Nor 1999].

A versão inicial do *Snap*, descrita por North [Nor 2000], foi implementada na plataforma Windows com a linguagem de programação Visual Basic; a comunicação empregou o modelo Microsoft COM/Active X; e o acesso a base de dados – Access ou Oracle – foi feito via ODBC. Para ser incorporada a essa arquitetura, uma ferramenta de visualização deve implementar uma API que tem definidas as seguintes formas de comunicação: (1) *Snap* envia dados para a ferramenta via memória, arquivo ou ODBC, (2) a ferramenta informa o *Snap* sobre eventos decorrentes de ações dos usuários, (3) *Snap* solicita a uma ferramenta a realização de uma determinada ação.

A estratégia de coordenação de visualizações segue o modelo apresentado por North [Nor 2001], no qual o esquema de múltiplas visualizações coordenadas é formalizado pelos seguintes elementos:

- A definição de uma visualização individual como uma representação visual (um diagrama de dispersão, por exemplo) que exhibe itens de dados de um conjunto. Uma coordenação entre duas visualizações estabelece uma relação entre os itens de dados exibidos em duas visualizações distintas;
- Uma linguagem para especificar a coordenação entre duas visualizações, com as seguintes sintaxe e semântica:
 - o ((visA,açãoA,itensA)(visB,açãoB,itensB)) : Uma ação do tipo “açãoA” em uma visualização “visA” envolvendo os itens “itensA” provoca uma ação correspondente “açãoB” na visualização “visB” envolvendo os itens “itensB”. O

caminho inverso de “visB” para “visA” também é válido, pois a coordenação tem propriedade comutativa;

- Uma taxonomia para os tipos de ações:
 - o Seleção. O usuário seleciona itens de dados de interesse, que são destacados. Opcionalmente, para ressaltar os itens selecionados, os demais podem ser apagados total ou parcialmente;
 - o Navegação. O usuário navega em uma visualização para focar em itens de interesse ou exibir esses itens. Exemplos de ações de navegação são: abrir arquivo, carregar dados, ampliar/reduzir, deslizar, particionar, rotacionar, e outras;

North e Shneiderman [Nor 2000a] ressaltam a importância e a necessidade de testar a estratégia de múltiplas visualizações coordenadas junto aos usuários. Em testes realizados com o *Snap*, foi observado que a construção de cenários com múltiplas visualizações coordenadas requer que os usuários compreendam o modelo de coordenação. Entretanto, a interface da versão inicial (ver Figura 2.19) não deixa esse modelo muito visível, demandando um certo tempo para a familiarização. Em versões posteriores, a limitação foi tratada com uma interface que proporciona maior visibilidade do modelo de coordenação, explicada a seguir.

A versão inicial do Snap [Nor 2000] foi aperfeiçoada por Conklin [Con 2002], buscando resolver algumas limitações:

- Necessidade de codificação e recompilação do ambiente quando da incorporação de novas visualizações; exigência da instalação de todas as ferramentas de visualização antes do uso do ambiente, dificultando a sua disponibilização para potenciais usuários;
- Ausência de suporte para seleção múltipla, impossibilitando a atuação simultânea sobre várias tuplas;
- Ausência de uma interface visual para a construção de coordenações e representação gráfica dos elementos envolvidos (tabelas/visualizações e relacionamentos/coordenações), o que ajudaria o usuário a compreender o modelo de coordenação.

A nova solução também adota o modelo de banco de dados relacional para definir a coordenação de visualizações, entretanto, o elemento central, denominado *Snap Visualization Server*, é implementado em Java e acessado via web [Nor 2002a]. Um cliente, por meio de um navegador (ver exemplo na Figura 2.21), utiliza o *Snap Control Panel*, um *applet* Java, para

conectar com bases de dados e exibir/manipular um ‘Esquema de Visualização’ (*Visualization Schema*). Esse esquema é análogo ao esquema de dados para visualizações e coordenações e é representado por um grafo. Isso torna o modelo de coordenação mais visível e permite a manipulação direta de seus elementos (nós = visualizações e arcos = coordenações). A instanciação de uma visualização, que faz um nó ser acrescentado ao grafo no esquema de visualização, é feita no ‘Espaço de Visualização’ (*Visualization Workspace*). Nesse espaço, as visualizações são organizadas pela divisão horizontal ou vertical de espaço [Nor 2002].

Figura 2.21. Interface web de acesso ao *Snap Visualization Server* [Con 2002].

Três camadas compõem a arquitetura da versão web, as quais podem ser observadas no diagrama da Figura 2.22-a. O acesso à base de dados é via JDBC, além de ODBC, com possibilidade de conexão com bases de dados remotas (diagrama 2.22-d). A medida em que o usuário constrói o esquema de visualização é gerado um grafo de coordenação (diagrama 2.22-c), que permite gerenciar a coordenação entre as visualizações. Componentes de visualização, em conformidade com a API do *Snap*, podem ser adicionados ao sistema e disponibilizados aos usuários, e se comunicam com o gerenciador de coordenações por meio de “Adapters” (diagrama 2.22-b), o que permite incorporar componentes implementados com diferentes tecnologias [Nor 2002].

Figura 2.22. Camadas da arquitetura do *Snap Visualization Server*. Adaptado de [Nor 2002].

Uma extensão que incorpora algoritmos de mineração ao modelo do *Snap* integra o modelo *Fusion* [Nor 2003], que mantém o modelo relacional de dados, a interface com esquemas de visualização e a construção de espaços de visualização. O uso e a evolução do *Snap/Fusion* mostra que a estratégia de múltiplas visualizações coordenadas é muito poderosa para a exploração e análise visual de dados. Além disso, evidencia a variedade de parâmetros envolvidos no uso e implementação de múltiplas visualizações coordenadas e, principalmente, que uma solução de aplicabilidade prática requer a integração de instrumentos visuais, analíticos e de armazenamento/recuperação de dados.

2.3.2 PROJETO SPIN!

O SPIN! – *Spatial Mining for Data of Public Interest* – é um projeto coordenado pelo Fraunhofer AIS, *Autonomous Intelligent Systems* (anteriormente, GMD, *German National Research Center for Information Technology*) – com vários parceiros europeus, cujo objetivo é desenvolver uma plataforma para mineração de dados espaciais, baseada na *web*, integrando funcionalidades de um sistema de informações geográficas (SIG) e de algoritmos de mineração de dados. Além dessas funcionalidades, o sistema deve garantir escalabilidade e independência de plataforma [May 2002].

Um dos componentes do SPIN! é o Descartes, um sistema para visualização exploratória de dados estatísticos referenciados espacialmente – por exemplo, dados sócio-econômicos – que oferece recursos visuais tipicamente presentes em SIGs. Esse sistema possui um conjunto de heurísticas sobre cartografia temática que permite gerar automaticamente mapas temáticos para a apresentação de dados estatísticos. Os mapas são escolhidos pelo sistema considerando a caracterização dos dados, e podem ser manipulados interativamente, e as modificações são refletidas dinamicamente. Outras representações visuais estão disponíveis, como gráficos de dispersão e coordenadas paralelas, e podem ser coordenadas entre si e com os mapas. Operações de ampliação/redução (*zooming*), deslocamento (*panning*) e ativação/desativação de camadas de informação geográfica, comuns em SIGs, podem ser realizadas. O servidor desse sistema, implementado em C++ e disponível para Unix e Windows, é acessado por clientes *applets* Java, e sua arquitetura é esquematizada na Figura 2.23. Um problema potencial dessa configuração é que alguns *firewalls* podem barrar a comunicação por *sockets* [And 2000][And 1999a], o que foi resolvido com o direcionamento do projeto SPIN! para o uso do CommonGIS⁷, em que a comunicação ocorre via protocolo HTTP.

A interface do SPIN!, mostrada na Figura 2.24, disponibiliza componentes de acesso a base de dados, de algoritmos de mineração e de visualização, e aqueles selecionados pelo usuário para uma tarefa de análise formam um ‘Espaço de Trabalho’ (*workspace*), que é apresentado em forma de árvore (*workspace tree*) e em forma de grafo (*workspace tree* e *point and click tool for defining analysis tasks*, respectivamente, na Figura 2.24). Cada componente tem suas

⁷ www.commongis.de

propriedades (conjunto de parâmetros para seu funcionamento) assinaladas por meio de um ‘Editor de Propriedades’ (*property editor* na Figura 2.24).

Figura 2.23. Funcionamento do Descartes [And 1999a].

Uma análise interativa pode utilizar múltiplas visualizações coordenadas, ilustrada na Figura 2.25, e a coordenação (representada pelas flechas no grafo da Figura 2.24) é feita por meio de identificadores dos itens de dados e de propagação de eventos dentro de um espaço de trabalho [May 2002a].

Figura 2.24. Elementos da interface do SPIN! [May 2002a].

Figura 2.25. Visualizações resultantes de uma configuração de espaço de trabalho no SPIN! [May 2002a].

A arquitetura do SPIN! é baseada no *Enterprise Java Beans* (EJB), e configurada em esquema cliente/servidor com múltiplas camadas (*n-tier client/server architecture*), cujos componentes são apresentados na Figura 2.26. O cliente '*Client*' é uma interface gráfica, que contém o espaço de trabalho com a configuração definida pelo usuário, implementada como uma aplicação ou *applet* Java, que se comunica com o servidor por meio de RMI, CORBA, ou utiliza o protocolo HTTP para acessar *servlets*. Outros subsistemas são o servidor de aplicações (*application server*) e o servidor de banco de dados (*database server*) [May 2002].

Figura 2.26. Elementos da arquitetura do SPIN! [May 2002a].

Integrando funcionalidades de SIG com algoritmos de mineração de dados espaciais, o projeto SPIN! pode oferecer maneiras alternativas de apoiar tarefas de exploração e análise de dados geo-referenciados, tais como buscar explicações para o agrupamento espacial de doenças,

ou calcular o impacto no clima devido à mudança no padrão de uso ou ocupação da terra, ou ainda estabelecer uma segmentação espacial de clientes de uma empresa. O projeto também define uma API que viabiliza a incorporação de novos componentes ao ambiente.

2.3.3 GEOVISTA STUDIO

O *GeoVISTA Center*, da *Pensylvania State University*, desenvolve pesquisas em Geo-computação, que trata do desenvolvimento e uso de ferramentas e técnicas computacionais para o tratamento de dados geográficos. No centro está sendo desenvolvido o *GeoVISTA Studio*, um ambiente de programação visual que integra recursos de processamento analítico e de visualização para apoiar atividades de análise. O *Studio* deve oferecer diversas funcionalidades, apoiando de forma integrada as diferentes atividades de análise de dados geográficos, desde a formulação de hipóteses, passando por tarefas de classificação, até a construção de modelos [Gah 2000][Tak 2002].

A implementação do *Studio* utiliza recursos da plataforma Java. Programas são criados visualmente pela conexão de módulos independentes disponibilizados em bibliotecas de componentes *JavaBeans*, seguindo o modelo de fluxo de dados, dispensando a escrita de código fonte. Veja o diagrama na Figura 2.27-b. Os módulos disponíveis permitem implementar diversas funcionalidades, tais como recursos de visualização, tratamento estatístico e técnicas de mineração. A Figura 2.27-a mostra que componentes *JavaBeans* podem ser incorporados ao *Studio*. Os requisitos desejados do ambiente incluem uma arquitetura não proprietária e recursos para a integração de componentes e de programação [Tak 2002a]. O componente ‘*coordinator bean*’ atua como um intermediário entre componentes cujos comportamentos devem ser coordenados, isto é, estabelece um forte acoplamento entre componentes, provendo uma funcionalidade adicional ao ambiente, além do ‘*pipeline*’ do esquema de fluxo de dados [Mac 2001].

Figura 2.27. Arquitetura e programação visual no *Studio* [Mac 2001].

Componentes são organizados e apresentados ao usuário, no painel principal (Figura 2.28), agrupados, em guias individuais, segundo suas funcionalidades. O ‘Quadro de Projeto’ (DesignBox, na Figura 2.29) permite escolher os componentes e estabelecer as conexões, construindo um ambiente cuja organização pode ser observada visualmente durante a construção, no ‘Quadro de Interface Gráfica’ (GUIBox), conforme ilustra a Figura 2.29. O GeoVISTA *Studio* pode ser acessado por meio do *Java Web Start* [Mac 2001].

Figura 2.28. Organização de componentes no painel principal do *Studio* [Tak 2002a].

O *Studio* atende três tipos de usuários: (1) o desenvolvedor de ferramentas, que pode criar novos módulos individuais e agregá-los ao ambiente; (2) o criador de aplicativos, que cria uma aplicação específica utilizando os módulos disponíveis; e (3) o usuário de aplicativos, que utiliza uma aplicação já criada e que atenda suas necessidades. Uma vez construído e testado um ambiente, é possível disponibilizá-lo na forma de *applets*, aplicativos ou componentes *JavaBeans*, o que torna o *Studio* uma ferramenta para a criação de soluções voltadas para usuários/tarefas específicos, sem necessidade de escrita de código.

Figura 2.29. Criação visual de um projeto no *Studio* [Tak 2002a].

2.3.4 CONSIDERAÇÕES GERAIS SOBRE AS SOLUÇÕES

Os ambientes descritos nesta Seção oferecem flexibilidade na definição da coordenação entre múltiplas visualizações e garantem portabilidade entre diferentes sistemas operacionais e configurações de hardware. Outro aspecto importante é a possibilidade de (re)aproveitamento de módulos já existentes, diminuindo o esforço de desenvolvimento e permitindo a evolução gradativa dos ambientes.

Embora algumas iniciativas tenham inicialmente adotado outras soluções, a opção pela plataforma Java é comum a todas. Utilizando diferentes mecanismos de implementação, cada ambiente oferece recursos próprios aos seus usuários. A versão estendida do *Snap* adota um servidor *web* cujos recursos de visualização são acessados por *applets*, e uma interface que torna visível ao usuário o modelo de coordenação de visualizações. *SPIN!* e *GeoVista Studio* enfatizam o tratamento de dados espaciais e adotam a tecnologia de *JavaBeans*, permitindo a incorporação gradual de novos componentes. No ambiente *SPIN!*, funcionalidades de SIG e algoritmos de mineração são integradas, e a interface com o usuário, que pode ser um aplicativo Java ou um *applet*, oferece acesso a mensagens reportadas pelos módulos, constituindo uma forma de registro e documentação de atividades. No *GeoVista Studio*, um componente *JavaBeans* coordenador arbitra a comunicação entre componentes, apoiando a coordenação entre representações visuais, e o esquema de fluxo de dados (*dataflow*) é adotado para permitir a

criação de um cenário para uma dada tarefa de análise, o qual que pode ser disponibilizado posteriormente como um aplicativo Java, um *applet*, ou um componente *JavaBeans*.

Os ambientes descritos convergem para o objetivo de possibilitar a criação de cenários personalizados para resolver questões de exploração e análise de dados, atendendo aos requisitos específicos de usuários e/ou tarefas. A personalização de soluções é viabilizada pela flexibilidade oferecida na configuração da coordenação entre visualizações, e para a integração de componentes analíticos e de visualização escolhidos segundo as necessidades do usuário. Completando os pontos comuns, cada ambiente apresenta características próprias, que são importantes no contexto de múltiplas visualizações coordenadas:

- **Visibilidade de modelo.** A interface do *Snap* deixa o modelo de coordenação visível;
- **Registro de atividade.** A interface do SPIN! dá acesso a mensagens de alguns componentes;
- **Formas de disponibilização de um cenário.** O *Studio* tem mecanismos para gerar aplicativos, *applets* ou *beans*;

2.4 CONSIDERAÇÕES FINAIS

Neste Capítulo, foi estabelecida a relação entre Visualização, Descoberta de Conhecimento e Mineração de Dados, e apresentado um panorama das técnicas de Visualização de Informação. Foi abordado, também, o uso de técnicas de visualização em conjunto de técnicas de interação para a exploração e análise de dados, com ênfase em soluções que empregam a estratégia de múltiplas visualizações coordenadas.

Foi apresentada uma iniciativa de formalizar o processo de coordenação entre visualizações, no escopo de desenvolvimento do *Snap*, em que foram propostos um modelo de coordenação baseado no esquema relacional, uma linguagem de especificação e uma taxonomia de ações para a coordenação. A experiência acumulada na concepção e desenvolvimento dos ambientes descritos, que incorporam as experiências de estudos anteriores (não necessariamente restritos a múltiplas visualizações coordenadas), evidencia a necessidade de um modelo para o desenvolvimento de plataformas com múltiplas visualizações coordenadas.

No Capítulo 3 é descrito o modelo *AdaptaVis*, proposto nesta tese, que tem por objetivo definir uma arquitetura de software independente de tecnologia e um conjunto de diretrizes para o desenvolvimento de ambientes adaptáveis às necessidades de usuários/tarefas no contexto de exploração e análise de dados usando a estratégia de múltiplas visualizações coordenadas. Ainda, no mesmo capítulo, são introduzidas duas representações visuais para dados com atributos temporais propostas no âmbito deste trabalho. O modelo proposto e as representações visuais temporais foram instanciados em uma plataforma de visualização, e sessões de uso foram realizadas em domínios de dados com atributos espaciais e temporais, em conjunto com um componente de visualização para atributos espaciais (mapa) e uma tabela para a exibição de dados em formato textual.

DEFINIÇÃO DE REPRESENTAÇÕES VISUAIS TEMPORAIS E CONCEPÇÃO DO MODELO *Adapta Vis*

Recursos de visualização interativa, pela capacidade de ampliação do processo cognitivo, são potencialmente úteis em diversas etapas do processo de descoberta de conhecimento [Kim 2004] e sua aplicabilidade abrange domínios variados. O uso integrado e coordenado de múltiplas representações visuais também traz benefícios, aumentando esse potencial [Car 1999][Kos 2003]. Fatores como o tipo dos dados, a especificidade do domínio, as características do usuário e as tarefas realizadas determinam as representações visuais e as formas de coordenação adequadas.

Dispor de uma plataforma de software única que atenda os requisitos de diferentes aplicações apresenta vantagens, tais como o (re)uso de técnicas já implementadas e a redução do tempo de treinamento e de desenvolvimento. É vantajosa, também, a possibilidade de incorporar novas técnicas com baixo impacto sobre os componentes já implementados [Nor 2001][May 2002][Tak 2002]. Tal contexto é fundamental para viabilizar e encorajar o uso das técnicas em múltiplos domínios de aplicação e por diversos tipos de usuário, com diferentes dados e tarefas, possibilitando avaliações, amadurecimento, contato entre usuários e a generalização dos componentes implementados e a criação de novos [Oli 2004]. Neste Capítulo são propostas duas novas técnicas para a visualização exploratória de dados temporais, bem como um modelo que permite a sua integração em uma plataforma de visualização flexível e extensível.

As técnicas propostas podem ser enquadradas como orientadas a pixel (ver Subseção 2.2.1), procurando aproveitar algumas vantagens proporcionadas por essa categoria de técnicas: (1) apresentação de um grande volume de dados, pela maximização do uso da área do dispositivo de exibição; (2) bom desempenho interativo devido à facilidade de implementação, pois não requerem renderizações complexas; e (3) são intuitivas e de fácil entendimento, principalmente pela ordenação natural associada a atributos temporais. Elas foram posteriormente utilizadas com as representações de mapa, que transmite o componente espacial da informação, e tabela, como detalhado no Capítulo 5. Manco, Pizzuti e Talia [Man 2004] utilizam, no Eureka!, uma ferramenta de mineração visual de dados, o mapeamento de cor para pixels ou pontos, maximizando o uso do espaço da interface para a exibição de vários atributos.

Foi mencionado, no Capítulo 2, que representações visuais de dados podem ser utilizadas de forma estática, dinâmica isolada ou dinâmica coordenada. A primeira forma de uso é adequada à apresentação de descobertas já feitas, permitindo, no caso de uma visualização bem projetada, sumarizar a informação de maneira intuitiva e de fácil percepção. Nas outras duas formas, recursos de interação permitem ao usuário atuar sobre as representações visuais manipulando-as diretamente, e observando o resultado imediatamente. A interação pode afetar tão somente a representação visual diretamente manipulada (forma dinâmica isolada), ou a ação sobre uma representação visual pode ser propagada e refletida em outras representações semanticamente relacionadas (forma dinâmica coordenada). Para estabelecer uma coordenação entre duas ou mais representações visuais é estabelecido um ‘forte acoplamento’ [Nor 1997] entre elas, um processo de interação também denominado *linking* [Cle 1993]. Basicamente, essas formas resumem a evolução do uso de representações visuais em processos de análise de dados.

Um modelo formal, como o *Snap* [Nor 2001] para a coordenação entre visualizações, bem como um conjunto de recomendações sobre quando e como utilizar múltiplas visualizações coordenadas [Bal 2000][Bro 2003], são importantes para prover diretrizes ao desenvolvedor de tais plataformas. Boukhelifa, Roberts e Rodgers [Bou 2003] enfatizam que um modelo de coordenação é importante para definir, testar e comparar estratégias de coordenação, e avaliar sistemas que incorporam coordenação; eles, assim como Pattison e Philips [Pat 2001], acreditam que aspectos mais gerais, como a movimentação de janelas, podem ser coordenados.

O atual estágio de desenvolvimento da área indica a necessidade e a viabilidade de um modelo de arquitetura de software para uma plataforma de visualização adaptável, a partir da qual seja possível criar ambientes personalizados. Tal arquitetura pode garantir liberdade e flexibilidade de escolha das representações visuais e de coordenação das mesmas [Jer 2003]. Kimani et al. [Kim 2004] destacam a necessidade e utilidade de um modelo, um arcabouço (*framework*) aberto e extensível para um sistema de mineração visual de dados. A independência de tecnologia é uma característica importante no modelo que suporta tal plataforma, conferindo liberdade ao desenvolvedor quanto à forma de implementação. Com base na evolução do uso de representações visuais e no conhecimento das soluções descritas na Seção 2.3, identifica-se um conjunto de características que reflete pontos de convergência nas decisões de projeto de ambientes que integram múltiplas técnicas de visualização operando de maneira coordenada. O modelo *AdaptaVis*, aqui introduzido, identifica os elementos essenciais, suas funcionalidades e conexões que garantem a adaptabilidade e flexibilidade da plataforma.

Nas próximas seções, são descritas as representações visuais temporais propostas, denominadas ‘Variação Temporal Uni-escala’ e ‘Variação Temporal Multi-escala’, e o modelo *AdaptaVis*, que caracterizam as principais contribuições deste projeto. Na Seção 3.1, são apresentados os fundamentos na definição das representações visuais temporais. Em seguida, na Seção 3.2, a concepção do modelo *AdaptaVis* é detalhada.

3.1 REPRESENTAÇÕES VISUAIS PARA ATRIBUTOS TEMPORAIS

A concepção das visualizações temporais foi inspirada nos resultados de alguns estudos iniciais, descritos no Apêndice B, realizados com dados de precipitação pluviométrica, que contaram com a participação de um especialista no domínio dos dados. Os dados descrevem medidas diárias de precipitação pluviométrica, obtidas ao longo de vários anos (elemento temporal) em estações de coleta definidas por suas coordenadas geográficas [Shi 2003] (as estações constituem, portanto, objetos espaciais), conforme o esquema ilustrado na Figura 3.1.

Figura 3.1. Objeto espacial com várias medidas de um atributo ao longo do tempo.

Os estudos iniciais permitiram identificar a necessidade de apoiar as seguintes tarefas genéricas: (1) análise de relações decorrentes do posicionamento espacial relativo entre os elementos representados pelos itens de dados; e (2) análise do comportamento temporal de um atributo dos dados, e de sua possível relação com o posicionamento espacial [Shi 2003]. Na plataforma *InfoVis*, que constitui uma instanciização do modelo *AdaptaVis* descrita no Capítulo 4, foram incluídas as seguintes representações visuais: uma tabela que exibe os dados em formato textual, um mapa que apresenta os itens de dados espacialmente, e as duas representações aqui definidas, que exibem o comportamento temporal de um determinado atributo. A tabela e o mapa são formas tradicionais de apresentação de texto e localização espacial, respectivamente. As duas técnicas para visualização de atributos com variação temporal foram concebidas no âmbito deste projeto considerando os requisitos das tarefas acima. Observa-se que, em muitas situações, essas tarefas são realizadas de forma simultânea ou intercalada.

As técnicas propostas são genéricas, e podem apoiar as tarefas (1) e (2) acima para dados e aplicações em diferentes domínios. Concebidas inicialmente para apoiar tarefas de exploração de um conjunto de dados de precipitação pluviométrica, algumas das técnicas foram empregadas para explorar dados de transações comerciais que também apresentam atributos temporais e espaciais. A opção pela estratégia orientada a pixels permite o mapeamento de atributos dos itens de dado para cor, saturação e brilho, sendo possível, teoricamente, exibir até três atributos simultaneamente em cada pixel.

As visualizações temporais propostas mapeiam os valores de um único atributo dos dados para a cor dos pixels, os quais são organizados segundo a ordem cronológica natural dos

itens de dados que representam. A escala de tempo adotada pode variar de acordo com os requisitos de usuários e tarefas. Além dos valores do atributo mapeado para cor, as visualizações devem ser capazes de exibir informações adicionais [Shi 2004]:

- **Intervalo temporal**, mostrando o tempo inicial e final da presença de um atributo, delimitando sua duração;
- **Interrupção no intervalo**, caso existam interrupções na presença de valores do atributo entre o tempo inicial e final;
- **Posicionamento cronológico**, i.e., intervalos de igual duração podem ter tempos inicial e final diferentes, o que pode constituir uma informação importante;
- **Ciclos temporais**, i.e., a existência de repetições ou padrões que podem ser explicadas pelo atributo temporal.

As visualizações concebidas com as características acima são detalhadas a seguir.

3.1.1 VARIAÇÃO TEMPORAL UNI-ESCALA

Essa representação é adequada para exibir a variação temporal de um determinado atributo, segundo uma escala de tempo (dia, ano, hora, trimestre,...), e uma tabela de cores definida pelo usuário que mapeia intervalos de valores de interesse em cores. Os valores de um atributo temporal são representados por uma sequência de pixels, ou pontos, esquematizados como quadrados na Figura 3.2-a.

Figura 3.2. Construção da representação visual 'Variação Temporal Uni-escala'.

Cada sequência de pixels é associada a um objeto espacial, e pode ser disposta horizontal ou verticalmente, sobre um eixo temporal de referência, compondo uma linha ou coluna, respectivamente. A justaposição de várias linhas (Figura 3.2-b) ou colunas (Figura 3.2-c) cria uma representação visual que permite a observação simultânea de informações relativas a vários objetos espaciais. Considerando que cada ponto da linha/coluna é posicionado sobre o eixo temporal definido em uma escala de tempo, informações adicionais (intervalo temporal, interrupção no intervalo e ciclos temporais) podem ser percebidas. A imagem gerada lembra as imagens produzidas pela técnica Table Lens [Rao 1994], mas a representação aqui descrita exibe um único atributo, e sua construção e interpretação são diferentes. Em sua implementação, essa representação visual recebeu as denominações ‘Intervalo em Anos’ e ‘Intervalo/Categoria’, e os detalhes de sua interface e suas funcionalidades estão descritas na Tabela 4.3 do próximo capítulo.

Essa visualização foi implementada com base em estudos iniciais conduzidos com dados de precipitação pluviométrica, buscando uma forma visual efetiva de exibir a qualidade de dados de precipitação coletados, em uma escala mensal (ver Figuras B.3 e B.4 no Apêndice B).

3.1.2 VARIAÇÃO TEMPORAL MULTI-ESCALA

Durante os estudos iniciais, também foi utilizada uma visualização (ver Figura B.5) que permitia focar em um único objeto espacial (estação de coleta), observando a variação temporal de um atributo (no caso, qualidade do dado coletado) com maior grau de detalhe (escala diária). O formato da representação visual foi o mesmo adotado na representação uni-escala, com cada valor diário mapeado para cor e representado por um pixel, cuja sequência formava uma linha. Contudo, cada uma das linhas se referia a um determinado ano, e a justaposição delas indicava a variação do valor de interesse em determinada estação de coleta. A experiência com escalas de tempo variadas mostrou que a visualização em várias escalas simultaneamente seria útil.

Assim, a visualização multi-escala exibe a variação temporal de um determinado atributo simultaneamente em escalas temporais distintas. Os valores em um nível da escala podem ser gerados a partir da soma dos valores do nível anterior. Por exemplo, valores de totais anuais (de chuva, de transações, etc.) são calculados a partir de totais mensais que, por sua vez, são obtidos pela somatória de valores diários registrados. Para cada escala, o usuário deve definir uma tabela

de mapeamento de cor. É possível navegar pelos diferentes níveis da escala temporal, explorando os valores simultaneamente em diferentes graus de detalhamento.

A disposição dos dados nessa representação é mostrada na Figura 3.3-a, assumindo as escalas diária, mensal e anual. Os valores referentes a um determinado ano são apresentados em uma coluna, distribuídos em elementos retangulares. A escala diária é exibida nos doze retângulos superiores em cada coluna, sendo que cada retângulo consiste de uma grade 5x7 (suficiente para exibir 28, 29, 30 e 31 dias, conforme o mês e o ano) cujos elementos são preenchidos da esquerda para a direita, e de cima para baixo. O dia 1 fica no canto esquerdo superior e a distribuição dos demais dias lembra um calendário, como ilustrado na Figura 3.4. A escala mensal é composta pelos doze retângulos intermediários em cada coluna, e a escala anual pelo retângulo inferior. Essa nova disposição em retângulos proporciona uma nítida percepção das localizações das unidades de tempo, quando comparada à disposição linear utilizada nas visualizações nos estudos iniciais.

A justaposição das colunas, seguindo a ordem cronológica, produz uma representação visual esquematizada na Figura 3.3-b. A coloração de um retângulo da escala diária (ver Figura 3.4) pode levar, pela disposição das cores, a uma figura que pode ser interpretada como um ícone, o que pode auxiliar tarefas de busca por padrões. Informações adicionais, como intervalo temporal (quantidade de colunas) e interrupção (áreas sem coloração) também são perceptíveis.

Figura 3.3. Elementos de construção de uma representação visual ‘Variação Temporal Multi-escala’.

Figura 3.4. Divisão do retângulo para representação da escala diária.

Wijk e Selow [Wij 1999] empregaram um esquema semelhante para exibir os agrupamentos de padrão diário (única escala) de consumo de energia gerados a partir de uma série temporal, pintando os retângulos de um calendário. A coloração de cada retângulo é determinada pelo agrupamento ao qual o dia do mês pertence. Essa representação permitiu observar a existência de concentração de padrões, mostrando comportamentos de consumo em períodos diversificados (quinzenas, finais de semana, ...).

Uma representação visual pode exibir a variação temporal de um determinado atributo em duas situações: (1) fixado um objeto espacial (por exemplo, uma estação de coleta), variar os períodos no tempo (ano), como ilustrado na Figura 3.5-a; ou (2) fixar um período da escala de tempo, um dado ano, por exemplo, e variar nas colunas os objetos espaciais, cidades, por exemplo, como ilustrado na Figura 3.5-b. Nessa última situação, a finalidade é verificar o comportamento temporal de um atributo (um ano fixo, por exemplo), em relação a vários objetos espaciais. O eixo que representa os níveis de agregação permanece inalterado em ambas as situações.

Figura 3.5. Opções de organização da representação visual 'Variação Temporal Multi-escala'.

A implementação segundo a organização dada pela Figura 3.5-a recebeu a denominação ‘Dados de Chuva’, cujos detalhes também estão na Tabela 4.3 do próximo capítulo, e foi utilizada nas sessões de uso com dados pluviométricos. Essa implementação foi adaptada para refletir a organização dada pela Figura 3.5-b, e uso no segundo domínio de aplicação, com dados de transações comerciais.

3.2 O MODELO *ADAPTA VIS*

O *AdaptaVis* é um modelo de arquitetura de software cuja finalidade é estabelecer um arcabouço para nortear o desenvolvimento e a implementação de uma plataforma computacional extensível, a partir da qual ambientes personalizados de apoio ao processo de descoberta de conhecimento, integrados por múltiplas visualizações coordenadas, possam ser usados. A plataforma deve oferecer recursos para atender os requisitos de diferentes aplicações, permitindo criar novos ambientes a partir da escolha do leque de representações visuais julgadas adequadas pelo usuário, bem como do conjunto de ações coordenadas entre as representações. A possibilidade de utilizar múltiplas técnicas de forma coordenada, bem como a facilidade para incorporar novos recursos de visualização, são pontos comuns aos ambientes descritos na Seção 2.3, apesar das diferentes soluções de implementação. O modelo *AdaptaVis* busca identificar e definir os elementos mínimos para viabilizar ambientes com as características descritas, sem contudo, restringir ou sugerir uma forma de implementação, ou a tecnologia a ser utilizada. Por exemplo, deve ser conferida flexibilidade para decidir por uma implementação distribuída ou baseada na *Web*.

Outro aspecto importante da flexibilidade na criação de ambientes personalizados é a possibilidade de ampliar o universo de potenciais usuários e acompanhar o uso que estes fazem dos recursos, viabilizando a avaliação do uso de visualização exploratória em diferentes contextos. É muito importante capturar informações relativas ao uso para a melhoria contínua dos recursos visuais oferecidos para apoiar processos de descoberta de conhecimento. Graham, Kennedy e Benyon [Gra 2000] recomendam a manutenção de um histórico de atividades (*log*) realizadas no ambiente, em conjunto com anotações gerais e telas capturadas, como uma ferramenta eficiente para a captura de detalhes das tarefas executadas em ambientes dessa natureza, permitindo verificar o comportamento e o modelo mental [Spe 2001] do usuário durante suas atividades. O histórico de atividades é uma das tarefas identificadas por

Shneiderman [Shn 1996], em adição às aquelas do ‘*VIS mantra*’. Além da anotação de observações dos usuários, a avaliação feita por especialistas constitui uma importante fonte de informação [Kos 2003]. O registro das atividades realizadas e das combinações experimentadas é um instrumento poderoso para o acompanhamento, planejamento, avaliação e documentação das ocorrências nos ambientes gerados. Um mecanismo para o registro de atividades realizadas para a configuração de ambientes e durante o seu uso foi incluído no *AdaptaVis*.

Com base no exposto, extrai-se que a adaptabilidade da plataforma é garantida pela capacidade de atender as necessidades específicas de usuários de diferentes domínios, permitindo a escolha de representações visuais adequadas e oferecendo mecanismos para estabelecer, em tempo de execução, coordenações entre ações executadas sobre essas representações. O registro de atividades também contribui para o item adaptabilidade, na medida em que oferece condições para um refinamento sucessivo das funcionalidades da plataforma. No contexto do *AdaptaVis*, o termo plataforma se refere ao conjunto completo de componentes e as respectivas ações de interação, e o termo ambiente se refere a um subconjunto configurado para atender às necessidades específicas de um usuário e suas tarefas. Na implementação dessa plataforma adaptável, com reflexo nos ambientes gerados, os critérios a seguir devem ser observados [Oli 2004]:

- **Extensibilidade**. Cada técnica de visualização é implementada como um módulo independente e auto-suficiente, controlando seus parâmetros e a sua interação com o usuário, segundo o conceito de componente de software. Complementando a independência e a auto-suficiência, cada módulo deve ser capaz de trocar mensagens e interagir com outros módulos, aderindo a um protocolo padrão que implementa um conjunto de comportamentos comuns. As características de independência e auto-suficiência permitem que módulos existentes sejam atualizados com baixo impacto sobre a plataforma. A conformidade com um padrão garante um funcionamento uniforme dos ambientes gerados a partir da plataforma, e facilita a criação, inserção e teste de novos módulos, além de permitir criar diversos ambientes pela combinação de diferentes módulos, o que constitui um importante fator para o critério de configurabilidade, descrito a seguir;
- **Configurabilidade**. Diferentes tarefas, mesmo em uma única aplicação, podem demandar seqüências diversificadas de ações coordenadas entre diferentes representações visuais. Por isso, o usuário deve ter liberdade para coordenar diferentes visualizações em tempo de

execução, durante o uso do ambiente, segundo suas necessidades, preferência e experiência. Como complemento, deve ser possível definir vários ambientes independentes e distintos, cada um disponibilizando um conjunto de técnicas visuais estabelecido de acordo com as necessidades de uma aplicação específica. Uma personalização maior do ambiente pode ser alcançada pelo uso de termos do domínio de aplicação nas operações de manipulação das representações visuais. Como cada representação visual admite um certo conjunto de ações de interação, a composição do conjunto de tais representações define o conjunto de ações coordenadas que podem ser estabelecidas em um ambiente;

- **Monitorabilidade.** A facilidade de definir ambientes personalizados cria uma fonte de informação potencial para a documentação e avaliação de diferentes cenários de uso de múltiplas visualizações coordenadas e, também, do processo de descoberta de conhecimento. O registro de informações sobre as representações visuais utilizadas, sobre as ações coordenadas definidas, e sobre a seqüência de ações executadas e seus respectivos parâmetros é um instrumento estratégico em processos de extração de informação, e pode oferecer informações úteis para pesquisadores em visualização. Ainda, um usuário particular pode revisar e avaliar suas formas de atuação com base nesse registro, e refinar seu processo de análise caso julgue necessário. A plataforma deve oferecer mecanismos para capturar e armazenar essas informações e outras de interesse. No âmbito da operação de um ambiente, as informações capturadas podem ser utilizadas para desfazer ou repetir ações, e também para a criação de macro operações, isto é, uma seqüência de ações que definem uma tarefa ou sub-tarefa.

O conjunto de representações visuais é classificado na literatura segundo diferentes critérios; por exemplo, a classificação de Keim [Kei 2002] considera como critérios ou eixos de classificação o trio ‘técnicas de visualização, técnicas de interação e distorção, e tipos de dados’; a classificação de Shneiderman [Shn 1996] considera, além dos tipos de dados que uma técnica suporta, os tipos de tarefa identificados no seu ‘*Vis mantra*’. A classificação de North [Nor 2000b], para sistemas de múltiplas visualizações coordenadas, considera a flexibilidade dos sistemas quanto ao tratamento dos dados, a escolha de técnicas de visualização, e a configuração de coordenação. Os critérios apresentados anteriormente – extensibilidade, configurabilidade, monitorabilidade – podem ser utilizados para a classificação de sistemas de múltiplas visualizações coordenadas, complementando as classificações existentes.

Além de estabelecer os três critérios acima, o modelo proposto nesta Seção decompõe a plataforma segundo os módulos identificados na Figura 3.6; o módulo ‘outros recursos’ deixa explícito que um ambiente de análise exploratória deve incluir não apenas técnicas de visualização, mas também ferramentas de acesso a bancos de dados, técnicas de mineração e de instrumentos estatísticos. A integração entre instrumentos visuais e analíticos, como apontado por Mackinlay [Mac 2000], torna a plataforma mais completa, aumentando sua aplicabilidade.

Figura 3.6. Módulos que integram o modelo *AdaptaVis*.

A Figura 3.6 identifica, basicamente, dois grupos de funcionalidade na plataforma. No primeiro grupo, abaixo da linha tracejada, estão as funcionalidades de gerenciamento providas pela plataforma, que garantem o funcionamento integrado dos ambientes configurados a partir dela. Essas funcionalidades estão presentes em todos os ambientes. Já as funcionalidades do segundo grupo são aquelas associadas à exibição e manipulação de representações visuais e acesso aos dados, e variam de um ambiente para outro. No modelo *AdaptaVis*, os elementos da arquitetura são denominados de componentes ‘Básicos’ ou ‘Funcionais’, caso implementem funcionalidades do primeiro grupo ou do segundo, respectivamente. Segundo o modelo, os componentes de um ambiente se agrupam em dois núcleos, o ‘Núcleo Básico’ e o ‘Núcleo Funcional’, conforme ilustrado na Figura 3.7. As funcionalidades de gerenciamento, pertinentes ao Núcleo Básico, são realizadas por quatro componentes básicos:

- **Gerenciador de Interface (Acesso)**. Responsável pela apresentação da interface de acesso à plataforma de visualização: dá ao usuário acesso aos recursos dos diferentes componentes dos núcleos Básico e Funcional, também é responsável pela inicialização dos demais componentes básicos. Solicita aos componentes funcionais a apresentação de suas interfaces;
- **Gerenciador de Componentes (Integração)**. Controla a ativação dos componentes funcionais, e mantém a lista de referência daqueles que estão ativos. Quando um componente funcional é finalizado, este gerenciador procede à sua desativação;
- **Gerenciador de Ações (Coordenação)**. Propaga a ação executada sobre uma representação visual para as demais segundo as ações coordenadas estabelecidas pelo usuário que configurou o ambiente. A coordenação entre as ações é feita dinamicamente pelo usuário, de acordo com as necessidades de cada tarefa;
- **Gerenciador de Históricos (Monitoramento)**. Registra as ocorrências dentro do ambiente, os componentes ativados, as ações coordenadas definidas e suas ativações com os respectivos parâmetros. Esses registros mantêm uma descrição dos vários cenários configurados e das estratégias de atuação adotadas ao longo de um processo de exploração.

Os componentes funcionais podem ser agrupados em diferentes categorias, de acordo com o tipo de funcionalidade oferecida. Essa categorização delimita o que cada componente deve implementar, preservando o caráter genérico dos componentes. Inicialmente, identificamos as seguintes categorias para os componentes funcionais:

- **Fonte de Dados**. Implementam recursos para acesso a arquivos de dados ou gerenciadores de bases de dados externos, bem como recursos para filtragem, amostragem, seleção, e pré-processamento dos dados;
- **Visualização**. Implementam técnicas de visualização e seus recursos de interação;
- **Interação**. Implementam estratégias de interação genéricas, aplicáveis a diferentes técnicas de visualização (como Consultas Dinâmicas [Shn 1994][Ahl 1992], por exemplo);
- **Recursos Analíticos**. Implementam recursos complementares aos de visualização; por exemplo, algoritmos de mineração ou instrumentos estatísticos.

Figura 3.7. Composição dos núcleos Básico e Funcional no AdaptaVis.

A integração e a coordenação entre os componentes do ambiente ocorrem pela troca de mensagens entre componentes básicos e funcionais, de acordo com um protocolo de serviços descrito na Tabela 3.1. Esse protocolo define, também, as regras de aderência de um componente funcional ao modelo *AdaptaVis*.

Tabela 3.1. Descrição funcional do protocolo de serviços para a comunicação entre os componentes.

Serviço	Descrição
Ativação de Componente Funcional	O usuário pode ativar um componente acessando na interface principal um menu que lista o conjunto de componentes funcionais disponíveis na plataforma. Quando isso ocorre, o Gerenciador de Interface ativa o Gerenciador de Componentes, que invoca o serviço de ativação do componente funcional correspondente. Em seguida o Gerenciador de Interface envia uma requisição para o componente funcional exibir sua interface.
Desativação de Componente Funcional	Quando um componente funcional é encerrado pelo usuário, o próprio componente envia uma mensagem para o Gerenciador de Componentes que, por sua vez, atualiza o ambiente para refletir a nova configuração. Por exemplo, ações coordenadas registradas que envolvem esse componente são removidas da configuração do ambiente.
Registro de Ações de Interação Suportadas por um Componente Funcional	Cada componente funcional implementa as ações de interação com o usuário, e informa essas ações para coordenação com outros componentes junto ao Gerenciador de Ações. Quando o usuário deseja definir uma coordenação entre ações em diferentes componentes funcionais, o Gerenciador de Ações é responsável por consultar cada componente funcional envolvido, e por registrar a coordenação.
Propagação de Ação	Um componente funcional envia uma mensagem para o Gerenciador de Ações informando sobre as ações do usuário, juntamente com os parâmetros relacionados, de forma que ações correspondentes em componentes coordenados possam ser invocadas. Uma ação só é propagada se houver registro de coordenação com outras ações, evitando a troca desnecessária de mensagens.

Solicitação de Ação	O Gerenciador de Ações invoca esse serviço para requisitar a realização de uma ação no componente funcional alvo, após receber a informação de propagação de ação do componente fonte.
Registro de Histórico	Cada componente informa o Gerenciador de Históricos sobre as atividades ocorridas dentro do seu escopo de controle.

3.3 CONSIDERAÇÕES FINAIS

Neste Capítulo, foram apresentadas as definições de duas representações visuais temporais – ‘Variação Temporal Uni-escala’ e ‘Variação Temporal Multi-escala’ – e a concepção do modelo *AdaptaVis*. O objetivo do modelo é guiar o desenvolvimento e a implementação de uma plataforma que agregue múltiplas visualizações coordenadas e que possa ser empregada em diversas aplicações, além de oferecer suporte para a avaliação de seu uso. O Núcleo Básico do modelo foi instanciado para permitir a criação de um cenário de múltiplas visualizações coordenadas, na qual ambas as visualizações temporais são utilizadas de forma integrada às visualizações em mapa e tabela, para a observação conjunta de relações temporais e espaciais. As visualizações foram implementadas como componentes funcionais aderentes ao modelo *AdaptaVis*, compondo o Núcleo Funcional. A instanciação gerou a plataforma *InfoVis*, que é descrita no próximo capítulo.

PLATAFORMA *InfoVis* : UMA INSTANCIÇÃO DOS NÚCLEOS BÁSICO E FUNCIONAL DO *AdaptaVis*

Com base no modelo *AdaptaVis*, foi implementada, utilizando a linguagem Java, a versão inicial de uma plataforma de recursos de visualização, denominada *InfoVis*. Inicialmente, essa plataforma oferece um conjunto de técnicas de visualização concebidas para a manipulação de dados com atributos temporais e espaciais. A implementação do Núcleo Básico e seus componentes, bem como dos componentes que integram o atual Núcleo Funcional estabelece o conjunto de componentes básicos e funcionais disponíveis que permite definir, a partir dessa plataforma, ambientes para visualização exploratória de dados temporais e espaciais.

Cada um dos componentes básicos (Gerenciadores de Interface, de Componentes, de Ações e de Históricos) e funcionais (representações visuais temporais uni- e multi-escala, mapa e tabela) foi implementado como uma classe Java distinta. As técnicas de visualização do Núcleo Funcional foram utilizadas inicialmente em sessões de uso com dados de precipitação pluviométrica e, posteriormente, foram adaptadas e estendidas para uso na exploração de dados de transações comerciais. Os resultados das sessões de uso com usuários são abordados no próximo capítulo.

Neste Capítulo são descritos a estrutura geral da plataforma *InfoVis* e os detalhes das implementações dos núcleos Básico e Funcional. Características da ferramenta (Java) adotada para o desenvolvimento, que embasam a escolha, são apresentadas ao final.

4.1 ESTRUTURA GERAL

A plataforma *InfoVis* é o conjunto composto pelos componentes do Núcleo Básico e todos os componentes disponíveis no Núcleo Funcional, bem como pelos ambientes que podem ser configurados a partir desse conjunto, agregando o Núcleo Básico, um subconjunto de componentes funcionais, e registrando um conjunto de ações coordenadas entre os componentes funcionais. Existem duas formas alternativas para a configuração de ambientes, ilustradas na Figura 4.1, que as exemplifica considerando os ambientes hipotéticos A e B. Na primeira, chamada ‘local’ na figura, a plataforma é um módulo único instalado no computador do usuário, a partir da qual são configurados os ambientes, de acordo com as necessidades impostas pelas tarefas do usuário, que escolhe os componentes funcionais e registra as ações coordenadas. Na segunda possibilidade (‘remota’), os componentes de um ambiente são buscados remotamente (via Internet, por exemplo) e podem formar instâncias completamente isoladas (note que cada ambiente tem o seu Núcleo Básico). Observe também que, a partir do acesso à plataforma remota, é possível gerar uma plataforma que opere de forma local, trazendo o Núcleo Básico e todos os componentes funcionais disponibilizados.

Figura 4.1. Configurações local e remota de plataforma/ambiente no *InfoVis*.

Por ser comum em todos os ambientes, o Núcleo Básico poderia, se conveniente, ser implementado como um único módulo englobando todas as funcionalidades de seus componentes. Já os componentes funcionais devem ser implementados, de acordo com o modelo *AdaptaVis*, como módulos separados e independentes.

Os componentes básicos e funcionais (as classes Java), são distribuídos em uma estrutura de diretórios descrita na Figura 4.2. O diretório *<raiz>* pode ser um diretório ou subdiretório, e contém arquivos de configuração da plataforma *InfoVis* e o módulo que ativa a plataforma, isto é, ativa o componente básico de interface com o usuário (Gerenciador de Interface). O subdiretório *<basico>* é o local de armazenamento dos componentes do Núcleo Básico, e o seu nome não pode ser alterado porque os componentes fazem referência a ele. Os demais diretórios podem ter seus nomes atribuídos pelo usuário, e contêm os componentes funcionais; a organização mostrada na Figura 4.2 foi definida para armazenar as técnicas de visualização para dados temporais e o mapa em *<vis>*, e a tabela em *<dados>*.

Figura 4.2. Estrutura de diretórios da plataforma *InfoVis*.

4.2 NÚCLEO BÁSICO

A Figura 4.3 mostra a seqüência de inicialização da plataforma *InfoVis*, válido tanto para plataformas locais como remotas. O componente básico ativado pelo usuário é o ‘Gerenciador de Interface’ (GI), que se encarrega de ativar os demais gerenciadores do Núcleo Básico (GC, GA e GH).

Figura 4.3. Passos para a ativação da interface principal da plataforma *InfoVis*.

O menu de opções apresentado pelo GI é definido em um arquivo de configuração que recebe a denominação *menu.cfg*, e um exemplo é apresentado na Figura 4.4. O quadro à esquerda apresenta o conteúdo do arquivo, e o quadro à direita explica seu formato. Os menus correspondentes, como apresentados na interface da plataforma *InfoVis* são ilustrados, na figura, abaixo dos quadros. A opção ‘Numérico’ é um exemplo de acesso a componentes funcionais implementados segundo o modelo *AdaptaVis*, enquanto a opção ‘Externo’ exemplifica a possibilidade de chamar um utilitário externo (um software matemático que implemente algoritmos de interesse, por exemplo). A coordenação de ações com um utilitário externo, dentro da plataforma, quando possível, pode requerer a implementação de um componente funcional nos moldes do modelo *AdaptaVis* para intermediar a comunicação entre o utilitário e os demais componentes funcionais.

Figura 4.4. Formato do arquivo de configuração *menu.cfg* e o resultado no *InfoVis*.

O acesso aos componentes funcionais é feito a partir do menu de opções exibido na interface com o usuário (ver Figura 4.5), o qual é montado pelo ‘Gerenciador de Interface’ acessando a informação no arquivo texto de configuração, conforme indicado pelas setas mais escuras na Figura 4.3. A medida em que novos componentes são acrescentados à plataforma, o arquivo de configuração de menus é alterado apropriadamente e a interface é atualizada na próxima inicialização, sem necessidade de re-compilação dos componentes já disponíveis. A localização de cada componente funcional na estrutura de diretórios (Figura 4.2) é especificada na variável de ambiente ‘CLASSPATH’ própria da plataforma Java (o termo plataforma é aqui utilizado para referenciar a ferramenta de desenvolvimento).

O recurso de Java que permite que componentes sejam adicionados ao ambiente com uma compilação parcial é a reflexão. Este recurso provê mecanismos para a instanciação de objetos e a invocação de seus métodos em tempo de execução, dispensando o conhecimento de tais elementos em tempo de compilação. A verificação da existência da classe e da correta

chamada aos métodos também é feita em tempo de execução [Lem 2001]. Mais detalhes de Java estão na Seção 4.4.

Figura 4.5. Interface Principal do *InfoVis* e as áreas dos Gerenciadores.

Observe ainda, na Figura 4.5, os espaços reservados na interface aos gerenciadores da plataforma: *Componentes Ativos* para o ‘Gerenciador de Componentes’, na parte superior, exibe informações sobre os componentes funcionais que estão ativos; *Ações Coordenadas* para o Gerenciador de Ações, na parte intermediária, registra e exibe os pares de ações coordenadas; e *Histórico* para o Gerenciador de Históricos, na parte inferior, exibe as atividades e os eventos ocorridos na plataforma durante as interações. Em função da configuração das ações coordenadas ocorrer em tempo de execução, pela escolha das visualizações e ações de interação envolvidas, dois ambientes distintos podem estar ativos simultaneamente, entretanto, na atual versão, o registro de histórico é único.

O registro de um par de ações coordenadas e de histórico é ilustrado na Figura 4.6. O botão ‘Registrar’ ativa a interface para a seleção dos componentes ‘acionador’ e ‘acionado’ e as respectivas ações de interação, cuja sequência é dada pelas setas. Com o botão ‘Adicionar’, o registro é concretizado. As ações realizadas e seus parâmetros ficam registrados no espaço reservado para ‘Histórico’, que podem ser armazenados em arquivo (botão ‘Salvar’).

Figura 4.6. Registro de ações coordenadas e histórico.

4.3 NÚCLEO FUNCIONAL

Como mencionado na Seção 3.2, os componentes do Núcleo Funcional devem aderir ao modelo *AdaptaVis* seguindo um protocolo de serviços que assegura o funcionamento integrado e coordenado dos componentes nos ambientes. O protocolo é implementado por um conjunto de métodos que possibilita a comunicação entre os diferentes componentes. Esse conjunto de métodos é descrito na Tabela 4.1.

Tabela 4.1. Métodos a serem implementados por um componente funcional.

Método	Descrição
Construtor	Na atual versão, o construtor é implementado sem parâmetros.
doAction	Recebe mensagem do Gerenciador de Ações para realizar uma ação coordenada ligada a uma ação do componente acionador.
showGUI	Torna o componente visível exibindo sua GUI, atendendo uma requisição do Gerenciador de Interface.
setManagers	Recebe as referências aos “Gerenciadores” para comunicação com eles.
getComponentActions	Informa as ações de interação que o componente pode realizar.

getComponentCategory	Informa a categoria a qual o componente pertence.
setIndex	Recebe o identificador do componente na plataforma, que é um índice atribuído pelo Gerenciador de Componentes
addRegisteredAction removeRegisteredAction	Informação para controle de ação coordenada, utilizada para verificar a necessidade de informar o Gerenciador de Ações, para que este solicite uma ação ao componente acionado

A seguir é apresentado um trecho de código genérico para um componente funcional, mostrando detalhes dos métodos listados na Tabela 4.1, juntamente com as variáveis necessárias, as quais podem receber outros nomes. Com base nesse arcabouço, outros componentes funcionais podem ser desenvolvidos e incorporados ao ambiente atualizando o arquivo de configuração de menus descrito na Figura 4.4.

Trecho de código genérico para um componente funcional na plataforma InfoVis

```

import basico.*; // Núcleo Básico, os componentes devem estar em um diretório denominado basico

public class NomeComponenteFuncional extends JFrame implements ActionListener
{
    private LogManager lmanager;           // Gerenciador de Históricos
    private ComponentsManager cmanager;    // Gerenciador de Componentes
    private ActionsManager amanager;       // Gerenciador de Ações

    private String[] actionsid = {"Escolher", "Destacar"}; // Lista de Rótulos das Ações de Interação
    private final int NRACTIONS = actionsid.length;      // Quantidade de Ações de Interação Disponíveis para o Componente
    private int[] actionsregd = new int[NRACTIONS];      // Sinalizador: <>0, ação está ativada para coordenação; =0, não está
    private Vector actions = new Vector();               // Lista de Rótulos informados ao Gerenciador de Ações

    private final String cmptitle = "Componente Funcional"; // Rótulo Identificador do Componente Funcional
    private String cmptitleindex; // Rótulo do Componente acrescido de seu índice
    private int indice=-1; // Índice do Componente, atribuído pelo Gerenciador de Componentes
    private Integer intclass; // ~índice, estrutura utilizada para aproveitar a reflexão do Java

    public NomeComponenteFuncional() { // Construtor
        int i;
        for (i=0; i<NRACTIONS; i++) {
            actionsregd[i] = 0; // Inicialmente, a ação não faz parte de nenhuma lista de ações coordenadas
            actions.add(actionsid[i]); // Preparação da Lista de Ações de Interação informada ao Gerenciador de Ações
        } // Estrutura utilizada para usar a característica de reflexão do Java
    }

    public void actionPerformed(ActionEvent e) { // Captura os eventos deste componente, por exemplo, opção de menu ativada
        // códigos específicos para cada evento
    }

    public void doAction(String s, Vector v) { // Executa uma 'ação coordenada chamada' (ação-2 do par) requisitada pelo
        // Gerenciador de Ações
        // s: identificador da ação a ser realizada, v: parâmetros para a ação (lista de ids dos itens de dados)

        if (s == "Destacar") { // Exemplo de realização da ação 'Destacar'
            // Realiza a ação
            if ( 'sinalizador da ação é diferente de 0' ) {
                amanager.setActionDone(indice, "Destacar", v); // Sinalizador <> 0, o Gerenciador de Ações é informado para que ele
                // requisite a um terceiro componente uma ação coordenada com esta.
                // indice: índice do componente
            }
            lmanager.writeLogRecord(logline, indice); // Gerenciador de Históricos registra
            // logline: frase sobre a ação, indice: índice do componente
        }
    }
}

```

```

public void showGUI() {
    cmanager.appendDescription("<" + cmptitleindex + ">",indice); // Descrição passada ao Gerenciador de Componentes,
                                                                // apresentada por este na Interface Principal
                                                                // índice: índice do componente
    lmanager.writeLogRecord("GUI <" + cmptitleindex + "> exibida",indice); // Gerenciador de Históricos registra
                                                                // índice: índice do componente
    addWindowListener ( // Usuário desativa este componente
        new WindowAdapter() {
            public void windowClosing(WindowEvent e) {
                painelcontrole.bye();
                lmanager.writeLogRecord("Componente em desativação",indice); // Gerenciador de Históricos registra
                cmanager.deactivateComponent(intclass); // Gerenciador de Componentes informado sobre a desativação
                                                                // intclass ~ índice, usado para utilizar a reflexão no Java
            }
        }
    );
    show();
}

public void setManagers(ComponentsManager cm, ActionsManager am, LogManager lm) {
    cmanager = cm; // Este componente recebe as referências para os Gerenciadores do ambiente
    amanager = am;
    lmanager = lm;
}

public Vector getComponentActions() {
    return actions; // Este componente informa os rótulos das ações de interação que pode realizar
}

public Integer getComponentCategory() {
    return new Integer(indice_categoria); // Este componente informa sua categoria de componente funcional
}

public void setIndex(Integer i) { // Este componente recebe do Gerenciador de Componentes seu índice no ambiente
    indice = i.intValue(); // e realiza os processamentos necessários
    intclass = new Integer(indice);
    cmptitleindex = cmptitle + "-" + indice;
    setTitle(cmptitleindex);
}

public void addRegisteredAction(String actid) { // Este componente é informado pelo Gerenciador de Ações que sua
    int i = actions.indexOf(actid); // ação 'actid' foi registrada em um par de ações coordenadas
    actionsregd[i]++;
}

public void removeRegisteredAction(String actid) { // Este componente é informado pelo Gerenciador de Ações que um par
    int i = actions.indexOf(actid); // de ações coordenadas envolvendo sua ação 'actid' foi removido
    actionsregd[i]--;
}

//*****
// classes e métodos adicionais específicos
//*****
} fim da classe NomeComponenteFuncional

```

A Tabela 4.2 lista as principais atividades executadas no âmbito da plataforma e as principais mensagens trocadas pelos diferentes componentes da plataforma para a execução das variadas atividades. Esses componentes incluem: o Gerenciador de Interface (GI) que dá acesso aos recursos da plataforma; os demais componentes (gerenciadores) do Núcleo Básico (GC, GA e GH); e os componentes funcionais (CF) responsáveis por visualização e acesso a dados. Os métodos ativados são apresentados delimitados por '['e ']'. Note que os métodos dos componentes funcionais devem ser codificados exatamente com os identificadores apresentados para conformidade com o protocolo padrão da plataforma *InfoVis*. Observe, também, que alguns métodos listados não foram descritos na Tabela 4.1, pois são métodos implementados nos componentes básicos.

Tabela 4.2. Mensagens trocadas entre os componentes em diversas atividades na plataforma InfoVis.

Siglas GI : Gerenciador de Interface GC : Gerenciador de Componentes GA : Gerenciador de Ações GH : Gerenciador de Históricos CF : Componente Funcional (mn) : mensagem número <i>n</i>		
Atividade: Inicialização da GUI Principal da Plataforma		
Emissor	Receptor(es)	Mensagem
GI	GC	Informação sobre o número máximo de componentes ativos simultaneamente [Construtor do GC] (m1)
GI	GC, GA, GH	Cada gerenciador recebe as referências dos outros gerenciadores [setComponentsManager] (m2) [setActionsManager] (m3) [setLogManager] (m4)
GI	GC, GA, GH	Solicitação das áreas de visualização de informação de cada gerenciador [getActiveComponentsInfoViewer] (m5) [getCoordinatedActionsInfoViewer] (m6) [getLogBoard] (m7)
GI	GC	Informações sobre os itens do menu da GUI principal [setMenuItemsInfo] (m8)
Diagrama de troca de mensagens <pre> graph TD GI[GI] -- "m2, m3, m7" --> GH[GH] GI -- "m1, m3, m4, m5, m8" --> GC[GC] GI -- "m6, m2, m4" --> GA[GA] </pre>		
Atividade: Ativação de um Componente		
Emissor	Receptor(es)	Mensagem
GI	GC	Solicitação de ativação de um componente definido pela escolha no menu da GUI principal [activateComponent] (m1)
GI	GC	Solicitação de informação sobre o tipo de item de menu (aplicativo externo ou componente) [getMenuItemType] (m2)
GC	CF	Informação sobre o índice que foi atribuído ao componente [setIndex] (m3)
CF	GC	Solicitação para o acréscimo de descrição sobre o componente no quadro de visualização do GC [appendDescription] (m4)
GC	CF	Solicitação da categoria a qual o componente pertence [getComponentCategory] (m5)
GI	CF	Referências dos gerenciadores [setManagers] (m6)
GI	CF	Solicitação para exibição da GUI do componente [showGUI] (m7)

Diagrama de troca de mensagens

Atividade: Desativação de um Componente

Emissor	Receptor(es)	Mensagem
CF	GC	Aviso sobre a desativação do componente [deactivateComponent] (m1)
GC	GA	Aviso sobre a desativação do componente [setDeactivatedComponent] (m2)
GA	GC	Solicitação da referência do componente [getComponentRef] (m3)
GA	CF	Solicitação para atualização do controle de ação coordenada com o componente desativado [removeRegisteredAction] (m4)

Diagrama de troca de mensagens

Atividade: Registro de uma Ação Coordenada

Emissor	Receptor(es)	Mensagem
GI	GA	Solicitação para ativação do processo de registro de ação coordenada [registerCoordinatedActions] (m1)
GA	GC	Solicitação da lista de componentes ativos [getComponentStatusList] (m2)
GA	GC	Solicitação da referência do componente [getComponentRef] (m3)
GA	CF	Solicitação das ações disponibilizadas pelo componente [getComponentActions] (m4)
GA	CF	Solicitação ao componente acionador para atualização do controle de ação coordenada [addRegisteredAction] (m5)

Diagrama de troca de mensagens

Atividade: Exclusão de Ação Coordenada

Emissor	Receptor(es)	Mensagem
GI	GA	Solicitação para remoção de registros de ação coordenada [removeCoordinatedActions] (m1)
GA	GC	Solicitação de referência do componente acionador [getComponentRef] (m2)
GA	CF	Solicitação para atualização do controle de ação coordenada [removeRegisteredAction] (m3)

Diagrama de troca de mensagens

Atividade: Coordenação de Ação		
Emissor	Receptor(es)	Mensagem
CF	GA	Aviso sobre a realização de uma ação coordenada [setActionDone] (m1)
GA	GC	Solicitação de referência do componente acionado [getComponentRef] (m2)
GA	CF	Solicitação ao componente acionado o realização da ação coordenada [doAction] (m3)

Diagrama de troca de mensagens

Atividade: Registro de Histórico (ocorre em todas atividades)		
Emissor	Receptor(es)	Mensagem
GI	CH	Solicitação de registro de histórico oriundo da Interface Principal [writeLogFromMI] (m1)
GC	CH	Solicitação de registro de histórico oriundo do Gerenciador de Componentes [writeLogFromCM] (m2)
GA	CH	Solicitação de registro de histórico oriundo do Gerenciador de Ações [writeLogFromAM] (m3)
CF	CH	Solicitação de registro de histórico oriundo de um componente [writeLogRecord] (m4)

Diagrama de troca de mensagens

4.3.1 TÉCNICAS DE VISUALIZAÇÃO IMPLEMENTADAS

Nesta Subseção são descritos os componentes funcionais implementados visando atender usuários que manipulam dados com atributos temporais e espaciais. A Tabela 4.3 apresenta os componentes (‘Distribuição Espacial’ ou ‘Mapa’, ‘Intervalo/Categoria’ ou ‘Intervalo em Anos’, ‘Dados de Chuva’ e ‘Tabela de Dados’) e descreve as suas principais características, bem como as ações de interação suportadas. Um conjunto mínimo de ações de interação foi disponibilizado para a ação coordenada entre esses componentes (ações ‘Escolher’ e ‘Destacar’). Cada elemento do conjunto de dados (item de dado) é identificado por um índice seqüencial que corresponde à sua posição dentro arquivo. Os índices são passados nas mensagens, dentro da lista de parâmetros nos métodos ‘doAction’ e ‘setActionDone’ (ver trecho de código), trocadas entre os componentes envolvidos em uma ação coordenada. As funcionalidades implementadas em cada componente são apresentadas mais detalhadamente na descrição das sessões de uso, no próximo capítulo.

Tabela 4.3. Características dos componentes funcionais implementados.

Componente <i>InfoVis</i>	Características	
Distribuição Espacial ou Mapa	Os itens de dados são representados por pontos (pixels) e podem ser posicionados sobre um mapa de acordo com os atributos espaciais de latitude e longitude, que estão no formato de graus e minutos. Um outro atributo é mapeado por cor, através de uma tabela de mapeamento definida pelo usuário. Mapas são de uso comum para dados dessa natureza e, portanto, de fácil entendimento. Os pontos podem ser exibidos como um único pixel ou como uma matriz quadrada de pixels, de acordo com a escala (ampliada/reduzida) de apresentação escolhida.	
	Ações de Interação	“Escolher”
		“Destacar”

INTERFACE PRINCIPAL DO COMPONENTE FUNCIONAL DISTRIBUIÇÃO ESPACIAL

The screenshot shows the main interface of the 'Distribuição Espacial' component. It consists of two main windows. The top window, titled 'Distribuição Espacial-0 <postos.dat>', displays a map with numerous colored points. The bottom window, titled 'Painel de Controle Distribuição Espacial-0 <a1alt1.cor>', contains a control panel. This panel includes a table for defining value intervals and corresponding colors, a scale control (set to 2), and buttons for selecting position attributes (Latitude, Longitude, Altitude) and a 'Destacar' (Highlight) button. A label 'INTERFACE PRINCIPAL DO COMPONENTE FUNCIONAL DISTRIBUIÇÃO ESPACIAL' points to the top window.

Intervalo em Anos ou Intervalo/Categoria (uma implementação de Variação Temporal Uni-escala) Ações de Interação	Os itens de dados são representados por uma seqüência de pontos formando uma linha horizontal. A cor de cada ponto indica a qualidade do dado e o comprimento da linha mostra o tamanho do intervalo em anos com presença de dados. Uma linha vertical guia pode ser posicionada sobre um determinado ano, para uma referência mais precisa.	
	“Escolher”	Com o arrasto do mouse na direção vertical com um dos botões pressionado, escolhe-se as linhas contidas no intervalo de arrasto.
	“Destacar”	Um segmento de reta prolongando uma linha é usado para destacar tal linha.

Dados de Chuva (uma implementação de Variação Temporal Multi-escala)	Este componente mostra a variação temporal de um atributo, no caso chuva, relacionado com um único item de dado (objeto espacial), nas escalas diária, mensal e anual. O atributo é mostrado em um conjunto de janelas retangulares alinhados verticalmente, ou seja, uma coluna de 25 janelas representa um determinado ano. Nas 12 janelas superiores, em cada retângulo são apresentados os valores de chuva, dia-a-dia, para o mês correspondente. Nas 12 janelas seguintes, cada retângulo é pintado com uma única cor que representa o total mensal de chuva. No último retângulo, ou janela, a cor indica o total anual. Cada uma das escalas tem uma tabela de cores para mapeamento de valores. A variação do tamanho do ponto (escala de ampliação/redução) é interessante para observar com maior detalhe a escala diária. Contornos retangulares, horizontal e vertical, são usadas como guias para posicionar sobre um determinado mês e ano, respectivamente.	
Ações de Interação	Ações “Escolher” “Destacar”	Não implementados na atual versão. A generalização no formato das mensagens trocadas para coordenação de ações, em versões futuras, permitirá as suas implementações.

Tabela de Dados	Lê o conjunto de dados e mostra os atributos na sua forma original, texto ou número, em formato de tabela. Cada linha representa um item de dado, e cada coluna um atributo. Uma tabela auxiliar permite observar um determinado atributo ordenadamente, sem contudo, reorganizar a tabela toda. Itens de dados podem ser selecionados na tabela auxiliar e, posteriormente, marcados na tabela principal. Como esta marcação é cumulativa, várias tabelas auxiliares podem ser usadas simultaneamente, simulando uma operação OU. Uma operação E, envolve o uso de um arquivo intermediário. Os elementos marcados na tabela principal podem ser gravados em arquivo, assim como o seu complemento, ou seja, os não marcados.	
Ações de Interação	“Escolher”	Os elementos de dados são escolhidos pelo mouse, marcando-se intervalos de elementos ou individualmente. Em virtude da marcação ser cumulativa, a propagação das escolhas é feita manualmente.
	“Destacar”	As linhas da tabela são destacadas com uma cor de fundo diferenciada.

tabela auxiliar ativada

atributo bacia, do tipo string, ordenado

coluna direita contém os índices (posição na tabela principal) dos itens ordenados

tabela principal

item marcado

Salvar escolhidos Salvar não escolhidos Propagar escolhas marcados: 0

itens selecionados (marcados) são propagados manualmente

Na atual versão da plataforma *InfoVis*, cada componente funcional deve carregar o conjunto de dados em seu espaço de memória. Ainda, as ações de interação que podem ser coordenadas são aquelas que envolvem o mesmo conjunto de dados ou conjuntos que tenham os mesmos itens de dados, ainda que com alguns atributos diferentes. Versões futuras deverão oferecer maior flexibilidade nesses dois pontos.

4.4 FERRAMENTA DE DESENVOLVIMENTO

A plataforma Java foi escolhida como ferramenta de implementação por oferecer não apenas uma linguagem de programação, mas um conjunto de soluções integradas adequadas para tratar vários problemas críticos na arquitetura da plataforma *InfoVis*. Apesar de extensa e complexa, espera-se que a escolha de Java permita um bom aproveitamento do esforço dispendido na concepção de uma solução flexível e genérica.

Dentre os recursos disponíveis estão o Java3D [Sow 1998] para implementar as operações gráficas, o JDBC [Lem 2001] para conexão com bancos de dados externos, e o suporte aos protocolos XML/SOAP [Lib 2001][Scr 2000] para troca de informação entre diferentes componentes. Ainda, RMI [Hor 2001][Eck 2000] pode oferecer os recursos necessários à distribuição dos componentes e aplicativos. Os aspectos determinantes para a escolha da plataforma Java estão diretamente relacionados aos requisitos e critérios apresentados na Seção 3.2, tais como independência e auto-suficiência dos módulos e baixo impacto na incorporação de novos módulos, que são naturalmente satisfeitos com a opção pelo desenvolvimento puramente em Java.

Uma outra vantagem bastante importante decorrente da opção pelo Java é a independência de plataforma, que provê condições para tratar aspectos de escalabilidade, com o desenvolvimento em equipamentos de configurações relativamente modestas e uso real em ambientes variados e de maior porte e desempenho. Um código fonte Java é transformado em uma representação intermediária – bytecodes – que é convertida em tempo de execução, pela JVM (*Java Virtual Machine*), para o respectivo código nativo, provendo portabilidade ao código. Entretanto, quando a conversão é feita em um esquema puramente interpretado, o desempenho é baixo, e esse é um dos pontos mais criticados na plataforma Java, um aspecto que tende a ser ainda mais crítico em se tratando de um ambiente de visualização exploratória. Entretanto, novos recursos vêm sendo incorporados à plataforma para melhorar esse aspecto e aumentar a gama de aplicações que podem ser implementadas [Arm 1998][Man 1998][Sun 2001]. Em particular, melhor desempenho, mantendo a portabilidade, pode atualmente ser obtido com um compilador JIT (*Just-In-Time*) ou com a máquina virtual *HotSpot*, disponível a partir da versão 1.3 do Java.

Um compilador JIT converte os bytecodes para código nativo, e introduz algum grau de otimização na primeira vez que sua execução é requisitada, mantendo em memória o resultado da conversão para uso posterior e evitando assim a necessidade de nova conversão. Nesse caso, diferente do esquema de um compilador tradicional, a conversão é feita em tempo de execução, e apenas dos trechos requisitados, podendo ser até 50 vezes mais rápida do que em um esquema de interpretação puro [Arm 1998]. Mangione [Man 1998] apresenta resultados de testes comparativos entre C++ , Java (JIT) e Java (interpretação de bytecodes), cujos tempos de execução de algumas operações são listados na Tabela 4.4, considerando apenas o tempo de execução das instruções relacionadas.

Tabela 4.4. Tempos, em segundos, para a execução de tarefas por códigos gerados por um compilador C++, Java com compilador JIT e Java com interpretação de bytecodes [Man 1998].

Descrição do teste	Tempo (segundos)		
	C++	Java (JIT)	Java (interpretação)
Repetição de divisão com inteiro : 10.000.000 vezes	1.8	1.8	4.8
Repetição de divisão com ponto-flutuante : 10.000.000 vezes	1.6	1.6	8.7
Repetição de chamada a método estático que contém uma divisão inteira	1.8	1.8	6.0
Repetição de chamada a método membro que contém uma divisão inteira	1.8	1.8	10.0

No esquema do HotSpot, a conversão para código nativo ocorre de forma mista. Na primeira vez, os *bytecodes* são processados pelo interpretador, e a execução do programa é monitorada e analisada procurando identificar “*hot spots*” para aplicar, adaptativamente, um esquema de compilação com otimização [Sun 2001][Arm 1998]. Para uma melhor adaptação a diferentes características de desempenho, o compilador *HotSpot* é disponibilizado para dois perfis: cliente e servidor. O primeiro é uma atualização da máquina virtual clássica e de compiladores JIT de versões anteriores do Java SDK, e oferece um desempenho melhorado para aplicações e *applets* pela redução de tempo de inicialização e ocupação de memória. Para o perfil *server*, a máquina virtual contém um compilador adaptativo que se mostra mais adequado a

aplicações em servidores. Ambos os compiladores operam sobre uma máquina virtual comum, como indicado na Figura 4.7 [Sun 2001].

Nossa visão é que o acréscimo desses recursos torna menos crítica a questão do fraco desempenho de tempo de execução do Java, e que as vantagens de utilizar a plataforma superam a necessidade de tratar uma eventual perda de desempenho. Essa questão do desempenho, na verdade, vai além da discussão relativa ao código interpretado versus compilado, já que o desempenho é influenciado por outros fatores, como algoritmos otimizados, configuração de hardware e sistema operacional, recursos de paralelismo e distribuição. Esses aspectos são tratáveis na plataforma Java.

Figura 4.7. Configuração da máquina virtual *HotSpot* que permite a escolha de um perfil para melhor adequação à aplicação [Sun 2001].

4.5 CONSIDERAÇÕES FINAIS

Neste Capítulo foi detalhada a implementação de uma instância do modelo *AdaptaVis*, a plataforma *InfoVis*. Foram descritos as alternativas de configurações local ou remota, a estrutura de diretórios, a configuração do menu de opções, os métodos do protocolo padrão de serviços, o funcionamento do Núcleo Básico, o arcabouço inicial do Núcleo Funcional, e o mecanismo de troca de mensagens para a coordenação de ações entre múltiplos componentes. Os gerenciadores e as representações visuais temporais, juntamente com um mapa e uma tabela, foram implementados como classes Java independentes, e integrados em uma plataforma para permitir a definição de ambientes específicos. A implementação em classes separadas facilita a disponibilização das configurações local e remota, aproveitando os componentes sem nenhuma alteração, requerendo apenas uma organização para acesso.

Com a implementação dos núcleos Básico e Funcional da plataforma *InfoVis*, instanciando o modelo *AdaptaVis*, foi possível conduzir sessões de uso envolvendo usuários e tarefas de exploração de conjuntos de dados com atributos temporais e espaciais em dois domínios de aplicação. Essas sessões de uso são descritas no próximo capítulo.

VALIDAÇÃO DO MODELO E DAS REPRESENTAÇÕES VISUAIS TEMPORAIS

A plataforma *InfoVis*, descrita no Capítulo 4, e ambientes de visualização exploratória gerados a partir dela foram validados empiricamente em sessões de uso realizadas com os objetivos de: (1) validar a plataforma enquanto prova de conceito para o modelo base, verificando se a arquitetura proposta é capaz de efetivamente viabilizar a configuração de ambientes com múltiplas visualizações coordenadas dinamicamente de maneira útil e flexível para potenciais usuários, a partir de componentes independentes; (2) verificar a utilidade das técnicas de visualização temporais propostas para apoiar a exploração e análise de conjuntos de dados espaço-temporais; (3) verificar a extensibilidade da plataforma para domínios diversos por meio da adequação de componentes existentes.

As sessões de uso foram realizadas em encontros agendados com usuários de dados gerados nos domínios de Climatologia e Negócios, os quais tinham uma relação de atividades pré-definidas a serem executadas envolvendo a análise dos dados. No domínio de Climatologia, as sessões de uso previam a execução de tarefas integrantes da fase inicial de um processo de classificação climática (uma descrição geral desse processo é apresentada no Apêndice A), utilizando dados reais de precipitação pluviométrica, contando com a participação de um especialista do domínio, acompanhado pelo especialista em visualização (o autor desta tese). O processo de classificação climática já havia sido conduzido anteriormente pelo especialista do domínio sem o emprego de técnicas de visualização exploratória. A execução das tarefas com o

apoio dessas técnicas permitiu estabelecer uma comparação empírica entre os resultados obtidos com ambas as abordagens, bem como colher as impressões do usuário especialista. Os mesmos componentes funcionais também foram aplicados em uma empresa, para analisar conjuntos de transações comerciais contendo atributos temporais e espaciais. O objetivo das sessões de uso foi explorar os dados para descobrir novas informações, completando ou detalhando aquelas advindas de processamento numérico tradicionalmente aplicado. Em ambos os casos, foram apreciadas as vantagens e desvantagens envolvidas na adoção de uma abordagem com visualização interativa, no contexto de cada domínio de aplicação.

Neste Capítulo, são apresentadas as sessões de uso conduzidas, descrevendo as tarefas realizadas, os resultados obtidos e as respectivas conclusões, analisando as facilidades e dificuldades no uso da plataforma/ambientes e das visualizações, e discutindo possíveis adaptações e melhorias desejáveis.

5.1 EXPLORAÇÃO DE DADOS DE PRECIPITAÇÃO PLUVIOMÉTRICA

Essa aplicação foi escolhida para validar a plataforma por se tratar de um problema real e permitir a comparação com resultados já validados, bem como verificar se o uso de múltiplas visualizações coordenadas efetivamente proporciona benefícios do ponto de vista do usuário final. A exploração e análise dos dados de precipitação compõem uma etapa inicial cujos resultados são importantes para as etapas posteriores de uma metodologia de classificação climática [Flo 2000]. Essa etapa inicial demanda muito processamento e tempo, em função de exigir o tratamento de dados coletados ao longo de um período de, no mínimo, 30 anos. Além da grande quantidade de dados, existe, também, a necessidade de analisar simultaneamente vários atributos (posição, relevo, qualidade dos dados e variação temporal). Esses dois aspectos motivam o uso de múltiplas visualizações coordenadas buscando melhorar a produtividade (redução de tempo) e aumentar a confiabilidade (decisões melhor fundamentadas). As sessões de uso nesse domínio abrangeram os objetivos 1 e 2, ou seja, de validação da plataforma/ambientes e de verificação da utilidade das visualizações temporais propostas.

O Prof. Edílson Flores, no desenvolvimento de sua tese de doutorado em Climatologia [Flo 2000], analisou medidas de precipitação pluviométrica registrados em postos de coleta do Departamento de Águas e Energia Elétrica do Estado de São Paulo (DAEE), distribuídos em

diversos pontos do estado. Esses dados são disponibilizados no aplicativo BcDAEE (ver Figura A.1 no Apêndice A), sendo que o Prof. Flores analisou medidas da região oeste do estado para propor uma classificação climática para essa região. O produto final de uma classificação climática é um mapa identificando regiões com características climáticas homogêneas, definidas a partir da análise de variáveis de clima, tais como chuva, temperatura e pressão atmosférica. Cada posto de coleta é identificado por um prefixo e tem como atributos espaciais a latitude, longitude e altitude. A quantidade de chuva diária, em milímetros, registrada em cada posto ao longo de vários anos, armazenada na base de dados, define um atributo temporal. O período de medidas, em anos, varia de posto para posto, já que os anos de início ou de término de operação dos postos podem ser diferentes, e certos postos podem ter sofrido uma interrupção temporária de atividades. Algumas visualizações preliminares desses dados, descritas no Apêndice B, foram examinadas com o Prof. Flores, a fim de verificar a aplicabilidade e potenciais benefícios do uso de recursos de visualização.

Inicialmente, foram conduzidas as sessões de familiarização com a plataforma *InfoVis*. Nas demais sessões de uso, o especialista do domínio de aplicação teve liberdade para escolher as técnicas de visualização e definir as ações coordenadas mais adequadas, na sua opinião, aos objetivos de cada tarefa, embora o especialista em visualização tenha apresentado sugestões em algumas situações.

A partir da abordagem adotada originalmente na análise de medidas de chuva para o processo de classificação climática, foram identificadas três tarefas fundamentais, compostas por um único ou por múltiplos passos:

- **Tarefa 1: Definição de um conjunto de postos para processamento.** A Organização Mundial de Climatologia sugere que a análise seja conduzida em um período mínimo de 30 anos. Assim, nessa tarefa, o objetivo é encontrar postos cujo período de coleta satisfaça esse requisito, i.e., tenham um período de coleta comum, superior a 30 anos e cujos dados no período apresentem um bom grau de confiabilidade, ou seja, ou foram consistidos ou os valores podem ser conferidos ou inferidos de forma confiável. Na abordagem original, foi seguida a seguinte sequência de passos para executar essa tarefa:
 - **Passo 1.1.** Foram determinados os postos contidos na região de interesse (oeste do estado) pela delimitação manual, a lápis, sobre um mapa impresso da região

que continha os postos plotados como pontos, conforme exemplificado na Figura 5.1;

- **Passo 1.2.** Foi feita uma seleção dos postos determinados no passo anterior, descartando aqueles cujo período de coleta de dados é inferior a 30 anos. A escolha dos postos para permanência ou descarte foi feita consultando uma tabela organizada com os postos nas linhas e os anos de coleta nas colunas (ver Figura 5.2). As medidas podem ser classificadas, segundo a confiabilidade, como dados consistidos ou não consistidos. O primeiro tipo denota valores que foram conferidos e validados; enquanto que o segundo indica dados coletados, mas não validados. Na tabela, as colunas eram preenchidas com os símbolos 'X' e '*', representando anos com dados consistidos e não consistidos, respectivamente; e, por consequência, o não preenchimento (em branco) indicava inexistência de medidas para o ano. O processo de construção dessa tabela também foi manual, com base nas informações contidas no banco de dados de precipitação pluviométrica;
- **Passo 1.3.** Se o período de coleta dos postos escolhidos no passo 1.2 é adequado para os propósitos da classificação climática, pode ser feito um segundo processo de descarte de postos considerando as suas características geográficas. Por exemplo, se dois postos estão próximos e na mesma altitude, é possível considerar as medidas de apenas um deles na análise, já que provavelmente ambos têm comportamento semelhante quanto à precipitação. Nesse caso, pode-se escolher o posto com o maior período de coleta. A decisão de descarte não deve considerar apenas esses dois postos, mas também os seus vizinhos. Para realizar esse passo, foram utilizados o mapa do passo 1.1 e os atributos geográficos dos postos, informados no banco de dados;

A principal dificuldade nessa tarefa foi gerar, agregar e adequar, a partir dos dados brutos, as informações necessárias à realização dos passos, requerendo acesso ao banco de dados mesclado com processamento manual e criando dificuldades para tratar e relacionar vários atributos simultaneamente. Além da escolha dos postos com período de coleta adequado, o outro objetivo desta etapa é diminuir o volume de dados para tratamento nas tarefas seguintes. Os passos 1.2 e 1.3 podem vir a ser repetidos caso, em tarefas posteriores, seja detectado que a lista de postos deve ser revista, como por

exemplo, se for percebido que as informações de um posto descartado poderiam melhorar algum resultado;

- **Tarefa 2: Preenchimento de falhas.** A informação sobre o grau de confiabilidade dos dados, consistidos ou não, usada na tarefa anterior para selecionar postos, é relativa ao ano. Entretanto, no caso de dados não consistidos, pode haver dias nos quais os dados não foram coletados, requerendo que os valores correspondentes sejam inferidos por meio de métodos de interpolação. O preenchimento de falhas, ou valores ausentes (o termo falha é empregado pelo especialista da aplicação), é o objetivo da Tarefa 2, na qual foram utilizadas, na abordagem original, tabelas impressas contendo os valores diários de chuva e os totais mensais para cada posto. Um valor ausente pode ser inferido de duas formas:

- **Forma 2.1.** O valor é estimado com base em valores de precipitação de outros anos do próprio posto, que apresentam o mesmo ritmo ou comportamento de chuva do ano sendo analisado;
- **Forma 2.2.** O valor é estimado com base em valores registrados nos postos vizinhos, observando o comportamento da chuva no ano em questão. As tabelas dos postos eram dispostas sobre uma mesa de forma a refletir o seu posicionamento geográfico relativo, junto ao mapa com os postos plotados.

Nessa tarefa, a maior dificuldade consistiu justamente em comparar os postos, observando o comportamento e os valores de chuva, para decidir qual das duas formas empregar para inferir o valor de chuva a ser atribuído à falha. O comportamento da chuva era analisado com base no ritmo ou variação dos totais mensais, para viabilizar a análise. Os mesmos procedimentos foram utilizados para confirmar valores registrados aparentemente discrepantes, como por exemplo, uma variação abrupta inesperada. Em ambas as formas, após o cálculo dos valores, uma conferência poderia ser feita considerando postos distantes, mas em regiões de altimetria semelhante à região em análise;

- **Tarefa 3: Determinação de anos representativos.** Com os dados validados, o objetivo da Tarefa 3 é determinar três anos, comuns a todos os postos considerados, cujas precipitações registradas melhor representem as categorias: ano seco, ano chuvoso ou ano habitual. Na abordagem original, os anos representativos foram determinados com o auxílio de árvores de ligação (ver Figura A.2) geradas individualmente para cada posto a partir de um algoritmo de agrupamento. Tendo em vista que 150 postos foram analisados,

a grande dificuldade nessa tarefa foi inspecionar todas árvores de ligação e determinar os melhores representantes de cada categoria.

Figura 5.1. Mapa do aplicativo BcDAEE impresso, com delimitação manual de região de interesse, oeste do Estado de SP.

Prefixo	Nome do Posto	Alt.	Lat.	Long.	ANOS				
					58	59		96	97
					X	X	• • •	*	*
							• • •		
					•				
					•				
					•				
							• • •		

Figura 5.2. Modelo de tabela sintética sobre quantidade e qualidade de dados. Extraído de [Flo 2000].

A Figura 5.3 mostra a interface da plataforma *InfoVis* para apoiar as três tarefas descritas, as quais são realizadas em ambientes definidos a partir da escolha e coordenação entre visualizações. No âmbito deste projeto, o termo ‘plataforma’ denota, como discutido no Capítulo 4, o conjunto dos componentes básicos e de todos os componentes funcionais disponíveis, enquanto ‘ambiente’ denota uma configuração específica da plataforma. Aqui, a plataforma segue o esquema ‘local’ (ver Figura 4.1), e o conjunto de componentes funcionais disponíveis (visualizações temporais, mapa e tabela, descritos na Tabela 4.3) pode ser visto pelo conteúdo do arquivo de configuração *menu.cfg*. Os componentes funcionais são referenciados com os termos utilizados nas sessões com o usuário:

- ‘Intervalo/Categoria’ ou ‘Intervalo em Anos’ indica a implementação da visualização denominada ‘Variação Temporal Uni-escala’ (introduzida no Capítulo 3);
- ‘Dados de Chuva’ indica a implementação da visualização ‘Variação Temporal Multi-escala’ (introduzida no Capítulo 3);
- ‘Distribuição Espacial’ indica o posicionamento dos itens de dados no mapa;

- ‘Tabela de Dados’ indica a forma tabular de exibição dos itens de dados.

Figura 5.3. Configuração da plataforma *InfoVis* para as sessões de uso com dados de chuva.

A seguir são apresentados os detalhes das sessões de uso, e dos respectivos ambientes configurados para a execução dessas tarefas.

5.1.1. DEFINIÇÃO DE UM CONJUNTO DE POSTOS PARA PROCESSAMENTO

O objetivo desta tarefa é definir um conjunto mínimo de postos na região de estudo. Em um primeiro momento, todos os postos da região oeste do estado são considerados de interesse. Em seguida, o usuário verifica quais postos podem ser descartados, de maneira a diminuir o volume de dados a ser tratado. Os elementos considerados, em conjunto ou isoladamente, nas decisões de descarte são:

- **Concentração de postos.** A distribuição espacial dos postos permite verificar regiões densas e esparsas, buscando definir regiões onde é possível desconsiderar alguns postos;
- **Intervalo de anos e qualidade dos dados disponíveis.** A seleção final dos postos deve considerar aqueles cujos dados têm quantidade e qualidade adequadas, ou seja, que tenham dados coletados ao longo de vários anos e que sejam, em grande parte, consistentes;

- **Relevo.** A informação da altimetria dos postos provê uma noção do relevo de uma região, e auxilia na decisão de descarte de um posto pela comparação das altitudes relativas entre postos vizinhos.

A seleção inicial dos postos com base na localização (passo 1.1) pode ser feita de três formas. Na primeira, a seleção é feita pelos prefixos dos postos, aproveitando que cada posto é identificado por um prefixo formado por uma letra seguida de um número (C7, por exemplo), que identifica a sua região de localização, isto é, o posto está situado em uma área definida em uma grade bidimensional sobre o mapa, na qual a letra designa a linha e o número a coluna. Na segunda forma, a seleção é feita delimitando a região de interesse pelos valores de coordenadas de latitude e longitude. Na terceira forma, a seleção é feita visualmente, observando no mapa o posicionamento espacial dos postos, e demarcando diretamente no mapa a região de interesse. Os componentes funcionais disponibilizados na plataforma *InfoVis* permitem configurar ambientes para realizar a seleção por qualquer das formas descritas. A terceira forma foi a empregada e é detalhada a seguir.

O usuário seleciona os postos delimitando no mapa (componente ‘Distribuição Espacial’) a região que contém os postos de interesse. Na abordagem original, a região de estudo foi demarcada em um mapa impresso contendo a localização dos postos (Figura 5.1). Os componentes funcionais (‘Distribuição Espacial’ e ‘Intervalo/Categoria’) e as ações coordenadas usadas nessa forma de seleção são esquematizados na Figura 5.4. Na Figura 5.5 é exibida a interface principal com os elementos do ambiente configurado de acordo com esse esquema. As visualizações coordenadas permitem delimitar a região de interesse interativamente, observando a altitude dos postos (relevo), que é mapeada para cor, e a concentração (denso/esparso) de postos no componente ‘Distribuição Espacial’; e simultaneamente, a quantidade/qualidade das medidas dos postos, exibida no componente ‘Intervalo/Categoria’. A Figura 5.6 mostra uma região selecionada em ‘Distribuição Espacial’ (delimitada pelo retângulo), e a qualidade e quantidade de dados relativas aos postos selecionados destacados em ‘Intervalo/Categoria’ (linhas prolongadas em magenta). A facilidade da observação simultânea de diversos atributos e características, aumentando a riqueza de detalhe para a análise, conferiu ao usuário condições para decisões melhor fundamentadas.

Figura 5.4. Ambiente configurado para seleção visual de postos.

Figura 5.5. Interface da plataforma *InfoVis* com ambiente para seleção visual de postos.

Figura 5.6. Componentes funcionais para a seleção visual de postos.

Uma percepção decorrente das sessões de uso até este momento é que a tabela de cores para mapeamento dos atributos deve apresentar uma variação de cores capaz de facilitar a percepção das diferenças ou semelhanças entre os pontos, contribuindo para realçar padrões, por exemplo. Esse efeito está ilustrado na Figura 5.7, que mostra imagens produzidas com diferentes tabela de cores mapeando a altitude dos postos. As elipses e os retângulos identificam as mesmas regiões nos dois mapas. No ‘Mapa 1’, as variações de altitude entre os postos são mais evidentes que no ‘Mapa 2’ devido ao uso de uma tabela com cores de maior diferenciação entre os matizes. Quanto menor o ponto, mais notável é este efeito. Os pontos que representam os postos na atual implementação do componente ‘Distribuição Espacial’ podem ter dimensão variável controlada pelo parâmetro ‘escala’, que assume valores inteiros entre 1 e 4. Um ponto é representado por um único pixel se a escala é igual a 1, por uma matriz quadrada 2x2 de pixels se a escala é igual a 2, e assim por diante. Caso os itens de dados fossem representados visualmente por regiões, por exemplo, esse efeito talvez não fosse tão acentuado.

A tabela de cores é acessada pelo botão ‘Painel de Controle’ (ver Figura 5.7), sendo que o usuário pode carregar de um arquivo uma tabela salva anteriormente, ou pode definir a tabela. Embora o usuário especialista do domínio fosse familiarizado com o uso de cores, encontrar o mapeamento de cores mais adequado na visualização anterior exigiu diversas tentativas. Um conjunto de mapas de cores pré-definidos e sugestões de mapeamentos para situações específicas podem minimizar o esforço na definição de tabelas de cores [Tut 1998]. O ‘Painel de Controle’ permite, também, escolher o parâmetro ‘escala’ e informar os atributos de posição (latitude e longitude, no caso) dos postos.

Após a seleção inicial dos postos, o descarte de postos, descrito nos passos 1.2 e 1.3, é realizado ainda como parte dessa tarefa. O usuário verifica se alguns postos podem ser descartados, buscando diminuir o volume de dados a ser processado e evitar redundância. Um posto pode ser descartado quando:

- Suas medidas não oferecem contribuição relevante, pela quantidade ou pela qualidade;
- Existem postos próximos com grande quantidade de dados de boa qualidade na mesma faixa de altitude.

Figura 5.7. Influência da escala de cores na percepção da informação.

A Figura 5.8 ilustra como as visualizações espaço-temporais disponibilizadas na plataforma *InfoVis* apóiam o passo de descarte de postos. Observando a visualização exibida no componente ‘Intervalo/Categoria’ (a esquerda), o usuário percebe, pelo comprimento das linhas que indicam o intervalo temporal das medidas dos postos, que boa parte deles tem volume de dados suficiente. Percebe, também, que a faixa de anos com dados consistidos é coincidente na grande maioria dos postos, fato indicado pelo tom de azul mais escuro na região delimitada pelo retângulo amarelo. Essa constatação permite prosseguir sem analisar o início e o final do período de anos com dados consistidos, pois ao escolhermos dois postos com base em outra informação, provavelmente, os seus intervalos de anos com dados consistidos serão coincidentes.

Ainda considerando a Figura 5.8, foram utilizados três componentes do tipo ‘Distribuição Espacial’, para representar, respectivamente de cima para baixo, a altitude, a quantidade de anos com dados consistidos e a quantidade total de anos de coleta dos postos. Esses dois últimos atributos são os mesmos exibidos no componente ‘Intervalo/Categoria’, mas posicionados espacialmente e coloridos por faixas de valores. A tabela de cores que mapeia altitude está na parte superior da figura, e a tabela utilizada no mapeamento dos outros dois atributos (total de anos consistidos e total de anos de coleta) está na parte inferior. O valor ‘3’ que aparece na parte inferior das tabelas indica o fator de escala de exibição dos pontos. A região delimitada pelo retângulo vermelho é a que está visível nos três mapas. O retângulo verde engloba dois postos que estão em altitudes diferentes, e por isso, apesar de próximos, nenhum deles pode ser descartado; além disso, possuem grande quantidade de dados (observe o outro retângulo verde, as cores dos pontos indicam a faixa 20-30 e 30-40 de anos com dados coletados). Os retângulos brancos, superior e intermediário, mostram três postos que estão na mesma faixa de altitude e têm a mesma quantidade de dados consistidos, respectivamente. O atributo de decisão para descarte, nesse caso, é a quantidade de anos de dados coletados, que difere nos três postos, como pode ser observado no retângulo branco inferior. O posto com menor quantidade de dados é aquele delimitado pela elipse e, portanto, o eleito para descarte, considerando-se também que sua remoção não terá impacto sobre os demais postos fora do retângulo branco, pois a região tem boa quantidade de postos localizados adequadamente.

O exemplo anterior ilustra bem a capacidade de análise multidimensional proporcionada pelo uso de múltiplas visualizações. Além da faixa de valores dos atributos mapeados para cor,

podem ser percebidas a distribuição espacial dos postos e relações espaciais como concentração (denso/esparso) e o posicionamento relativo (próximo/acima/abaixo).

Figura 5.8. Componentes para suporte ao passo de descarte de postos.

Nesse segundo passo, a capacidade para um desenvolvedor acrescentar funcionalidades e, assim, adaptar um componente funcional para o domínio, foi importante. Essa capacidade é prevista pelo modelo *AdaptaVis* na definição de componentes independentes e auto-suficientes, e reforçada com a adoção da plataforma Java, que provê mecanismos que permitem acrescentar novos componentes ou versões atualizadas dos existentes com uma compilação desses módulos somente.

No decorrer das sessões percebeu-se a utilidade de acrescentar duas funcionalidades complementares ao componente ‘Tabela de Dados’ (Figura 5.9) para apoiar o descarte de postos; a primeira para salvar postos selecionados (botão ‘Salvar escolhidos’) e, a segunda para salvar os não selecionados (botão ‘salvar não escolhidos’). Certamente, o número de postos descartados é muito inferior ao de postos selecionados; dessa forma, é mais natural e eficiente selecionar os postos a descartar e salvar em arquivo os demais. Entretanto, é importante também manter em arquivo os dados dos postos descartados, pois pode-se concluir, posteriormente, que o conjunto de postos selecionados não foi adequado e, deve ser possível escolher entre os descartados aqueles que podem contribuir na formação do novo conjunto.

Figura 5.9. Componente funcional ‘Tabela de Dados’ após acréscimo de funcionalidades.

Em relação à abordagem original, a abordagem viabilizada pela plataforma *InfoVis* permitiu o tratamento simultâneo de mais atributos dos dados – distribuição espacial, intervalo de anos, altimetria –, o que é proporcionado pelo uso coordenado de múltiplas visualizações, oferecendo ao especialista uma visão macro e mais abrangente, levando a uma maior riqueza de informação já nesta etapa inicial de exploração. Na abordagem original boa parte do tempo e esforço dispendidos nessa tarefa foram dedicados à preparação dos dados, formatando-os e calculando alguns atributos – quantidade de anos com dados coletados e consistidos – que não estavam disponíveis nos dados brutos originais.

5.1.2. PREENCHIMENTO DE FALHAS

Após selecionar o conjunto de postos para processamento, o usuário verifica os dados de precipitação pluviométrica registrados em cada um dos postos selecionados, buscando inferir

medidas inexistentes ou corrigir as medidas aparentemente discrepantes. Tais valores podem ser preenchidos, ou substituídos por valores mais adequados, em uma abordagem que pode adotar dois caminhos: (1) procurar possíveis valores para preenchimento/substituição dentro do próprio posto, em períodos com padrão de comportamento chuvoso similar; ou (2) procurar valores em postos vizinhos cujo comportamento pluviométrico registrado seja semelhante ao do posto em questão. Para validar os valores inferidos, pode ser usada informação de postos de regiões distantes cuja configuração ou distribuição de altimetria, bem como o comportamento chuvoso, sejam semelhantes.

O ponto comum a essas duas abordagens é a busca por um comportamento chuvoso similar. Foram utilizadas as visualizações implementadas nos componentes funcionais ‘Distribuição Espacial’ e ‘Dados de Chuva’ (chuva diária, mensal e anual de um posto). Esta última representação é que permite comparar o comportamento de chuva inter e intrapostos. Como o número de postos comparados entre si não é grande, tipicamente três, os componentes podem ser posicionados na tela de modo a refletir o posicionamento espacial dos postos a que se referem. O componente ‘Distribuição Espacial’ fornece informações complementares, como altimetria, e uma percepção maior do posicionamento relativo entre os postos em comparação.

Na Figura 5.10 está sendo observado simultaneamente o comportamento pluviométrico de três postos, em diferentes instâncias do componente funcional ‘Dados de Chuva’. O posto C8-055 (prefixo no rótulo do componente) apresenta uma medida inexistente em 6/1979 (retângulo escuro no período destacado com bordas amarelas, ano na vertical e mês na horizontal), cujo preenchimento é tentado comparando-se os valores referentes à mesma data registrados nos postos C8-019 e C8-057, seus vizinhos acima e abaixo respectivamente. No mapa, é possível observar as altitudes, e os postos avaliados são destacados visualmente, depois de selecionados no componente ‘Tabela de Dados’. Na atual versão do *InfoVis*, apenas os componentes ‘Distribuição Espacial’ e ‘Tabela de Dados’ podem ter suas ações coordenadas, porque ambas representam o mesmo conjunto de dados (postos de coleta); os componentes ‘Dados de Chuva’ são ativados separadamente, sem relação com ações coordenadas. Uma generalização no formato das mensagens trocadas entre os componentes é uma alternativa para viabilizar a coordenação de ações mais genérica entre visualizações.

Figura 5.10. Avaliação da chuva em postos para preenchimento de falha.

Alternativamente, na Figura 5.11, os valores de precipitação são visualizados com um fator de escala maior e alinhados para facilitar a avaliação. Analogamente ao componente ‘Distribuição Espacial’, um fator de escala, variando de 1 a 4, permite controlar o tamanho do retângulo referente a um mês. O mês cujo valor deve ser preenchido está marcado com uma elipse, e o mês correspondente nos outros postos está marcado com um retângulo. Observando a distribuição das cores ao longo do ano de 1979, destacado em bordas amarelas, pode-se concluir que o comportamento da chuva é similar nos postos, incluindo o total anual, representado pelo mesmo tom de azul na última linha de cada posto. A tabela de cores na parte superior da Figura 5.11 é a que foi utilizada para mapear os totais mensais e, a outra tabela mapeia os valores dos totais anuais. Infere-se, ainda, que o valor adequado a ser atribuído está na faixa representada pela cor vermelha, isto é, entre 0 mm e 100 mm. Para essa tarefa, embora valores diários de

chuva estivessem disponíveis, foram considerados os totais mensais e anual, com ênfase no primeiro, devido à maior familiaridade e ao conhecimento acumulado pelo usuário sobre os mesmos.

Figura 5.11. Disposição alternativa para avaliação do comportamento de chuva.

Na abordagem original, foram utilizadas tabelas numéricas individuais com os totais mensais de chuva para cada posto. Essas tabelas foram geradas a partir do aplicativo BcDAEE (ver Apêndice A), e eram posicionadas sobre a mesa para refletir o posicionamento relativo dos postos, e assim, a avaliação era realizada comparando os valores numéricos.

Um componente funcional ‘Tabela de Dados’ permitiria verificar o valor numérico exato de uma medida de chuva. Entretanto, o especialista do domínio optou por prosseguir a análise somente no modo visual, utilizando a escala de cores para decidir qual faixa de valores melhor se encaixaria para substituir um valor inexistente. Isso pode ser um indicativo de que o modelo

mental do usuário formado a partir das visualizações interativas oferece condições para uma inferência adequada e com grau de confiança. A decisão tomada pela inspeção visual foi, posteriormente, referendada por uma consulta às tabelas com dados de chuvas dos postos envolvidos. Com base nessa observação, uma funcionalidade útil de ser implementada é permitir salvar a imagem gerada com o valor ausente preenchido com a cor que representa o intervalo escolhido, juntamente com os registros de histórico.

O componente funcional ‘Dados de Chuva’ mostrou-se adequado para apoiar a análise do comportamento local de chuva em torno do período cujo valor deve ser inferido, o que é útil para o preenchimento de falhas. A escolha de um mapeamento de cores é determinante para o sucesso da tarefa. Entretanto, não mostrou-se eficiente para analisar o comportamento de chuva global de múltiplos postos, que seria útil para determinar qual o posto mais chuvoso ou seco. Nesse caso, os valores diários de chuva contribuíram para a análise global, evidenciando a utilidade da visualização simultânea das medidas em múltiplos níveis de agregação (diário, mensal e anual) disponibilizado pelo componente.

A Figura 5.12 mostra uma configuração que exhibe os dados de chuva dos postos C8-055, D7-075 e D7-066, incluindo valores diários de chuva (parte superior das imagens, com predominância de tons vermelhos). No canto inferior esquerdo, estão as tabelas de cor para os valores diário (esquerda), mensal (meio) e anual (direita) de chuva. Para os totais mensais foi definida uma escala de cor monocromática azul, após testes com outras escalas de cor. A idéia é que o brilho indique o comportamento da chuva, entretanto a observação apenas da escala mensal não permite uma constatação eficiente sobre o comportamento do posto. Contudo, examinando também os dados das escalas diária e anual, é possível perceber que o posto C8-055 é mais seco que os demais.

Na atual implementação do *InfoVis*, cada componente controla a sua tabela de cores. As sessões de uso mostraram que a definição de um componente para manipulação de tabelas de cores traria benefícios para usuário. O componente ‘Tabela de Cores’ e os demais componentes poderiam trocar informações por meio de ações coordenadas; entretanto, isso requer, novamente, um formato de mensagem mais geral ou flexível do que o definido atualmente. Como essa flexibilização, certamente, cria condições para a implementação de tipos variados de ações de interação, sendo uma melhoria prioritária do modelo *AdaptaVis*.

Figura 5.12. Comparação do comportamento global de chuva dos postos.

Ainda na Figura 5.12, na marcação de pontos selecionados no mapa é importante não esconder ou dificultar a visualização das informações mapeadas para cor. Por isso, os marcadores devem ser discretos, como linhas laterais aos pontos no lugar de quadrados envolvendo-os. Os postos, cujos dados de chuva estão mostrados, são marcados com uma discreta linha vertical em ciano à esquerda do ponto. O atributo mapeado para cor, segundo a tabela na parte superior da figura, é altitude.

5.1.3. DETERMINAÇÃO DE ANOS REPRESENTATIVOS

Considerando o total anual de chuva em todos os postos selecionados na região oeste do Estado de São Paulo, com os dados devidamente preenchidos ou substituídos, o objetivo dessa tarefa é definir três anos representativos para ano seco, ano chuvoso e ano habitual. Para realizar essa

tarefas foram utilizados uma generalização do componente ‘Intervalo/Categoria’, adaptando o diagrama de dispersão de GGobi (ver Capítulo 2), e o componente ‘Distribuição Espacial’.

Na abordagem original, foi utilizado um algoritmo de agrupamento, cujo resultado foi apresentado graficamente como árvores de ligação para cada posto (dendogramas), as quais foram utilizadas pelo especialista para selecionar os anos representativos (ver Figura A.2 no Apêndice A). A escolha final recaiu sobre os anos 1983, 1984 e 1985, como representativos de ano chuvoso, habitual e seco, respectivamente. A generalização do componente ‘Intervalo/Categoria’ (Variação Temporal Uni-escala) permitiu validar essa escolha e verificar que outras escolhas seriam possíveis. A Figura 5.13 exibe três visualizações do tipo ‘Distribuição Espacial’ em que a cor mapeia os totais anuais de chuva para cada posto em um determinado ano, juntamente com o ‘Intervalo/Categoria Generalizado’ que exibe os totais dos postos em diferentes anos. Nessa última representação visual, cada linha se refere a um ano (ordenação crescente de baixo para cima) e cada ponto colorido corresponde a um posto. Nesse caso, olhar para uma linha equivale a olhar as árvores de ligação de todos os postos simultaneamente. A tabela de cor que mapeia os valores totais anuais de chuva nos mapas é exibida na parte superior da figura, o mesmo mapeamento de cores foi definido no aplicativo GGobi. Analisando as linhas contidas no retângulo branco, relativas aos anos 83, 84 e 85, e comparando-as entre si, pode-se inferir que 83 foi um ano *chuvoso* pela predominância das cores verde e amarela, 84 foi um ano *habitual* pela abundância da cor laranja, e 85 foi um ano *seco* pela maior presença das cores laranja e vermelho. Esta constatação pode ser comprovada pelas visualizações nos três mapas, que mostram uma diminuição do total de chuva de 83 para 85, de forma geral, na região em estudo. O retângulo azul mostra um possível conjunto alternativo de anos representativos.

A formação das classes pela árvore de ligação pode ser controlada pela escolha de níveis de corte, que refletem a perda de detalhe de informação para as classes. Esse controle pode ser simulado, na visualização ‘Intervalo/Categoria Generalizado’, variando a tabela de cores, isto é, os intervalos de valores mapeados para cor. Na imagem à esquerda na Figura 5.14, a identificação de padrões para a determinação de anos representativos é possível, mas é dificultada pela perda de detalhes em relação à imagem à direita, em virtude do intervalo de valores mapeado em uma única cor ser maior, 400 contra 200 da imagem à direita. O tamanho de cada intervalo controla a ocultação ou evidência de detalhes do atributo mapeado, e é variável em função do conjunto de valores presente nos dados. Isso reforça a necessidade de uma escolha

cuidadosa de mapas de cores em visualização, como já reforçado por diversos autores da área [Rog 2001].

Figura 5.13. ‘Distribuição Espacial’ e ‘Intervalo/Categoria Generalizado’ na escolha de anos representativos.

Figura 5.14. Ganho de detalhe controlado pela escala de cor.

O potencial das visualizações interativas em viabilizar e provocar uma participação mais ativa do usuário, apoiando o processo cognitivo, pôde ser evidenciada durante as sessões conduzidas com o especialista [Shi 2004][Shi 2003]. Na busca por uma escala de cores adequada para mapear o total de chuva no componente ‘Distribuição Espacial’, conhecimentos prévios do usuário, não diretamente relacionados ao objetivo da análise (comparação de total de chuva em anos distintos), vieram explicitamente à tona e modificaram o objetivo imediato da análise. O mecanismo de registro de histórico mostrou-se bastante útil nessas situações.

A Figura 5.15 ilustra um esquema no qual são empregados dois mapas, o da esquerda exibe o total de chuva em cada posto, e o da direita exibe altitude. Testes com tabelas de cores para o mapa da esquerda motivaram um raciocínio para explicar as quantidades de chuva percebidas, que levaram à ativação do componente da imagem a direita exibindo a altimetria para a região correspondente. Inicialmente, percebeu-se duas sub-regiões, A e B, com quantidades de chuva diferentes, e buscou-se uma explicação na altimetria, que influencia o deslocamento de massas de ar. De fato, a sub-região A é mais baixa que a sub-região B. Entretanto, observou-se que, em B, a parte inferior tem altimetria semelhante à de A, mas apresenta uma boa quantidade de chuva. Isso levou a definição de uma nova sub-região, C, e a explicação para a quantidade de chuva pode ser um maior volume de água devido à presença de um rio. As informações sobre deslocamento de massas de ar e presença do rio não estavam disponíveis no conjunto de dados para visualização, mas faziam parte do conhecimento do especialista.

Figura 5.15. Múltiplas visualizações como suporte ao processo cognitivo.

5.2. EXPLORAÇÃO DE DADOS DE TRANSAÇÕES COMERCIAIS

Dados reais de transações comerciais ofereceram um outro domínio para validar a plataforma *InfoVis* e as visualizações temporais, aproveitando as facilidades de extensão para domínios variados, um dos requisitos do modelo *AdaptaVis*, e contemplando os objetivos 2 (verificar a utilidade das visualizações temporais) e 3 (verificar a extensibilidade da plataforma), citados no início deste Capítulo. A origem dos dados é uma indústria do setor alimentício, que em função da atividade em diversas regiões do país, tem nos atributos temporal e espacial das transações registradas uma importante fonte de informação de apoio à tomada de decisões. Além dos atributos de data e local, em uma transação são registrados os diversos valores envolvidos (total da transação e taxas, por exemplo). Nesse caso, o especialista em visualização também participou ativamente das sessões de uso, mas o cenário da participação de usuários é diferente daquele das sessões com os dados de chuva. Vários usuários atuaram nas sessões, com um grupo, mais ligado à tecnologia da informação, explorando os dados e produzindo resultados iniciais, e outros grupos de usuários finais (gerentes e supervisores dos setores financeiro e comercial, por exemplo) avaliando e refinando os resultados.

Os componentes funcionais ‘Distribuição Espacial’ e ‘Tabela de Dados’ foram usados, de forma similar às sessões com dados de chuva, para verificar a distribuição espacial de um atributo, bruto ou calculado, mapeado para cor. Nesse caso, a única adaptação introduzida foi o uso, na interface principal, de termos específicos mais familiares aos grupos de usuários. Essa adaptação requereu apenas a edição do arquivo de configuração *menu.cfg* (ver Figura 5.3).

Aqui, também foi utilizada a generalização da visualização ‘Intervalo/Categoria’, com o gráfico de dispersão do aplicativo GGobi. A generalização segue os mesmos princípios da representação ‘Variação Temporal Uni-escala’ (apresentada na Subseção 3.1.1, ver Figura 3.2), contudo com uma organização diferente da adotada no domínio climatológico. A unidade de variação temporal (associada ao eixo horizontal) é diária, e os itens de dados, dispostos no eixo vertical, podem representar cidades ou produtos. Na Figura 5.16, cada linha representa um produto e cada ponto na linha corresponde a um determinado dia. A cor do ponto mapeia um atributo escolhido pelo usuário, por exemplo, o número de transações ou o valor monetário. A distribuição das cores no espaço do gráfico (tons azuis na base, por exemplo, indicam valores maiores), e a concentração/dispersão de pontos (esparsos na parte superior) fornecem

importantes informações para os grupos de usuários. A ausência de cor no gráfico significa que não houve transação de um determinado produto em um dado dia (o que pode ser uma ocorrência normal; um feriado, por exemplo). Essas sessões evidenciam a utilidade das visualizações em múltiplos domínios, corroborando a necessidade de generalização do componente funcional ‘Intervalo/Categoria’ para aumentar a sua aplicabilidade e, juntamente com as apreciações das sessões com dados de chuva, sugerem adaptações, como deixar variável a unidade temporal, por exemplo, que foi anual para a sessão de uso em climatologia e diária neste caso. A facilidade de aplicação das técnicas em domínios variados, viabilizada pela facilidade de adaptação proporcionada na plataforma *InfoVis*, cria condições para apoiar o refinamento dos componentes funcionais. Dadas as características do modelo *AdaptaVis*, esses refinamentos podem ser implementados com baixo impacto no ambiente global, desde que o código fonte do componente funcional esteja disponível.

Figura 5.16. Variação diária de transações de produtos.

O componente funcional ‘Dados de Chuva’ também foi utilizado com algumas pequenas adaptações, obedecendo aos princípios da representação ‘Variação Temporal Multi-escala’, apresentados na Subseção 3.1.2. Nos testes com dados de chuva, a visualização gerada por esse componente referia-se a um posto, cada coluna indicava um ano e o atributo de interesse (precipitação) era mapeado para cor de acordo com valores diário, mensal e anual. No segundo domínio, transações comerciais, a divisão em valores diário, mensal e anual permanece, mas

cada coluna representa uma cidade ou um produto associados a uma transação, e a visualização se refere a um determinado ano (ver o diagrama 3.5-b). A Figura 5.17 ilustra uma visualização, que representa o valor de transações (escalas diária, mensal e anual) em várias cidades (colunas) em um dado ano. A tabela de cores definida varia, em ordem crescente nos intervalos de valores, do vermelho escuro para o vermelho claro, branco, e depois do azul claro para o azul escuro, que é a mesma escala (pré-definida no GGobi) utilizada na generalização de 'Intervalo/Categoria'. Os intervalos nos extremos da tabela de cor têm cores escuras, vermelhas para o extremo inferior e azuis para o extremo superior; os intervalos intermediários têm cores mais claras, com a cor branca dividindo as faixas de cor.

A análise dessa representação permitiu aos usuários conceber um novo esquema de classificação para as cidades com base na frequência e volume de transações, buscando definir classes de comportamento para as cidades mais adequadas para apoiar o estabelecimento de metas. Anteriormente, o esquema de classificação utilizado era baseado apenas no valor total anual, que na figura é observado na última linha (parte inferior da imagem). Na Figura 5.17, na primeira coluna (marcações em laranja à esquerda) pode-se observar um total anual (elipse laranja) na faixa de maior valor (cor azul escuro), e analisando os valores diários nos 12 retângulos superiores (destacados em laranja), percebe-se que a frequência e os valores nas transações são altas (abundância de pontos e predominância da cor azul escuro, respectivamente). Entretanto, caminhando para a direita na figura, encontramos na oitava coluna (marcações em laranja à direita) um valor total também na faixa da cor azul escuro, mas o comportamento dos valores diários é bastante diferente, sem predominância da cor azul escuro, indicando faixas de valores menores, e também uma frequência menor. Essa constatação mostra claramente que o valor total, isoladamente, dá uma visão incompleta dos dados. Foi concebido, então, um esquema de classificação considerando o valor total, a frequência de ocorrência de transação e os valores dessas transações. Tal esquema está sendo testado, mas ainda não há resultados conclusivos sobre sua eficiência.

Figura 5.17. Transações de produtos em diversas cidades em um determinado ano.

A participação de diversos grupos de usuários propiciou um processo colaborativo, e percebeu-se que as visualizações geraram, inicialmente, interpretações diversas nos diferentes usuários, fato causado, provavelmente, pelas diferentes experiências individuais dos participantes e/ou pela falta de compreensão da construção de algumas representações visuais. Após uma breve introdução às representações, houve uma convergência das interpretações. A interatividade e o uso de múltiplas visualizações permitiram a percepção de padrões e, principalmente, a identificação de exceções seguidas das respectivas razões, com cada participante explicitando seu conhecimento prévio para os demais. Por exemplo, esperava-se, por hipótese, que os valores das transações em uma cidade estivessem relacionados com sua população, no entanto, a visualização não comprovou tal hipótese. Uma avaliação mais minuciosa, envolvendo especialistas do setor comercial da empresa, concluiu que características dos vendedores (experiência profissional, forma de atuação, tempo de vínculo com a empresa) e o tamanho da equipe de vendedores auxiliavam a explicar a visualização. Além disso, foram identificados grupos de cidades que mereceriam formas diferenciadas de atuação por parte da empresa.

As sessões de uso nesse domínio de aplicação também evidenciaram a utilidade do uso simultâneo de múltiplas representações visuais em tarefas exploratórias. A seguir, são reunidas as conclusões sobre as sessões de uso em ambos os domínios.

5.3. CONSIDERAÇÕES FINAIS

As sessões de visualização, com a execução de tarefas envolvendo dados reais e usuários finais nos dois domínios de aplicação, permitiram validar as características do modelo *AdaptaVis*/plataforma *InfoVis* e das representações visuais temporais. A participação de usuários finais com diferentes perfis foi um balizador importante no processo, principalmente, por se tratar de usuários reais. Nesta Seção são destacadas as principais conclusões das sessões de uso.

A possibilidade de criar um ambiente personalizado configurando um simples arquivo texto, que permite selecionar os componentes funcionais adequados, e acrescentá-los/removê-los dinamicamente, mostrou-se um importante catalisador para a adaptação dos usuários finais, em função do ambiente conter somente os componentes necessários ao problema sendo tratado. Em ambos os domínios, a simplicidade das representações visuais utilizadas, inspiradas nas técnicas orientadas a pixel, facilitou sua interpretação, o que permitiu aos usuários vislumbrar os potenciais de uso, e também contribuiu para uma boa receptividade e, portanto, para o bom andamento das sessões de uso. Esses fatores somados criaram as condições necessárias para a motivação e participação ativa dos usuários.

Durante as sessões detectou-se a necessidade de acrescentar funcionalidades a componentes funcionais, que tecnicamente é equivalente a criar novos componentes. No caso dos dados climáticos, o componente do tipo ‘Tabela de Dados’ recebeu duas novas funcionalidades, uma para gravar registros selecionados em arquivo, e outra para gravar os não selecionados. Essa solução foi possível porque o código fonte estava disponível; caso contrário, um novo componente funcional independente poderia ser implementado e acrescentado à plataforma, e as novas funcionalidades poderiam ser acessadas configurando ações coordenadas entre os componentes envolvidos. Mesmo com o código acessível, a opção por criar um novo componente seria uma alternativa para evitar o problema de efeitos colaterais ao manipular código pronto. Analogamente, a adaptação do componente ‘Dados de Chuva’ para a exploração de transações comerciais seguiu os mesmos procedimentos.

Ainda, com relação à adaptação de componentes, percebeu-se durante as sessões com usuários que a definição de tabelas de cores poderia ser tratada em um componente funcional independente, sendo que os mapas de cores seriam acessíveis por meio de ações coordenadas.

Isso evitaria a repetição na definição de mapas, e encorajaria o reuso de bons mapas. Entretanto, no atual estágio de implementação do *InfoVis*, as mensagens para o gerenciamento de ações coordenadas possuem como parâmetros somente os índices dos itens de dados a serem processados. Um modelo genérico para o processo de coordenação viabilizaria esse tipo de adaptação. A importância e a utilidade desse modelo genérico já foram apontadas na evolução do *Snap*, e por outras propostas, citadas no Capítulo 3, mais abrangentes. Tal generalização provê suporte para a criação de componentes funcionais melhor delimitados, na medida em que permite, por exemplo, seguir a categorização dos componentes funcionais em tipos, apresentada no Capítulo 3 (Seção 3.2), que busca definir claramente o escopo de cada componente.

As representações visuais temporais foram aplicadas com sucesso, e a integração no esquema de múltiplas visualizações interativas coordenadas apoiou efetivamente as tarefas de exploração de dados espaço-temporais [Shi 2003][Shi 2004], como ilustrado na execução das tarefas nos dois domínios de aplicação. Todavia, as formas como as visualizações são coordenadas varia de acordo com os objetivos específicos de uma tarefa, ainda que as representações visuais utilizadas sejam as mesmas. O recurso de configuração de ações coordenadas em tempo de execução foi bastante utilizado, tanto nas sessões com dados climáticos quanto nas sessões com dados comerciais. O recurso apóia o gerenciamento de diferentes tarefas em uma mesma sessão de uso do ambiente, estabelecendo apenas as coordenações necessárias.

Além da já mencionada interpretação intuitiva, a abordagem baseada em pixels utilizada é de simples implementação, o que se reflete positivamente no desempenho. As técnicas não requerem algoritmos elaborados de ordenação dos itens de dados e posicionamento dos pixels, nem algoritmos complexos de renderização, devido à ordenação temporal natural dos itens e à simples atribuição de cores aos pixels, respectivamente. O mapeamento de cores, intrínseco à abordagem, mostrou-se eficiente em tarefas que envolviam comparações locais, como, por exemplo, a comparação de medidas de chuva em torno de um determinado mês na tarefa de preenchimento de falhas (ver Figuras 5.10 e 5.11); entretanto, o uso de cores não se mostrou adequado para uma comparação global, tal como determinar qual o posto mais chuvoso analisando um longo período. Tal inadequação, nesse caso, foi minimizada pela exibição simultânea de várias escalas de tempo (ver Figura 5.12).

O caráter interativo e iterativo do processo exploratório no qual hipóteses são criadas e verificadas, leva ao estabelecimento de muitos cenários ou configurações. Nesse processo, o caminho é caracterizado mais pela aleatoriedade do que pela linearidade, com as descobertas dirigindo a construção de novos cenários. No *InfoVis*, o registro dos componentes utilizados em uma sessão, juntamente com a configuração de ações coordenadas e as ações realizadas, possibilita que os diferentes cenários sejam registrados e recuperados posteriormente. Esse instrumento foi utilizado, em ambos os domínios, para verificar estratégias adotadas e traçar novas, e mostrou-se útil nas situações em os períodos entre as sessões eram bem espaçados e, também, para evitar repetições de atividades. O registro dos cenários nas sessões permite, também, extrair informações para o refinamento de componentes, da plataforma e do próprio modelo.

Neste Capítulo foram apresentados os detalhes das sessões de uso com dados reais e usuários finais especialistas em dois domínios de aplicação, Climatologia e Negócios. As sessões visaram validar o modelo *AdaptaVis*, a plataforma/ambientes *InfoVis* e representações visuais temporais. No próximo capítulo, conclusões gerais e trabalhos futuros são apresentados.

CONCLUSÕES

O foco inicial deste projeto de doutorado foi o de propor técnicas de visualização para dados com atributos temporais e espaciais, e verificar como essas técnicas poderiam apoiar usuários em tarefas de exploração na etapa de pré-processamento dos dados em processos de descoberta de conhecimento nesse contexto. Em capítulos anteriores, foi ressaltada a importância da extração de informação a partir de dados espaço-temporais, quer pelo crescente volume de dados dessa natureza, quer pela sua aplicabilidade no suporte à tomada de decisão em diversos domínios. Além desses fatores, um forte motivador foi a existência de uma situação real para aplicação, isto é, o acesso a uma base de dados com medidas reais de chuva para processamento, e a perspectiva de colaboração com um usuário especialista em Climatologia, familiarizado com os dados. A complexidade associada ao tratamento simultâneo de vários atributos dos dados e a dificuldade em analisar todo o volume de dados necessário, identificados pelo usuário em processamentos anteriores, eram indicativos de que técnicas de visualização seriam úteis no processamento dos dados. Após uma resposta positiva do usuário em estudos preliminares, utilizando algumas visualizações experimentais adaptadas de técnicas tradicionais, a meta inicial era conceber técnicas de visualização para apoiar as tarefas do usuário e disponibilizá-las em um sistema de visualização orientado a dados com atributos temporais e espaciais.

Entretanto, experimentando alguns sistemas que implementam técnicas de visualização de dados multidimensionais e apóiam o uso coordenado de múltiplas visualizações, percebeu-se a necessidade de maior flexibilidade para o usuário adaptar o ambiente aos seus requisitos pela escolha das visualizações adequadas e configuração das coordenações entre elas. Notou-se que seria desejável incorporar ao sistema um mecanismo que permitisse a incorporação gradual de novas técnicas de visualização e a sua integração ao esquema de coordenação. Essa flexibilização possibilita que o alcance do sistema de visualização seja ampliado para diversos tipos de usuário, domínios de aplicação e tarefas de análise, provendo a esses sistemas um potencial maior de uso e crescimento, e proporcionando uma base de desenvolvimento mais sólida para os pesquisadores na área de visualização de informação e mineração visual de dados.

Sistemas de visualização descritos recentemente na literatura – alguns inclusive focados no tratamento de dados com atributos temporais e espaciais – têm buscado as características de flexibilidade citadas acima. O papel da coordenação entre múltiplas visualizações em processos exploratórios também tem ganhado destaque, merecendo uma conferência específica, a CMV – *Coordinated and Multiple Views in Exploratory Visualization*¹ –, que está em sua segunda edição em 2004. Nas soluções descritas na literatura, fica evidente a importância de uma padronização (para viabilizar a incorporação de novas técnicas, por exemplo) e o esforço de formalização (na definição de um modelo para a coordenação entre múltiplas visualizações, por exemplo). Assim, um modelo geral que identifique e defina os elementos de uma plataforma de software adaptável, provendo diretrizes para a implementação de uma solução computacional extensível, é essencial para permitir a expansão gradual dos ambientes, bem como para viabilizar futuramente uma avaliação controlada dos diversos aspectos e alternativas envolvidos no uso de técnicas de visualização (assim como de técnicas analíticas) em processos exploratórios e de descoberta de conhecimento.

Dessa forma, à proposta inicial de concepção e uso de técnicas de visualização para dados espaço-temporais somou-se a elaboração de tal modelo, e sua validação com a implementação de uma plataforma extensível quanto ao conjunto de técnicas visuais disponíveis, e flexível quanto à coordenação entre múltiplas técnicas. Assim, esta tese gerou duas contribuições principais:

¹ www.cvev.org/cmv2004/

- **A proposta de duas representações visuais temporais**, ‘Variação Temporal Uni-escala’ e ‘Variação Temporal Multi-escala’, inspiradas nas técnicas de visualização orientadas a pixel, aproveitando o espaço de visualização e a ordenação natural dos dados para exibir a distribuição, ao longo do tempo, dos valores de atributos pelo mapeamento de cor. Na primeira representação, é utilizada uma única escala temporal e características como comprimento, deslocamento, sobreposição e continuidade do intervalo temporal podem ser percebidas. Três escalas temporais agregadas são permitidas na segunda representação e, frequência, concentração e repetição de eventos temporais estão entre as características que podem ser observadas. Essas representações são aplicáveis em diferentes domínios envolvendo o tratamento de dados temporais, como pôde ser verificado por meio de sua aplicação, em colaboração com usuários e dados reais, para apoiar efetivamente tarefas de análise em dois domínios distintos: dados de precipitação pluviométrica, e dados de transações comerciais;
- **A concepção de um modelo de arquitetura de software**, o *AdaptaVis*, que norteou a implementação da plataforma *InfoVis*, que permite configurar, a partir de um conjunto de técnicas de visualização, ambientes adaptados aos requisitos de usuários e tarefas. Essa plataforma pode ser estendida com facilidade pela incorporação de novos componentes ou atualização dos existentes; a configuração de ações coordenadas é flexível e pode ser estabelecida pelo usuário em tempo de execução; e as ações do usuário realizadas ao longo da configuração e uso do ambiente são monitoradas e registradas. Essas características suportam uma evolução gradual da plataforma, que é útil para diversos tipos de usuário em vários domínios de aplicação. A escolha da ferramenta para implementação – Java – foi direcionada pelos requisitos de ‘extensibilidade’, ‘configurabilidade’ e ‘monitorabilidade’, determinados na definição do modelo *AdaptaVis*. Essa opção amplia, naturalmente, o leque de características, provendo também um grau de independência de sistema operacional e de tipos de equipamentos (computadores de mesa e estações de trabalho, por exemplo).

As sessões de uso com a plataforma *InfoVis* permitiram validar o modelo *AdaptaVis* e verificar a aplicabilidade das duas representações visuais temporais no tratamento de dados em dois domínios de aplicação. Resultados dos estudos de uso, alguns deles compatíveis com outros documentados na literatura, merecem destaque:

- A necessidade de treinamento prévio dos usuários para a compreensão das visualizações e correto uso da plataforma;
- A simplicidade das visualizações conferida por fatores tais como ordem natural dos dados, imagem bidimensional e mapeamento só para cor, foi importante para facilitar a explicação da visualização e sua interpretação visual, o que melhorou a receptividade pelos usuários;
- A simplicidade também afeta positivamente o desempenho, exigindo menor esforço computacional para a geração da imagem;
- O histórico de atividades (*log*) pode ser uma importante fonte de informação, de grande utilidade no apoio à interação entre o especialista no domínio de aplicação e o especialista em visualização;
- A importância de ter disponível uma plataforma que permita avaliar aspectos do uso das técnicas, em contextos diversos: uso de uma técnica de visualização específica em tarefas exploratórias, esforço cognitivo associado ao uso de múltiplas visualizações coordenadas e integração entre técnicas visuais e analíticas, dentre outras. Especificamente, o desenvolvimento deste trabalho permitiu verificar vantagens e desvantagens da abordagem baseada em pixels para apoiar a exploração de dados com atributos temporais e espaciais, como discutido no capítulo anterior.

Embora neste projeto a ênfase tenha sido na disponibilização de recursos de Visualização Exploratória, a solução concebida é suficientemente genérica para acomodar também técnicas de outras disciplinas, como Banco de Dados, Estatística, Inteligência Artificial e Interface Usuário-Computador, contemplando a natureza multi e interdisciplinar da descoberta de conhecimento, conforme ilustra a Figura 6.1. Uma instância futura do *AdaptaVis* deve oferecer recursos para a integração e coordenação de recursos visuais, analíticos e de interação, atendendo, de forma personalizada, aos requisitos de usuários, tarefas e domínios de aplicação.

Figura 6.1. Enfoque inter e multidisciplinar do modelo *AdaptaVis*.

Com base nos resultados alcançados, pode-se delinear alguns tópicos de continuidade para o aperfeiçoamento do modelo:

- **Modelo genérico de ações coordenadas.** No formato das mensagens definido atualmente no modelo, as informações transmitidas contêm a identificação da ação executada ou a ser executada e os índices dos itens de dados afetados, o que restringe o conjunto de ações que podem ser coordenadas. É necessário que um formato mais genérico seja definido, de forma a possibilitar que diversas outras ações de interação possam ser disponibilizadas, seguindo um modelo mais amplo de coordenação, como o proposto no âmbito do *Snap* [Nor 2001], ou modelos ainda mais genéricos, centrados não somente nos dados [Bou 2003][Pat 2001]. Por exemplo, uma mensagem na qual o conteúdo de seus parâmetros possa variar e a semântica desse conteúdo seja acessível possibilita ampliar os tipos de ações que podem ser coordenadas. A descrição semântica deve ser externa aos componentes, para evitar a necessidade de atualização dos componentes a cada definição de conteúdo;
- **Esquema colaborativo entre componentes funcionais.** Um formato de mensagem generalizado também cria condições para estabelecer um esquema de colaboração, bastante útil especialmente em tarefas exploratórias. Componentes funcionais associados a Núcleos Básicos diferentes poderiam ter suas ações coordenadas pela troca mensagens entre os seus respectivos gerenciadores de Componentes e Ações, operando em sincronia;
- **Personalização das interfaces.** Para diferentes grupos de usuários ou tipos de tarefas, pode ser interessante oferecer diferentes organizações da interface. Por exemplo, as

janelas das visualizações podem ser independentes ou estar contidas dentro de uma janela principal, ou mesmo interfaces diferentes para desenvolvedores e usuários finais. Assim, a própria interface principal poderia ser implementada como um componente funcional, permitindo que um mesmo Núcleo Básico comportasse diversas interfaces;

- **Extração de informação a partir dos históricos.** O registro de históricos é uma importante fonte de estudo de aspectos relacionados com a realização de tarefas com suporte de visualização, e uso integrado e coordenado de múltiplas visualizações, permitindo extrair informações para o aperfeiçoamento do modelo, adequação de componentes funcionais e aprimoramento de diretrizes sobre uso e desenvolvimento de soluções. Para facilitar a extração dessas informações, o registro de históricos deve ser estruturado de tal forma que a informação possa ser acessada com diversos níveis de detalhamento e em formatos apropriados;
- **Núcleo Básico adaptável.** As funções desempenhadas por cada um dos gerenciadores do Núcleo Básico também podem variar de acordo com modelos associados que podem ser implementados. Um Núcleo Básico adaptável abre a possibilidade de testar esses modelos através de uma única plataforma. Nesse tópico e nos anteriores, torna-se evidente que a flexibilidade no conteúdo das mensagens é essencial para viabilizar as várias alternativas.

Um desdobramento deste projeto foi o desenvolvimento de uma dissertação de mestrado [Bra 2003] que propõe outras estratégias de visualização de dados temporais, inspiradas em algumas técnicas de visualização utilizadas nos estudos preliminares com dados de chuva, as quais estão sendo incorporadas à plataforma *InfoVis*. Outro desdobramento interessante é que o modelo *AdaptaVis* serviu de base para a concepção do projeto ‘*Distributed, Linked Visualization and Sonification on a Peer-to-Peer, Web Services (P2P-WS) Architecture*’, cuja proposta foi encaminhada ao *National Science Foundation* (NSF), nos Estados Unidos, pelo professor Haim Levkowitz da *University of Massachusetts Lowell*. A proposta do projeto é implementar uma arquitetura que disponibilize técnicas de visualização e sonificação via tecnologia de *Web Services*.

Finalmente, espera-se que o desenvolvimento deste projeto tenha criado condições para o desenvolvimento evolutivo e sob demanda de soluções adaptadas para usuários, tarefas e domínios de aplicação no processo de descoberta de conhecimento, possibilitando que soluções mais localizadas contribuam para a elaboração e aperfeiçoamento de soluções mais gerais e

abrangentes, e contemplando a multidisciplinaridade. Como primeira tese de doutorado com foco em Visualização de Informação no Grupo de Computação Gráfica e Processamento de Imagens do ICMC/USP-São Carlos, espera-se que a plataforma *InfoVis* constitua um instrumento de pesquisa para grupo, como fonte geradora de informação para a construção de diretrizes, estabelecendo pontos fortes e fracos, vantagens e desvantagens, e precauções no uso de diferentes instrumentos e estratégias em tarefas de descoberta de conhecimento.

REFERÊNCIAS BIBLIOGRÁFICAS

- [Ahl 1992] Ahlberg, C.; Williamson, C.; Shneiderman, B. - Dynamic Queries for Information Exploration: An Implementation and Evaluation, **Proceedings ACM Conference on Human Factors in Computer Systems**, pp. 619-626, 1992
- [And 1999] Andrienko, G.; Andrienko, N. – Knowledge-Based Visualization to Support Spatial Data Mining, **IDA'99 LNCS 1642**, Hand, D.L., Kok, J.N., Berthold, M.R. (eds.), pp. 149-160, 1999
- [And 1999a] Andrienko, G.; Andrienko, N.; Voss, H. – Thematic Internet Maps for Cities: Data Analysis and Publishing with Descartes, **Proceedings of 21st UDMS'99 Conference**, pp. XII.1.1-XII.1.11, Urban Data Management Society, Venice, 21-23 April, 1999
- [And 2000] Andrienko, G.; Andrienko, N. – Java Tools for Exploratory Cartographic Visualization in the Internet, **The 10th International Conference on Computer Graphics & Vision (GRAPHICON'2000)**, pp. 18-25, Moscow State University, Russia, August,28 – September,2, 2000
- [Ank 1999] Ankerst, M.; Elsen, C.; Ester, M.; Kriegel, H.-P. – Visual Classification: An Interactive Approach to Decision Tree Construction, Conference on Knowledge Discovery in Data, **Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, pp. 392-396, San Diego, CA, 1999
- [Ank 2000] Ankerst, M. – **Visual Data Mining**, PhD Thesis, Faculty of Mathematics and Computer Science, University of Munich, August 2000
- [Aou 1983] Aouad, M.S. – **Tentativa de Classificação Climática para o Estado da Bahia (Uma análise quantitativa dos atributos locais associada à análise qualitativa do processo genético)** – IBGE, Rio de Janeiro, 1983
- [Arm 1998] Armstrong, E. - HotSpot: A new breed of virtual machine, Artigo Online, **JavaWorld**, International Data Group (IDG), Março 1998
URL: <http://www.javaworld.com/jw-03-1998/jw-03-hotspot.html>, acesso em 25.04.2004
- [Ayo 1991] Ayoade, J.O. – **Introdução à Climatologia para os Trópicos** – 3^a Edição, Tradução de Maria Juraci Zani dos Santos, Revisão de Suely Santos, Editora Bertrand Brasil, 1991

- [Bal 2000] Baldonado, M.Q.W.; Woodruff, A.; Kuchinsky, A. - Guidelines for Using Multiple Views in Information Visualization, **Proceedings of the Advanced Visual Interfaces**, pp. 110-119, 2000
- [Bou 2003] Boukhelifa, N.; Roberts, J.C.; Rodgers, P.J. - A Coordination Model for Exploratory Multi-View Visualization, **Proceedings of the International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV'03)**, pp. 76-85, London, England, July 15, 2003
- [Bra 2003] Branco, V.M.A. - **Visualização como Suporte à Exploração de uma Base de Dados Pluviométrica**, Dissertação de Mestrado, ICMC/USP, São Carlos-SP, abril/2003
- [Bro 1995] Brown, J.R.; Earnshaw, R.; Jern, M.; Vince, J. - **Visualization: Using Computer Graphics to Explore Data and Present Information**, John Wiley & Sons, 1995
- [Bro 2003] Brodbeck, D.; Girardin, L. - Design Study: Using Multiple Coordinated Views to Analyze Geo-Referenced High-Dimensional Datasets, **Proceedings of the International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV'03)**, pp. 104-111, London, England, July 15, 2003
- [Cab 1998] Cabena, P.; Hadjinian, P.; Stadler, R.; Verhees, J.; Zanasi, A. - **Discovering Data Mining: from concept to implementation**, Prentice Hall, New Jersey, USA, 1998
- [Car 1999] Card, S.K.; Mackinlay, J.D.; Shneiderman, B. (eds.) - **Readings in Information Visualization – Using Vision to Think**, Morgan Kaufmann, San Francisco, CA, 1999
- [Cha 2000] Chawla, S.; Shekhar, S.; Wu, W.; Ozesmi, U. - Extending Data Mining for Spatial Applications: A Case Study in Predicting Nest Locations, **ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery**, 2000
- [Che 1996] Chen, M.-S.; Han, J.; Yu, P.S. - Data Mining: An Overview from a Database Perspective, **IEEE Transactions on Knowledge and Data Engineering**, v. 8(6), pp. 866-883, December 1996
- [Cle 1993] Cleveland, W.S.; **Visualizing Data**, Hobart Press, Summit, NJ, 1993
- [Con 2002] Conklin, N. - **A web-based, run-time extensible architecture for interactive visualization and exploration of diverse data**, Virginia Polytechnic Institute and State University, Department of Computer Science, Master Thesis, December 2002
- [Der 1997] Derthick, M.; Kolojejchick, J.; Roth, S.F. - An Interactive Visualization Environment for Data Exploration, **Proceedings of Knowledge Discovery in Databases**, pp. 2-9, August 1997
- [Eck 2000] Eckel, B.; Thinking in Java, 2a. edição, Prentice Hall, 2000
- [Eic 2000] Eick, S.G. - Visual Discovery and Analysis, **IEEE Transactions on Visualization and Computer Graphics**, v. 6(1), pp. 44-58, January-March 2000
- [Eic 2000a] Eick, S.G.; Karr, A.F. - Visual Scalability, **Technical Report 106**, National Institute of Statistical Sciences (NISS), NC, USA, June 2000
- [Est 1999] Ester, M.; Kriegel, H-P.; Sander, J. - Knowledge Discovery in Spatial Databases, invited paper at **German Conference On Artificial Intelligence (KI'99)**, Bonn, Germany, 1999
- [Fay 1996] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. - Knowledge Discovery and Data Mining: Towards a Unifying Framework, **Proceeding of The Second International Conference on Knowledge Discovery and Data Mining**, pp. 82-88, 1996
- [Fay 1998] Fayyad, U. - Mining Databases: Towards Algorithms for Knowledge Discovery, **Data Engineering Bulletin**, v. 21(1), pp. 39-48, 1998

- [Fay 2002] Fayyad, U.; Grinstein, G.G.; Wiersem A. - **Information Visualization in Data Mining and Knowledge Discovery**, Morgan Kaufmann, Academic Press, 2002
- [Flo 2000] Flores, E.F. – **Modelagem em Climatologia Geográfica: Um Ensaio Metodológico Aplicado ao Oeste Paulista**, Tese de Doutorado, Instituto de Geociências e Ciências Exatas – Unesp – Campus de Rio Claro, Março de 2000
- [Gah 2000] Gahegan, M.; Takatsuka, M.; Wheeler, M.; Hardisty, F. - GeoVISTA Studio: A Geocomputational Workbench, **The Proceedings of the 5th International Conference on GeoComputation**, August 2000
- [Gan 1996] 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
- [Goe 1999] Goebel, M.; Gruenwald, L. - A Survey Of Data Mining And Knowledge Discovery Software Tools, **SIGKDD Explorations**, ACM SIGKDD, v 1(issue 1), June 1999
- [Gra 2000] Graham, M.; Kennedy, J.; Benyon, D. – Towards a methodology for developing visualizations, **International Journal of Human-Computer Studies**, v. 53(5), pp. 789-807, 2000
- [Gri 2001] Grinstein, G.; Trutschl, M.; Cvek, U. – High-Dimensional Visualizations, **Proceedings of the Visual Data Mining workshop (KDD'2001)**, 2001
- [Gri 2002] Grinstein, G.G.; Ward, M.O. – Introduction to Data Visualization, pp. 21-45, In: **Information Visualization in Data Mining and Knowledge Discovery**, Fayyad, U., Grinstein, G.G., Wierse, A. (eds), Morgan Kaufmann, Academic Press, 2002
- [Han 1998] Hand, D.J. – Data Mining: Statistics and More ?, **The American Statistician**, v. 52(2), May 1998
- [Hin 1999] Hinneburg, A; Keim, D.A; Wawryniuk, M. – HD-Eye: Visual Mining of High-Dimensional Data, **IEEE Computer Graphics and Applications**, v. 19(5), pp. 22-31, September/October 1999
- [Hof 1997] Hoffman, P.; Grinstein, G.; Marx, K.; Grosse, I.; Stanley, E. – DNA Visual And Analytic Data Mining, **8th IEEE Visualization'97 Conference**, Phoenix, AZ, October 1997
- [Hof 1999] Hoffman, P.E. – **Table Visualizations: A Formal Model and its Applications**, PhD Thesis, University of Massachusetts Lowell, USA, November 1999
- [Hor 2001] Horstmann, C.S.; Cornell, G. – **Core Java, Volume II – Recursos Avançados**, Makron Books, São Paulo, 2001
- [Ins 1997] Inselberg, A. – Multidimensional Detective, **Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis'97)**, pp. 100-107, 1997
- [Jer 2003] Jern, M.; Palmberg, S.; Ranlof, M.; Nilsson, A. - Coordinated Views in Dynamic Interactive Documents, **Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV'03)**, pp. 95-101, July 15, 2003
- [Kei 1994] Keim, D.A; Kriegel, H.-P.; Seidl, T. – Supporting Data Mining of Large Databases by Visual Feedback Queries, **Proc. 10th International Conference on Engineering**, pp. 302-313, Houston, Texas, 1994
- [Kei 1994a] Keim, D.A., Kriegel, H.-P. – VisDB: Database Exploration Using Multidimensional Visualization. **IEEE Computer Graphics and Applications**, pp. 40-49, September 1994
- [Kei 1995] Keim, D.A., Kriegel, H.-P. – Issues in Visualizing Large Databases, **Proceedings of the International Conference on Visual Database Systems**, pp. 203-214, 1995

- [Kei 1996] Keim, D.A.; Kriegel, H.-P.; - Visualization Techniques for Mining Large Databases: A Comparison, **IEEE Transactions on Knowledge and Data Engineering**, v. 8(6), pp. 923-938, December 1996
- [Kei 2000] Keim, D.A. – Designing Pixel-Oriented Visualization Techniques: Theory and Applications, **IEEE Transactions on Visualization and Computer Graphics**, v. 6(5), pp. 59-78, January-March 2000
- [Kei 2001] Keim, D. – Visual Exploration of Large Data Sets, **Communications of the ACM**, v. 44(8), pp. 38-44, 2001
- [Kei 2001a] Keim, D.; Hao, M.C.; Ladisch, J.; Hsu, M.; Dayal, U. – Pixel Bar Charts: A New Technique for Visualizing Large Multi-Attribute Data Sets without Aggregation, **IEEE Symposium on Information Visualization 2001 (INFOVIS'01)**, p. 113-120, San Diego, CA, October 22-23, 2001
- [Kei 2002] Keim, D. – Information Visualization and Visual Data Mining, **IEEE Transactions on Visualization and Computer Graphics**, v. 7(1), January-March, 2002
- [Kim 2004] Kimani, S.; Lodi, S.; Catarci, T.; Santucci, G.; Sartori, C. - VidaMine: a Visual Data Mining Environment, **Journal of Visual Languages and Computing**, pp. 37-67, v15(1), February 2004
- [Kop 1996] Koperski, K.; Adhikary, J.; Han, J. - Spatial Data Mining: Progress and Challenges Survey Paper, **Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery**, Montreal, Canada, 1996
- [Kop 1998] Koperski, K.; Han, J.; Adhikary, J. - Mining Knowledge in Geographical Data, **Communications of ACM**, 1998
- [Kos 2003] Kosara, R.; Healey, C.G.; Interrante, V.; Laidlaw, D.H.; Ware, C. – User Studies: Why, How, and When ?, **IEEE Computer Graphics and Applications**, pp. 21-25, July/August, 2003
- [Lem 2001] Lemay, L.; Cadenhead, R. - **Aprenda em 21 dias: Java 2 Professional Reference**, Editora Campus, 2001
- [Lib 2001] Liberty, J.; Kralej, M. – **Aprendendo a desenvolver documentos em XML para a Web**, Makron Books, São Paulo, 2001
- [Mac 1999] MacEachren, A.M.; Wachowicz, M.; Edsall, R.; Haug, D.; Masters, R. – CONSTRUCTING KNOWLEDGE FROM MULTIVARIATE SPATIOTEMPORAL DATA: Integrating Geographic Visualization (GVis) with Knowledge Discovery in Database (KDD) Methods, **International Journal of Geographic Information Science**, v 13(4), pp. 311-334, 1999
- [Mac 2000] Mackinlay, J.D. – Opportunities for Information Visualization, **IEEE Computer Graphics and Applications**, v. 20(1), pp. 22-23, January-February, 2000
- [Mac 2001] MacEachren, A. M.; Hardisty, F.; Gahegan, M.; Wheeler, M.; Dai, X.; Guo, D.; Takatsuka, M. - Supporting Visual Integration and Analysis of Geospatially-referenced Data Through Web-deployable, Cross--platform Tools, **Proceedings of the National Conference for Digital Government Research**, Los Angeles, CA, May 21-23, 2001
- [Mac 2001a] MacEachren, A.M.; Kraak, M-J. – Research Challenges in Geovisualization, **Cartography and Geographic Information Science**, pp. 3-12, v. 28(1), 2001
- [Man 1998] Mangione, C. - Performance Tests show Java as fast as C++, Artigo Online, **JavaWorld**, International Data Group (IDG), Fevereiro 1998
URL: http://www.javaworld.com/javaworld/jw-02-1998/jw-02-jperf_p.html,
acesso em 25.04.2004

- [Man 2004] Manco, G.; Pizzuti, C.; Talia, D. - *Eureka!*: an interactive and visual knowledge discovery tool, **Journal of Visual Languages and Computing**, pp. 1-35, v15(1), February 2004
- [May 2002] May, S.; Savinov, A. - An integrated platform for spatial data mining and interactive visual analysis, **3rd International Conference on Data Mining Methods and Databases for Engineering, Finance, and Other Fields**, Bologna, Italy, September, 2002
- [May 2002a] May, M. - SPIN! Spatial Mining for Data of Public Interest, slides de apresentação, **KDNet (Knowledge Discovery Network) European Project Presentation Day**, Helsinki, August 2002
- [Nor 1997] North, C.; Shneiderman, B. - A Taxonomy of Multiple-Window Coordination, University of Maryland Computer Science Department, **Technical Report #CS-TR-3854**, 1997
- [Nor 1999] North, C.; Shneiderman, B. - Snap-Together Visualization: Coordinating Multiple Views to Explore Information, **Technical Report CS-TR-4020**, 1999
- [Nor 2000] North, C. - **A User Interface for Coordinating Visualizations Based on Relational Schemata: Snap-Together Visualization**, University of Maryland, Computer Science Department, Doctoral Dissertation, May 2000
- [Nor 2000a] North, C.; Shneiderman, B. - Snap-Together Visualization: Can Users Construct and Operate Coordinated Views?, **International Journal of Human-Computer Studies**, Academic Press, 53(5), pp. 715-739, November 2000
- [Nor 2000b] North, C.; Shneiderman, B. - Snap-Together Visualization: A User Interface for Coordinating Visualizations via Relational Schemata, **Proceedings of the ACM Advanced Visual Interfaces 2000**, pp. 128-135, May 2000
- [Nor 2001] North, C. - Multiple Views and Tight Coupling in Visualization: A Language, Taxonomy, and System, **Proc. CSREA CISST 2001 Workshop of Fundamental Issues in Visualization**, pp. 626-632, June 2001
- [Nor 2002] North, C.; Conklin, N.; Idukuri, K.; Saini, V. - Visualization Schemas and a Web-based Architecture for Custom Multiple-View Visualization of Multiple-Table Databases, **Information Visualization**, Palgrave-Macmillan, December 2002
- [Nor 2002a] North, C.; Conklin, N.; Saini, V. - Visualization Schemas for Flexible Information Visualization, **Proceedings of the IEEE InfoVis 2002 Symposium**, October 2002
- [Nor 2003] North, C.; Conklin, N.; Indukuri, K.; Saini, V.; Yu, Q. - Fusion: Interactive coordination of diverse data, visualizations, and mining algorithms, **Proceedings of the ACM CHI 2003**, March 2003
- [Oli 2003] Oliveira, M.C.F.; Levkowitz, H. - From Visualization to Visual Data Mining: A Survey, **IEEE Transactions on Visualization & Computer Graphics**, pp. 378-394, v. 9(3), July-September, 2003
- [Oli 2004] Oliveira, M.C.F.; Shimabukuro, M.H. - A Conceptual Model for Adaptable and Extensible Visual Data Exploration, 12 p., **Proceedings of the Visualization and Data Analysis 2004**, v. 5295 (IS&T/SPIE Symposium on Electronic Imaging 2004), (January 8-22 2004), San Jose, CA, EUA, 2004
- [Pat 2001] Pattison, T.; Phillips, M. - View Coordination Architecture for Information Visualisation, Proceedings of the Australian Symposium on Information Visualisation (invis.au 2001), Sydney, Australia, In: **Conferences in Research and Practice in Information Technology**, v. 9, pp. 165-171, Eades, P.; Pattison, T (eds), 2001
- [Ram 2000] Ramirez, J.C.G.; Cook, D.J.; Peterson, L.L.; Peterson, D.M. - Temporal Pattern Discovery in Course-of-Disease Data, versão To appear in **IEEE Engineering in Medicine and Biology**, 2000

- [Rao 1994] Rao, R.; Card, S.K. – The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+Context Visualization for Tabular Information, pp. 318-322, **Proceedings CHI'94**, ACM, 1994
- [Rez 2003] Rezende, S.O.; Pugliesi, J.B.; Melanda, E.A.; Paula, M.F. – Mineração de Dados, Capítulo 12, pp. 307-335, In: Rezende, S.O. (organizadora), **Sistemas Inteligentes Fundamentos e Aplicações**, Editora Manole, Barueri, SP, 2003
- [Rhy 2003] Rhyne, T-M. – Does the Difference between Information and Scientific Visualization Really Matter?, **IEEE Computer Graphics and Applications**, pp. 6-8, May-June 2003
- [Rib 1999] Ribarsky, W.; Katz, J.; Jiang, F.; Holland, A – Discovery Visualization Using Fast Clustering, **IEEE Computer Graphics and Applications**, v. 19(5), pp. 32-39, September/October 1999
- [Rog 2001] Rogowitz, B.E.; Kalvin, A.D. - The Which Blair Project: A Quick Visual Method for Evaluating Perceptual Color Maps, pp. 183-190, **Proceedings of the Conference on Visualization '01**, IEEE Computer Society, San Diego, CA, 2001
- [Roh 1999] Rohrer, R.M.; Sibert, J.L.; Ebert, D.S. – A Shape-based Visual Interface for Text Retrieval, **IEEE Computer Graphics and Applications**, v. 19(5), pp. 40-46, September/October 1999
- [Sch 1998] Schroeder, W., Martin, L., Lorensen, B. - **The Visualization Toolkit An Object-Oriented Approach to 3D Graphics**, 2nd Edition, Prentice Hall, 1998
- [Scr 2000] Scribner, K.; Stiver, M.C. – Understanding SOAP – The Authoritative Solution, Sams Publishing, 2000
- [Shi 2003] Shimabukuro, M.H.; Branco, V.M.A.; Oliveira, M.C.F.; Flores, E.F. - Visual Exploration of Spatio-Temporal Databases, **GEOINFO 2003 V Simpósio Brasileiro de Geoinformática**, cd-rom, Campos do Jordão, SP, 2003
- [Shi 2004] Shimabukuro, M.H.; Flores, E.F.; Oliveira, M.C.F.; Levkowitz, H. – Coordinated Views to Assist Exploration of Spatio-Temporal Data: a Case Study, **CMV 2004**, 10 p. (aceito)
- [Shn 1994] Shneiderman, B. – Dynamic Queries for Visual Information Seeking, **IEEE Software**, pp. 70-77, v. 11(6), 1994
- [Shn 1996] Shneiderman, B. – The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations, **Proceedings of the 1996 IEEE Symposium on Visual Languages**, IEEE Computer Society, pp. 336-343, Boulder, CO, September 3-6, 1996
- [Sow 1998] Sowizral, H.; Rushforth, K.; Deering, M. – **The Java 3D API Specification**, Addison-Wesley, USA, 1998
- [Spe 2001] Spence, R. – **Information Visualization**, Addison-Wesley, 2001
- [Str 2000] Stroe, I.D.; Rundensteiner, E.A.; Ward, M.O. – Scalable Visual Hierarchy Exploration, **Computer Science Technical Report Series (WPI-CS-TR-00-06)**, Computer Science Department, Worcester Polytechnic Institute, March 2000
- [Sun 2001] The Java HotSpot Virtual Machine, **Technical White Paper**, Sun Microsystems, Palo Alto, CA, USA, May 2001
- [Swa 2001] Swayne, D.F.; Lang, D.T.; Buja, A.; Cook, D. - GGobi: XGobi redesigned and extended, **Proceedings of the 33th Symposium on the Interface: Computing Science and Statistics**, 2001
- [Swa 2002] Swayne, D.F.; Cook, D.; Buja, A.; Lang, D.T. – **ggobi Manual**, 40 p., February, 2002

- [Sym 2000] Symanzik, J.; Cook, D.; Lewin, N.; Majure, J.J.; Megretskaia, I. – Linking ArcView 3.0 and Xgobi: Insight Behind the Front End, **Journal of Computational and Graphical Statistics**, v. 9(3), pp. 470-490, 2000
- [Sym 2002] Symanzik, J.; Swayne, D.F.; Lang, D.T.; Cook, D. - Software Integration for Multivariate Exploratory Spatial Data Analysis, **Proceedings of the SCISS Specialist Meeting "New Tools for Spatial Data Analysis"**. Santa Barbara, California, USA. May 10-11, 2002
- [Tak 2002] Takatsuka, M.; Gahegan, M. - GeoVISTA Studio: A Codeless Visual Programming Environment For Geoscientific Data Analysis and Visualization, **The Journal of Computers & Geosciences**, 2002
- [Tak 2002a] Takatsuka, M. - An Open Component-Oriented Visual Programming Environment for Integrating Geospatial Data Analysis and Visualization Tools, **Specialist Meeting on Spatial Data Analysis Software Tools**, Upham Hotel, Santa Barbara, CA, May 10-11, 2002
- [Tej 2003] Tejada, E.; Minghim, R.; Nonato, L.G. - On Improved Projection Techniques to Support Visual Exploration of Multi-dimensional Data Sets, **Information Visualization Journal**, v. 2(4), Palgrave Mcmillan, December 2003
- [Tra 2002] Traina Jr., C.; Traina, A.J.M.; Faloutsos, C.; Seeger, B. - Fast Indexing and Visualization of Metric Datasets Using Slim-trees, **IEEE Transactions on Knowledge and Data Engineering**, v. 14, pp. 244-260, 2002
- [Tuf 1983] Tufte, E.R. – **The Visual Display of Quantitative Information**, Graphics Press, USA, 1983
- [Tuf 1990] Tufte, E.R. – **Envisioning Information**, Graphics Press, USA, 1990
- [Tut 1998] Tutida, S.M. – **Mapeamento de Cores Baseado em Aspectos de Percepção Visual**, Dissertação de Mestrado, ICMC-USP, São Carlos, SP, abril/1998
- [War 1994] Ward, M.O. - XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data. **Proceedings of Visualization '94**, pp. 326-33, 1994
- [War 2002] Ward, M.O.; Rundensteiner, E.A.; Yang, J.; Doshi, P.R.; Rosario, G. - Interactive Poster: XmdvTool: Interactive Visual Data Exploration System For High-dimensional Data Sets, **IEEE Symposium on Information Visualization 2002 (InfoVis 2002)**, pp 52-53, 2002
- [Weg 1996] Wegman, E.J.; Luo, Q. - High Dimensional Clustering using Parallel Coordinates and the Grand Tour, **Technical Report No. 124**, Center for Computational Statistics, George Mason University, 1996
- [Wij 1999] Wijk, J.J. van; Selow, E.R. van – Cluster and Calendar based Visualization of Time Series Data, **IEEE Symposium on Information Visualization (INFOVIS'99)**, San Francisco, CA, October 25-26, 1999
- [Won 1997] Wong, P.C.; Bergeron, R.D. - 30 Years of Multidimensional Multivariate Visualization, **Scientific Visualization – Overviews, Methodologies and Techniques**, Gregory M. Nielson, Hans Hagan, and Heinrich Muller (editors), pp. 3-33, IEEE Computer Society Press, Los Alamitos, CA, USA, 1997
- [Won 1999] Wong, P.C. - Visual Data Mining, **IEEE Computer Graphics and Applications**, v. 19(5), pp. 20-21, September/October 1999

CLASSIFICAÇÃO CLIMÁTICA E DADOS PLUVIOMÉTRICOS

Um dos conjuntos de dados utilizado neste projeto foi aquele que o Prof. Dr. Edilson Ferreira Flores, da Faculdade de Ciências e Tecnologia – Unesp - Presidente Prudente, usou para uma classificação climática, atividade essa relacionada à Climatologia, na sua tese de doutoramento [Flo 2000]. Neste Apêndice são apresentados na Seção A.1 alguns conceitos básicos de Climatologia e classificação climática, e na Seção A.2 duas etapas realizadas pelo Prof. Flores no processo de classificação. O conjunto de dados sobre precipitação pluviométrica usado nessa classificação é descrito na Seção A.3. Esse estudo foi necessário para ganhar um entendimento do respectivo domínio de aplicação, bem como para facilitar a comunicação com o usuário.

A.1 CLIMATOLOGIA E CLASSIFICAÇÃO CLIMÁTICA

De acordo com Ayoade [Ayo 1991] existe uma distinção entre clima e tempo, assim como entre Climatologia e Meteorologia. Tempo é entendido como um estado da atmosfera em um intervalo de tempo sobre um local definido, enquanto clima é a síntese do tempo em um dado local durante um período de aproximadamente 30-35 anos, referindo-se a características ou padrões de comportamento atmosférico inferidos a partir de observações contínuas coletadas por um longo período. Dessa forma, clima representa uma generalização, e tempo está relacionado a eventos específicos.

A Climatologia se baseia na Meteorologia, e os climatologistas empregam técnicas estatísticas na busca de informações sobre o clima a partir de dados meteorológicos. O termo ‘climatologia regional’ é usado na referência a clima sobre áreas na superfície terrestre [Ayo 1991].

Embora dois locais na superfície da Terra não tenham climas idênticos, é possível definir regiões climáticas cujos climas são relativamente uniformes. O objetivo da classificação climática é prover uma forma eficiente de organizar dados climáticos e entender as complexas variações do clima [Ayo 1991]. O fenômeno climático tem característica multivariada e, por isso, os climatologistas usam várias técnicas de estatística multivariada, tais como análise de *clustering* (agrupamento) e análise de componentes principais, para realizar uma classificação climática [Aou 1983].

Modelos climáticos gerados pelo tratamento estatístico considerando apenas atributos locais, sem relação com mecanismos da circulação atmosférica, são modelos espaciais que revelam somente a estrutura. O entendimento de tais mecanismos (processo de causa e efeito) é importante, e requer uma análise ao longo do tempo [Aou 1983]. Em um projeto apresentado por Aouad [Aou 1983] uma descrição climática foi gerada usando como parâmetros a temperatura, a pluviosidade, a umidade relativa e evapotranspiração potencial; na análise temporal foram considerados os valores anuais e extremos sazonais de parâmetros. Na Seção A.2, são descritas duas etapas da classificação climática, que foram estudadas com o objetivo de ter um panorama geral do processo conduzido pelo Prof. Flores. A primeira etapa descrita foi selecionada para estudo por ser uma etapa inicial do processo de classificação, e essa foi a etapa explorada com recursos de visualização neste projeto. Para se ter uma idéia dos parâmetros e outros dados envolvidos no processo de classificação, além dos dados de precipitação, bem como do uso desses parâmetros numa decisão é que foi estudada a segunda etapa.

A.2 CLASSIFICAÇÃO CLIMÁTICA REALIZADA NA TESE DO PROF. FLORES

A classificação climática gerada na tese do Prof. Flores abrangeu a região oeste paulista, e seu resultado é representado por um cartograma (Figura A.1), uma representação gráfica com diagramas aplicados sobre mapas acompanhada por uma explicação textual (não apresentada aqui). O conteúdo dessa Seção foi elaborado com base em uma entrevista com o Prof. Flores, na

qual foram explicados algumas das etapas realizadas na elaboração da classificação. Duas dessas etapas são descritas: (1) definição de anos representativos (seco, chuvoso, habitual) e (2) análise rítmica. A primeira etapa foi conduzida utilizando dados de chuva do banco de dados BcDAEE, descrito na Seção A.3. Os dados de chuva foram escolhidos pela sua disponibilidade e porque existem diversos trabalhos de pesquisa com esses dados; uma outra possibilidade seria usar a temperatura do ar, cuja variação está estreitamente relacionada com modificações na circulação atmosférica.

Figura A.1. Cartograma com uma classificação climática [Flo 2000].

A.2.1 DEFINIÇÃO DE ANOS REPRESENTATIVOS (SECO, CHUVOSO E HABITUAL)

Essa etapa pode ser dividida em três tarefas: (a) seleção de postos de coleta de chuva, (b) tratamento de dados ausentes, e (c) agrupamento de anos em classes representativas usando uma técnica de *clustering*. O objetivo da tarefa *a* é obter um conjunto de postos cujos dados sejam razoavelmente completos e confiáveis. Para cada posto selecionado, os dados ausentes são preenchidos na tarefa *b*. Na tarefa *c*, classes de anos representativos (seco, chuvoso e habitual) são definidas, e um ano representativo de cada classe é escolhido. Uma breve descrição dessas tarefas é apresentada a seguir:

- **Seleção de postos de coleta.** O Prof. Flores construiu uma tabela a partir de dados do BcDAEE, com postos colocados nas linhas e anos nas colunas. Cada célula correspondente ao par (posto,ano) foi preenchida com caracteres com os seguintes

significados: X = dado consistido, * = dado não consistido e “branco” = dado ausente. Após o preenchimento da tabela foram buscados postos com X ou * que tinham um intervalo de anos em comum. Inicialmente, foram selecionados 200 postos, e posteriormente, com uma análise mais detalhada, foram selecionados 150 postos;

- **Preenchimento de dados ausentes.** Essa tarefa pode seguir duas abordagens: (1) o dado ausente é preenchido com um valor de precipitação do mesmo dia/mês de um outro ano com comportamento similar de precipitação, dentro do mesmo posto; (2) o dado ausente é preenchido com um valor interpolado usando valores de postos vizinhos. Nessa última abordagem a informação espacial, relacionada à localização e à altitude de cada posto vizinho, pode ser útil para uma decisão mais precisa sobre os valores a usar;
- **Agrupamento de anos representativos (seco, chuvoso e habitual).** Nessa tarefa foi utilizado um aplicativo que implementa uma técnica de *clustering* para agrupar anos com base no valor de total anual de chuva. A Figura A.2 mostra um resultado de agrupamento representado por uma árvore de ligação, a partir da qual três anos, um para cada agrupamento, são escolhidos. Os anos estão ordenados na horizontal de acordo com a similaridade no total anual de chuva. Várias formas de agrupamento são possíveis conforme o grau de perda de detalhe admissível, indicado no eixo vertical. As linhas que conectam os anos dão uma indicação dos agrupamentos.

Figura A.2. Árvore de ligação com agrupamentos para anos representativos [Flo 2000].

A.2.2 ANÁLISE RÍTMICA

Pesquisadores em clima usam a análise rítmica para determinar a circulação atmosférica em base diária. São considerados simultaneamente vários parâmetros, como precipitação, temperatura e pressão do ar, e direção e velocidade do vento. Esses parâmetros são representados em um gráfico, ilustrado na Figura A.3; variações nesses parâmetros indicam modificações no clima. Essa análise foi realizada para cada um dos anos representativos escolhidos. A informação da última linha, circulação atmosférica, é deduzida pelo climatologista. Uma carta sinótica (Figura A.4), que mostra zonas de pressão, também é consultada pelo climatologista para concluir sobre a circulação atmosférica.

Figura A.3. Gráfico de Análise Rítmica [Flo 2000].

Figura A.4. Carta sinótica [Flo 2000].

A.3 BcDAEE – BANCO DE DADOS PLUVIOMÉTRICOS

O BcDAEE, Banco de Dados Pluviométricos do Estado de São Paulo, é implementado como um banco de dados do Microsoft Access. No banco estão armazenados os dados de precipitação diária (agrupados por mês) pluviométrica, em mm, de 1660 postos do DAEE (Departamento de Águas e Energia Elétrica) no estado de São Paulo. Cada posto é identificado por um prefixo, D8-003, por exemplo, que é o nome da tabela correspondente no banco de dados. Existe uma tabela adicional que associa a cada posto informações de latitude, longitude, altitude, nome do posto, etc. Os dados de precipitação, para cada mês, estão codificados em campo de 62 bytes, dificultando o acesso aos dados sem o uso do aplicativo que acompanha o banco. A versão utilizada aqui é a 1.0, atualizada até 1997. A base é disponibilizada sem custos.

A tela inicial que surge para o usuário é mostrada na Figura A.5. Cada ponto representa uma estação de coleta. Com o mouse, realiza-se a seleção do posto sobre o qual deseja-se informações.

Figura A.5. Tela inicial do aplicativo BcDAEE.

A Figura A.6 ilustra um *zoom* de uma região próxima a Presidente Prudente, identificada pela circunferência magenta. Para uma mesma cidade, podem estar relacionados diversos postos.

Figura A.6. Zoom em torno de uma região no aplicativo BcDAEE.

É possível ver os dados disponíveis para cada posto. Na Figura A.7 está um exemplo do posto D8-003, de Presidente Prudente.

Banco de dados DAEE - dados do posto

Prefixo: **D8-003**
 Nome: **PRESIDENTE PRUDENTE**
 Município: **PRESIDENTE PRUDENTE**
 Bacia: **MANDAGUARI**
 Altitude (m): **460,00**
 Latitude: **22° 05'**
 Longitude: **51° 22'**

Chuva diária
 Chuva mensal
 Máximos mensais
 Pluviograma médio diário
 Pluviograma acumulado mensal
 Pluviograma acumulado médio mensal
 Sair

Chuva diária (mm)
 Ano: **1942** 1936 1997 **Exporta dados (*.txt)**

	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
1	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	3.0*
2	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	0.0*
3	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	0.0*
4	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	0.0*
5	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	0.0*
6	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	0.0*
7	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	0.0*
8	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	31.0*
9	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	0.0*
10	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	21.0*
11	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	0.0*
12	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	0.0*
13	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	2.5*	0.0*
14	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	2.0*
15	???	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	???	0.0*	0.0*
16	???	???	???	???	???	???	???	???	???	???	0.0*	0.0*
17	???	???	???	???	???	???	???	???	???	???	0.0*	0.0*
18	???	???	???	???	???	???	???	???	???	???	0.0*	0.0*
19	???	???	???	???	???	???	???	???	???	???	0.0*	0.0*
20	???	???	???	???	???	???	???	???	???	???	11.0*	30.0*

Obs: valores com * representam dados não consistidos; células com ??? representam dados inexistentes

Figura A.7. Dados do posto D8-003 no aplicativo BcDAEE.

No exemplo da figura anterior estão sendo vistos os dados do ano de 1942. No caso deste posto estão disponíveis dados de 1936 a 1997, entretanto, para outros postos o intervalo pode ser

diferente. Cada dado pode estar em um dos estados: consistido (valor válido e confiável), não consistido (identificado por um '*') ou inexistente (identificado por '???').

As opções de visualização de pluviograma geram um histograma, exemplificado na Figura A.8.

Figura A.8. Exemplo de pluviograma no aplicativo BcDAEE.

Os dados podem ser exportados para um arquivo texto que apresenta o formato mostrado a seguir.

```
PREFIXO:      D8-003
NOME DO POSTO: PRESIDENTE PRUDENTE
MUNICÍPIO:    PRESIDENTE PRUDENTE
BACIA:        MANDAGUARI
ALTITUDE (m): 460,00
LATITUDE:     22° 05'
LONGITUDE:    51° 22'
```

Obs: valores com * representam dados não consistidos;
valores com ??? representam dados inexistentes.

Dados de chuva (mm)

Mês/Ano	1	2	3	...	Chuva_total	Chuva_máxima	Dias_com_chuva
12/1936	30,0*	18,0*	0,0*	...	286,0*	62,0*	14*
1/1937	2,0*	11,5*	32,5*	...	268,5*	80,0*	15*
2/1937	0,0*	0,0*	0,0*	...	169,0*	56,5*	14*
...	...	...	...	...	...	...	...

ESTUDOS INICIAIS COM DADOS PLUVIOMÉTRICOS

Algumas estratégias de visualização foram implementadas e utilizadas em colaboração com um potencial usuário na época, o Prof. Flores, de modo a obter uma impressão inicial para o andamento do projeto. Dados de chuva de alguns postos do banco de dados do BcDAEE, descrito na Apêndice A, foram exportados para o formato texto e usados para gerar as visualizações. O propósito desses estudos iniciais foi familiarizar o usuário com as formas de visualização e realizar uma apreciação da aplicabilidade dessas formas sob o seu ponto de vista. Os resultados foram satisfatórios, com o próprio usuário gerando conclusões após uma explicação das formas de visualização, indicando que uma abordagem visual teria potencial para auxiliar o usuário nas suas atividades que, para este caso específico estão brevemente descritas no Apêndice A. Para esses estudos iniciais foi considerado a etapa de seleção de anos representativos, um das etapas iniciais da classificação climática realizada pelo Prof. Flores, no qual ele escolhia, para cada posto de coleta, um ano representativo para as condições de chuva (seco, chuvoso e habitual). Essa tarefa foi escolhida para os estudos por ser uma etapa inicial do processo de classificação climática (Apêndice A) e envolver um grande volume de dados em diversas escalas (diário, mensal, anual).

Os botões na tela apresentada na Figura B.1 dão acesso às formas de visualização implementadas e funções associadas. O primeiro passo (botão “Carregar Postos”) a ser executado é o carregamento de alguns atributos (prefixo, nome, altitude, latitude, longitude,

primeira e última data de coleta de dados) que caracterizam cada posto de coleta. Para esses estudos iniciais foram selecionados 35 postos, quantidade que se mostrou suficiente para os propósitos do estudo. Os dois primeiros atributos, prefixo e nome, são apresentados no *listbox* logo abaixo do botão da função “Carregar Postos”. No *grid*, mais abaixo, aparece a lista completa dos demais atributos. Um posto selecionado no *listbox* também é destacado no *grid*. Na Figura B.1, os atributos dos postos já foram carregados. O mapa no canto inferior, no momento, é ilustrativo, mas pode vir a ser usado para mostrar a posição dos postos.

Figura B.1. Elementos da interface para os estudos iniciais.

Com a opção “Anos início-fim”, o usuário visualiza, por meio de uma variação da técnica Coordenadas Paralelas (ver Subseção 2.2.2), o intervalo entre a data da primeira coleta (coluna Início no *grid*) e a data da última coleta (coluna Fim no *grid*). O resultado dessa visualização está na Figura B.2. O atributo primeira data está associado ao eixo a esquerda e o atributo última data está associado ao eixo a direita. Ambos os eixos têm a mesma escala, com a base referindo-se a 1930 e o topo ao ano 2000. As visualizações são feitas no quadro de fundo cinza na parte superior direita da tela da Figura B.1, em que todos os postos no *listbox* são visualizados. O usuário detectou dois agrupamentos, delimitados por retângulos na Figura B.2, sugerindo que eles deveriam ser processados separadamente, em virtude de apresentarem intervalos de anos em comum.

Figura B.2. Visualização da opção “Anos início-fim”.

A visualização gerada ativando o botão “Total Mensal” utiliza o parâmetro configurado com a barra deslizante logo abaixo do botão (ver Figura B.1), que permite indicar o ano a partir do qual serão selecionados os dados para a visualização. No caso da Figura B.3, o ano inicial escolhido é 1930, delimitado por uma elipse na figura. Para cada um dos postos, informação sobre os totais mensais de chuva registrados em um intervalo de 30 anos (nesse caso, 1930-1959) é exibida. Cada linha representa um posto e cada pixel nessa linha informa a condição de um dado de precipitação (consistido, não consistido, inexistente) do total de chuva medido em um determinado mês/ano, com os meses ordenados da esquerda para a direita. As cores utilizadas foram sugeridas pelo usuário, sendo que azul indica dado consistido, amarelo indica dado não consistido, vermelho indica dado inexistente (não coletado). Detalhes sobre os dados são fornecidos no Apêndice A. Na visualização da Figura B.3, percebe-se uma predominância da cor amarela, ou seja, de dados não consistidos, o que pode indicar que o intervalo de anos selecionado para os postos em questão não é adequado. Pode-se observar, também, que muitos postos não têm dados referentes aos anos próximos a 1930 (indicado pela linha vertical na Figura B.3), pois os pixels mais a esquerda não foram pintados (seta na Figura B.3), permanecendo com a cor do fundo, isto é, não há registro para aquele mês/ano no banco de dados. Vale salientar que o dado inexistente, indicado pela cor vermelha, refere-se a um atributo registrado no banco de dados, mas cujo valor não foi coletado.

Figura B.3. Totais mensais a partir de 1930.

Na Figura B.4, o ano inicial escolhido foi 1968. O número de pixels com a cor azul aumentou, indicando que o intervalo pode ser mais conveniente do que o anterior. Nessa aplicação para estudos, os valores possíveis para o ano inicial estão no intervalo 1930-1968, previamente estabelecido por uma verificação da base de dados. Ainda, é possível ter uma idéia dos postos cujos dados podem ser potencialmente mais úteis, ou seja, aqueles que têm o intervalo de anos com dados representados predominantemente pela cor azul.

Figura B.4. Totais mensais a partir de 1968.

As visualizações apresentadas até aqui mostram os dados de todos os postos carregados. Para visualizar os dados referentes a um posto em particular, é preciso selecioná-lo no *listbox* e, em seguida, clicar no botão “Chuva Diária”. Como o rótulo da função sugere, a escala é diária. Nessa forma de visualização, os anos mais antigos (1930, por exemplo) estão na parte superior do quadro de visualização, e o posicionamento da faixa colorida dá noção do intervalo de anos para os quais existem dados naquele posto. Quanto mais próximos da parte inferior ou base do quadro de visualização, mais recentes são os dados. A Figura B.5 é uma visualização desse tipo para o posto de prefixo D8-003. Uma linha representa um ano, e cada ponto (pixel) da linha informa o *status* (consistido, não consistido, inexistente) do dado referente a um dia/mês. Percebe-se alguns pontos escuros, formando uma linha vertical, próximos ao início da imagem (seta na Figura B.5). Esses pontos se referem ao dia 29 de fevereiro, que não aparece em todos os anos.

Figura B.5. Visualização de dados de um único posto em escala diária.

A opção “Múltiplas Visualizações” estende o tipo de visualização descrito acima para a exibição de até 6 postos simultaneamente, possibilitando uma análise comparativa entre postos. Os postos selecionados são aqueles listados no *listbox* abaixo do *grid*. Um posto é selecionado pela função “Selecionar Posto”, que copia os dados do posto destacado no *listbox* superior para o *listbox* inferior. Para formar uma nova lista, primeiro a lista atual deve ser removida (“Limpar Seleção”). A Figura B.6 mostra o resultado dessa visualização. A ordem dos quadros de visualização é da esquerda para direita, e de cima para baixo, como indicado na Figura B.6. Nesse exemplo, a questão do posicionamento das faixas coloridas é mais evidente. Pelo posicionamento vertical das faixas pode-se perceber se os dados são recentes ou antigos, e pela extensão vertical da faixa é possível perceber se o intervalo de anos com dados é grande ou pequeno.

Figura B.6. Múltiplas visualizações em escala diária.

Finalizando, é importante ressaltar que muitas das conclusões apresentadas neste Apêndice foram do próprio usuário especialista do domínio. Apesar do conjunto de dados utilizado para os estudos ser pequeno e o sistema implementado ter recursos mínimos de interação, a perspectiva da aplicação de representações visuais, integrando-as nas atividades desenvolvidas pelo usuário, se mostrou promissora.