Visualização de dados multidimensionais referenciados utilizando projeções multidimensionais e animação

Tácito Trindade de Araújo Tiburtino Neves

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito: 21 de outubro de 2011

Assinatura:_____

Visualização de dados multidimensionais referenciados utilizando projeções multidimensionais e animação

Tácito Trindade de Araújo Tiburtino Neves

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

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre na área de Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

USP – São Carlos
Outubro de 2011

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

N511v

Neves, Tácito Trindade de Araújo Tiburtino
Visualização de dados multidimensionais referenciados utilizando projeções multidimensionais e animação / Tácito Trindade de Araújo Tiburtino Neves; orientadora Maria Cristina Ferreira de Oliveira -- São Carlos, 2011.
88 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2011.

1. Visualização de informação. 2. Dados variantes no tempo. 3. Dados geo-referenciados. I. Oliveira, Maria Cristina Ferreira de , orient. II. Título.

Dedicatória

Aos meus pais, pelo amor e apoio não somente durante o período de execução deste trabalho, mas durante toda a minha vida.

Agradecimentos

A Deus, por tudo que vivenciei durante esta etapa de minha vida. Nos momentos bons e difíceis, e foram muitos, Ele esteve sempre comigo.

Aos meus pais Hortogami e Girliane, aos meus irmãos Theomar e Nívea, além de todos os meus familiares. Sem o apoio, o incentivo, as cobranças e o amor nada disso seria possível.

À minha orientadora, Drª Maria Cristina Ferreira de Oliveira, pela paciência e sobretudo pelos ensinamentos transmitidos. Sua constante disposição na orientação deste trabalho, mostrando sempre o caminho que deveria ser trilhado, com abnegação e gosto pelo que faz, tornou-se um grande exemplo a ser seguido.

Aos demais professores que contribuíram com minha formação, no CEFET-AL, na UFAL e aqui no ICMC-USP. De um modo muito especial ao Prof. Dr. Alejandro Frery, que abriu-me as portas da pesquisa, através da iniciação científica. Também aos professores: Dr. Fernando Paulovich e Drª Rosane Minghim pelas contribuições neste trabalho.

Aos amigos que dividiram não somente o mesmo teto, mas também todas as minhas alegrias e tristezas, Fernando Sá (Feu) e Edgar Peixoto. Obrigado por suportarem os momentos em que fui muito chato.

Aos amigos de longa data: Andrezza, Cícero, Daniel, Everson (Cabeção), Higor, Jhonatan Guedes (Gaio), Jhonathan Gomes, Paulo Henrique, Rafael, Thiago, Ericka, Robson e todos os outros que estiveram presentes em minha vida mesmo que à distância.

Aos amigos que fiz em São Carlos: Gabriel (Ceará), Cássio Prazeres, Maycon, Danilo Coimbra, José Augusto (Tim), Danilo Eler, Ricardo Cerri, Rafael Messias, Rodolfo, Renato Rodrigues, Renato (Esquimó), Adalberto (Daga), Willian (Japa), Aretha, Marcel Yugo, Francisco, David, Guilherme (Baiano), Felipe (Mineiro), Nelson Ponce, Karina, Fernando Paulovich, todos os colegas de curso e do laboratório do grupo de pesquisa em Visualização, Imagens e Computação Gráfica (VICG).

A todos que, direta ou indiretamente, contribuíram para que tudo isto fosse possível.

À banca que gentilmente aceitou o convite para essa defesa.

Ao CNPq e FAPESP pelo apoio financeiro.

Resumo

TIBURTINO, Tácito. **Visualização de dados multidimensionais referenciados utilizando projeções multidimensionais e animação**. 2011. Dissertação (Mestrado) – Instituto de Ciências Matemática e de Computação, Universidade de São Paulo, São Carlos, 2011.

Ferramentas e técnicas de visualização promovem uma análise de dados mais efetiva pelo fato de explorar a capacidade humana na percepção de padrões, principalmente em representações gráficas. Muitos fenômenos são associados a algum tipo de referência, temporal ou geográfica, que pode oferecer informação importante quando são submetidos a processos de análise. Este trabalho aborda representações visuais de dados geradas por técnicas de projeção multidimensional, e propõe uma estratégia para o tratamento diferenciado das referências temporais ou geográficas presentes em conjuntos de dados, no processo de gerar uma projeção multidimensional. Foi proposta e implementada uma variação da técnica *Least Square Projection* (LSP) que evidencia a informação das referências e permite ao usuário interagir com os mapas visuais gerados, bem como diversas funcionalidades que auxiliam no processo de análise exploratória. A nova abordagem é ilustrada por meio de estudos de caso envolvendo bases de dados temporais e com referências geográficas, em que foi possível observar o comportamento global dos elementos, bem como comportamentos de elementos ou grupos de elementos de interesse. Limitações da estratégia proposta também são discutidas.

Palavras-chave: Visualização de informação, projeções multidimensionais, dados variantes no tempo, dados geo-referenciados.

Abstract

TIBURTINO, Tácito. **Referenced multidimensional data visualization using multidimensional projections and animation**. 2011. Dissertation (Master) – Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, 2011.

Visualization tools and techniques promote more effective data analysis by exploiting the human visual perception capabilities in detecting patterns in graphical representations. Many phenomena generate data that include temporal or geographical references, which are likely to provide important information in data analysis procedures. This work addresses data visualizations generated with multidimensional projections, proposing a strategy to handle temporal and geographical references present in multidimensional data sets, when generating multidimensional projections. The Least Squares Projection (LSP) technique was extended to explicitly handle the reference information and represent it in the visual maps, and a set of supporting analysis functions have been implemented. The proposed approach is illustrated through case studies on multidimensional data sets, in which it was possible to observe the global behavior of the elements, as well as individual behavior of elements or groups of elements of interest.

Keywords: Information Visualization, multidimensional projections, time-varying data, geo-referenced data.

Sumário

1 Introdução **1**
1.1 Contexto e motivação . . . 1
1.2 Objetivos . . . 5
1.3 Organização da dissertação . . . 5

2 Visualização de dados multidimensionais **7**
2.1 Considerações iniciais . . . 7
2.2 Técnicas geométricas . . . 9
2.2.1 Coordenadas Paralelas . . . 9
2.2.2 Matriz de gráficos de dispersão . . . 10
2.2.3 *TableLens* . . . 11
2.3 Técnicas iconográficas . . . 12
2.3.1 *Starplot* . . . 12
2.3.2 *Chernoff faces* . . . 14
2.4 Técnicas orientadas a pixels . . . 14
2.5 Técnicas de projeção multidimensional . . . 17
2.5.1 *Interactive Document Map* (IDMAP) . . . 18
2.5.2 *Least Square Projection* (LSP) . . . 18
2.5.3 *Projection Explorer* (PEx) . . . 21
2.6 Considerações finais . . . 23

3 Visualização de dados referenciados **25**
3.1 Considerações iniciais . . . 25
3.2 Visualização de dados geo-referenciados . . . 26
3.2.1 Coordenadas Paralelas . . . 26
3.2.2 Representação cartográfica . . . 27
3.2.3 Aplicações de visualização de dados geo-referenciados . . . 28
3.2.3.1 Power Map Explorer . . . 28
3.2.3.2 LAHVA . . . 30
3.2.3.3 Taggram . . . 32
3.3 Visualização de dados variantes no tempo . . . 32
3.3.1 Coordenadas Paralelas . . . 33
3.3.2 Técnica *Cluster and Calendar* . . . 34

3.3.3 Técnicas orientadas a pixels . . . 35
3.3.4 Projeções multidimensionais . . . 35
3.3.5 Animação apoiando a visualização . . . 37
3.4 Considerações finais . . . 39

4 *Least Square Projection* com pontos de controle fixos **41**
4.1 Considerações iniciais . . . 41
4.2 *Least Square Projection* com pontos de controle fixos . . . 42
4.2.1 Animação dos pontos . . . 46
4.2.2 Visualização dos rastros . . . 50
4.2.3 LSP_{FCP} na *Projection Explorer* . . . 53
4.3 Visualizando séries temporais geo-referenciadas . . . 56
4.4 Considerações finais . . . 62

5 Resultados **63**
5.1 Considerações iniciais . . . 63
5.2 Índice de Desenvolvimento Humano . . . 65
5.3 Dados do *Gapminder* . . . 70
5.3.1 Dados do grupo economia . . . 71
5.3.2 Componentes do IDH . . . 75
5.4 Considerações finais . . . 75

6 Conclusões **81**
6.1 Contribuições . . . 81
6.2 Limitações . . . 82
6.3 Trabalhos futuros . . . 83

Referências Bibliográficas **85**

Lista de Figuras

1.1 Ataques no Rio de Janeiro e reação da polícia. 2
1.2 Ilustração de rota utilizando o *Google Maps*. 3
1.3 Modelo de referência de visualização traduzido de Card et al. (1999). 4

2.1 Ilustração de um gráfico de dispersão bidimensional. 8
2.2 Ilustração de um gráfico de dispersão estendido. 8
2.3 Representação da técnica de coordenadas paralelas. 9
2.4 Ilustração do problema de oclusão visual da técnica de coordenadas paralelas. . . 11
2.5 Representação da técnica matriz de gráficos de dispersão. 12
2.6 Representação visual gerada na ferramenta *TableLens*. 13
2.7 Representação visual gerada utilizando *Starplot*. 14
2.8 Representação visual gerada utilizando *Chernoff faces*. 15
2.9 Ilustração de um arranjo de janelas para uma base com seis atributos. 15
2.10 Formato circular para janelas. 16
2.11 Ilustração de esquemas de arranjos de *pixels*. (Fonte: Keim e Kriegel (1996)) . . 16
2.12 Aplicativo *Market Radar*. 17
2.13 Mapa visual projetado com a técnica IDMAP. 19
2.14 Mapa visual projetado com a técnica LSP. 20
2.15 Relação de vizinhança para um conjunto de 6 pontos. 21
2.16 Matriz A resultante (com pontos de controle p_3 e p_6). 22
2.17 Visão geral da ferramenta *Projection Explorer*. 22
2.18 Processo de projeção de bases de dados multidimensionais da PEx. 23

3.1 Técnica de coordenadas paralelas para visualização de dados geo-referenciados. . 27
3.2 Representações gráficas sobre o mapa (Andrienko e Andrienko, 2005). 28
3.3 Ferramenta *Power Map Explorer* (Silva et al., 2007). 29
3.4 Ferramenta LAHVA. 30
3.5 Representações gráficas da ferramenta LAHVA (Maciejewski et al., 2008). 31
3.6 Representações gráficas da ferramenta LAHVA (Maciejewski et al., 2008). 31
3.7 Representações gráficas da ferramenta *Taggram* (Nguyen e Schumann, 2010). . . 32
3.8 Exemplo de um gráfico de linha. 33
3.9 Exemplo da técnica Coordenadas Paralelas 34
3.10 Exemplo da técnica *Cluster and Calendar View* (Wijk e Selow, 1999). 35
3.11 Exemplo da técnica GROOVE (Lammarsch et al., 2009). 36

3.12 Exemplo da técnica de projeção multidimensional (Alencar et al., 2007). 36
3.13 Tela da ferramenta *Temporal-PEx* (Alencar et al., 2007). 37
3.14 Ferramenta *Gapminder Trendalyzer*. 38
3.15 Evolução temporal do Brasil utilizando a ferramenta *Gapminder Trendalyzer*. . . 39

4.1 Etapas da técnica LSP_{FCP}. 43
4.2 Mapa cartesiano ilustrando as coordenadas geográficas. 43
4.3 Ilustração da justaposição da sequência temporal de projeções. 45
4.4 Diferentes abordagens na geração de sequências temporais de projeções. 47
4.5 Gráfico *Neighborhood Hit* 48
4.6 Gráfico *Neighborhood Preservation* 48
4.7 Projeções dos dados do IDH em 2000 e 2005 utilizando a LSP. 49
4.8 Projeções dos dados do IDH em 2000 e 2005 utilizando a LSP_{FCP}. 50
4.9 Ilustração da animação dos pontos. 51
4.10 Ilustração do problema de oclusão visual. 52
4.11 Ilustração da utilização de transparência para indicar a direção de movimento. . 52
4.12 Ilustração da utilização de setas para indicar a direção de movimento. 53
4.13 Projeções dos dados do IDH utilizando a técnica LSP_{FCP}. 53
4.14 Projeções dos dados do IDH utilizando a técnica LSP_{FCP}. 54
4.15 *Projection wizard*: escolha da base de dados. 54
4.16 *Projection wizard*: escolha da técnica e da métrica de distância. 55
4.17 Escolha dos parâmetros da técnica LSP_{FCP}. 55
4.18 Projeções geradas utilizando a técnica LSP_{FCP} justapostas. 57
4.19 Exemplo ilustrando a transição dos pontos entre as projeções. 57
4.20 *Projection Wizard*: escolha da base de dados. 58
4.21 *Projection Wizard*: escolha da técnica e da métrica de distância. 59
4.22 *Projection Wizard*: escolha dos parâmetros da técnica LSP_{FCP}. 59
4.23 Ferramenta *Temporal-PEx* com a visualização da informação geográfica. 60
4.24 Visualização da informação geográfica. 60
4.25 Projeção utilizando linhas com gradiente para o traçado do caminhos dos pontos. 61

5.1 Módulo *PExParser*. 64
5.2 Módulo PExParser: criação do arquivo de pontos de controle. 64
5.3 Sequência temporal de projeções para os dados do IDH. 67
5.4 Legenda das cores nos mapas da sequência temporal. 68
5.5 Formação de dois grupos de países. 68
5.6 Observação de Portugal e da Coreia do Sul ao longo do tempo na base do IDH. . 68
5.7 Portugal e Coreia do Sul apresentam comportamento diferente. 69
5.8 Análise dos dados de Portugal e Coreia do Sul. 69
5.9 Comparação do IDH em 1980 e 2005 utilizando geografia. 70
5.10 Legenda das cores nos mapas da sequência temporal. 71
5.11 Sequência temporal de projeções para os dados de economia. 72
5.12 Efeito do PIB no posicionamento. 73
5.13 Efeito do PIB no posicionamento, grupo misto. 74
5.14 Ilustração do rastro de Macau (China) entre os anos de 1995 e 2005. 75
5.15 Variação do PIB de Macau entre 1995 e 2005. 76
5.16 Sequência temporal de projeções para os dados do IDH do *Gapminder*. 77
5.17 Renda per capita determinando o posicionamento. 78
5.18 Ilustração da tendência de alguns países. 79

6.1 Ilustração da aplicação de *bundling* nas arestas da sequencia temporal. 84

Lista de Tabelas

4.1 Escolha dos pontos de controle para gerar as sequências temporais. 45
4.2 Exemplo de formato do arquivo LCP. 56

Lista de Siglas

API	*Application Programming Interface*
CDM	*Compression-based Dissimilarity Measure*
CSV	*Comma-Separated Values*
DTW	*Dynamic Time-Warping*
ECN	*Netherlands Energy Research Foundation*
GROOVE	*Granulary Overview Overlay*
ICMC	Instituto de Ciências Matemáticas e de Computação
IDH	Índice de Desenvolvimento Humano
IDMAP	*Interactive Document Map*
LCP	*Local Control Points*
LSP	*Least Square Projection*
LSP$_{FCP}$	*Least Square Projection* com Pontos de Controle Fixos
PEx	*Projection Explorer*
PIB	Produto Interno Bruto
PNUD	Programa das Nações Unidas para o Desenvolvimento
OECD	*Organization for Economic Co-operation and Developmet*
ONU	Organização das Nações Unidas
SDK	*System Development Kit*
Temporal-PEx	*Temporal Projection Explorer*
WWJ	*World Wind Java*

Capítulo

1

Introdução

1.1 Contexto e motivação

A Visualização Computacional tem por objetivo traduzir dados e informações para um formato gráfico, gerando representações visuais de conjuntos de dados que permitam aos seres humanos extrair, de forma interativa e exploratória, informações úteis (Oliveira e Levkowitz, 2003). Basicamente, ela pode ser dividida em três sub-áreas (Oliveira e Minghim, 1997):

- **Visualização Científica:** Trata da visualização de dados que possuem uma representação geométrica inerente. Normalmente estes dados são medidos ou simulados e representam fenômenos físicos, químicos ou biológicos.
- **Visualização de Informação:** Estuda representações gráficas para dados que não possuem uma geometria associada e, geralmente, são heterogêneos, multidimensionais e volumosos.
- **Visualização de Software:** Estuda ferramentas que auxiliam o desenvolvimento de programas e sistemas de *software* em todas as suas etapas, por meio de recursos visuais.

As dificuldades associadas à gestão de dados permeiam diferentes disciplinas e domínios, tanto no cenário da pesquisa acadêmica como nos setores de serviços, industrial e agropastoril, e a visualização assume um papel importante nesse contexto.

Nas últimas décadas, a quantidade de informação que temos que processar tem aumentado enormemente. Porém, o que temos realmente é uma explosão de dados, e não uma explosão de

informação (Mazza, 2009). A informação, tão valiosa e importante em nossas vidas, é construída através deste constante fluxo de dados que recebemos passiva ou ativamente. Assim, tornam-se necessários métodos eficientes que apoiem tomadas de decisão a partir da análise destes dados.

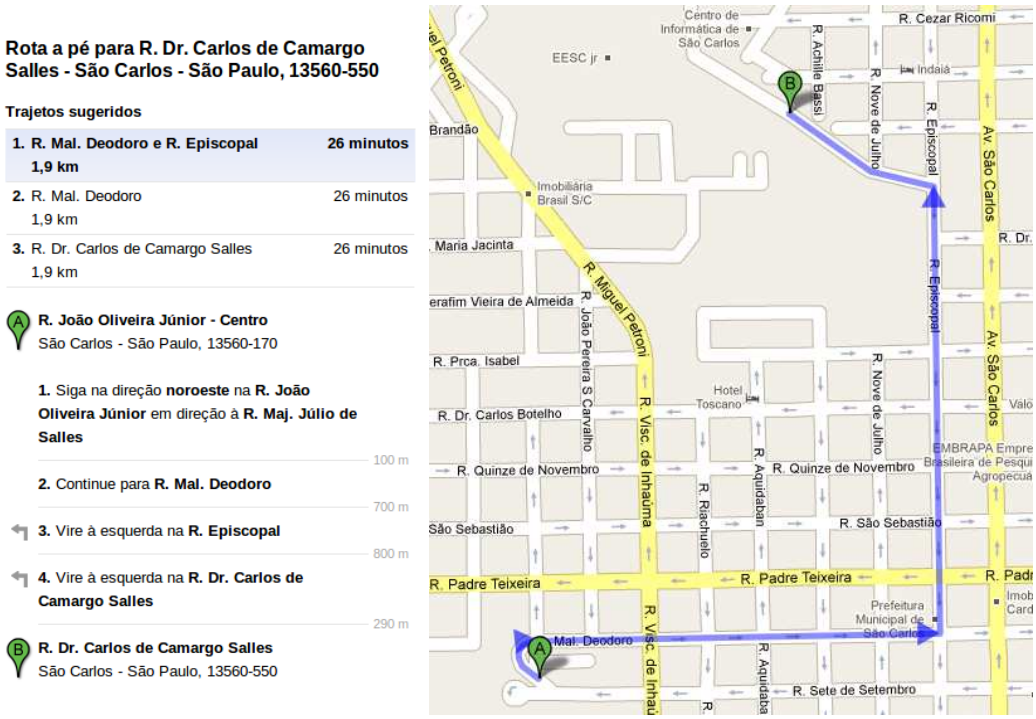
Andrienko e Andrienko (2005) associam duas componentes aos dados: referências e atributos. Os atributos refletem observações ou medições, enquanto as referências especificam o contexto em que estas observações ou medições foram feitas.

Muitos fenômenos importantes apresentam, naturalmente, evolução temporal, cuja compreensão é determinante para entender e prever o seu comportamento. Assim, não é surpreendente o fato deste tipo de dados ser recorrente, já que são obtidos como resultado em várias áreas, como biologia, finanças, geologia, análise de movimentos humanos, entre outras. Sua ampla presença em áreas de aplicação diversas confere a esse tipo de dado uma grande importância. A alta dimensionalidade é um dos fatores que dificultam a análise de dados dessa natureza, em sua forma bruta.

A geografia, ou o local da informação, também é muitas vezes relevante para um melhor entendimento dos dados e tomar decisões mais precisas e mais informadas. Transações feitas com cartões de crédito, ligações telefônicas e fenômenos meteorológicos são exemplos de domínios que envolvem questões de geografia. Como exemplo, a Figura 1.1 ilustra acontecimentos de uma onda de violência promovida por criminosos no Rio de Janeiro e a resposta da polícia. A informação geográfica permite identificar as regiões e locais onde os ataques se concentraram.

Figura 1.1: Ataques no Rio de Janeiro e reação da polícia.
(Fonte: http://oglobo.globo.com/rio/info/ataques-no-rio/ acessado em 20 de setembro de 2011)

Graças a nossa percepção visual, uma representação gráfica pode ser muito mais eficiente do que representações textuais (Mazza, 2009). Uma pessoa que precisa fazer uma viagem de carro, por exemplo, pode obter a rota da viagem de forma escrita ou visual (ver Figura 1.2). O site *Google Maps*[1] fornece um serviço muito útil ao mostrar as melhores rotas possíveis. Uma rota é representada de duas maneiras: em uma descrição textual com as distâncias (ver Figura 1.2(a)), os nomes das rodovias e as direções a seguir, e em uma representação gráfica no mapa (ver Figura 1.2(b)).



(a) (b)

Figura 1.2: Ilustração de rota: em (a) na forma escrita, e (b) na forma visual.
(Fonte: http://maps.google.com.br acessado em 20 de setembro de 2011)

Mazza (2009) nos diz que cada representação tem sua utilidade, porém a representação visual tem a vantagem de usar propriedades gráficas que podem ser rapidamente processadas pelo canal visual. Atributos como cor, tamanho, proximidade e movimento são imediatamente absorvidos pela habilidade de percepção visual mesmo antes do complexo processo cognitivo da mente humana começar a funcionar.

Muitas vezes, as pessoas querem observar um conjunto de dados sem ter um objetivo específico em mente (Brodbeck et al., 2009). Através do exame visual dos dados as pessoas podem aprender mais sobre eles, obter um conhecimento mais profundo, descobrir fatos inesperados, ou seja, obter um *insight*. O cérebro humano é uma máquina que processa continuamente uma grande quantidade de dados e informação. Representações visuais nos permitem entender sistemas complexos, tomar decisões e encontrar informação que, de outra maneira, ficaria escondida

[1] http://maps.google.com.br acessado em 20 de setembro de 2011

nos dados. Ferramentas de análise mais efetivas assumem um papel fundamental nesse contexto, sob a perspectiva da gestão pelo ser humano.

Segundo Card et al. (1999), a visualização pode ser descrita como uma sequência de mapeamentos ajustáveis de dados para uma representação visual. A Figura 1.3 ilustra esta sequência. Primeiramente, os dados em sua forma original são transformados em um conjunto de relações. A partir do conjunto de relações pode-se fazer a escolha de estruturas visuais para mapear a informação contida nos dados. O usuário deve ser capaz de interagir com as visões, assim como em todo o seu processo de geração.

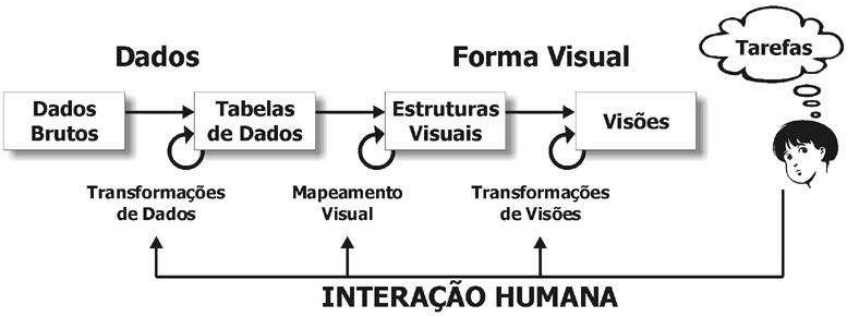


Figura 1.3: Modelo de referência de visualização traduzido de Card et al. (1999).

A área de visualização de informação trata da visualização de dados abstratos (Card et al., 1999). Estes dados podem variar de coleções de registros a dados não estruturados, como documentos, imagens, séries temporais e resultados de busca na Web. Em um primeiro passo, estes dados são convertidos em espaços vetoriais de alta dimensionalidade, tabelas formadas por itens (linhas) e atributos (colunas).

Técnicas de visualização que permitem criar mapas visuais de conjuntos de dados multidimensionais em geral vem sendo desenvolvidas e investigadas. Grinstein et al. (2001) apresentam algumas das técnicas mais clássicas: Gráficos de dispersão (Cleveland, 1993), *Table lens* (Rao e Card, 1994), Coordenadas Paralelas (Inselberg e Dimsdale, 1990), técnicas iconográficas (Chernoff, 1973), técnicas orientadas a *pixel* (Keim e Kriegel, 1996; Keim, 2000) etc. A maioria delas mapeia cada elemento do conjunto de dados em um elemento gráfico correspondente, podendo ser um *pixel*, um segmento de reta, um ícone, entre outros.

Como alternativa aos métodos de visualização tradicionais, as técnicas de projeção de dados multidimensionais, ou apenas técnicas de Projeção Multidimensional (Tejada et al., 2003), vêm sendo aplicadas com sucesso. O desenvolvimento e investigação deste tipo de técnica tem sido o foco principal do grupo de pesquisa em Visualização de Informação e Computação Gráfica do ICMC (Paulovich et al., 2011a,b; Siqueira et al., 2010; Eler et al., 2009a,b; Felizardo et al., 2009; Paulovich et al., 2008).

Os mapas gerados pelas técnicas de Projeção Multidimensional são representações visuais interativas criadas a partir do posicionamento de pontos no espaço 2D. As projeções geram um layout 2D das instâncias multidimensionais que reflete sua similaridade, e os mapas obtidos podem ser explorados por usuários interativamente em busca de grupos de interesse. Com isso,

um usuário pode explorar uma grande coleção de dados sem necessidade de um processamento sequencial dos elementos, em busca dos que realmente lhe interessam.

As técnicas de Projeção Multidimensional oferecem um arcabouço genérico para a visualização de dados de alta dimensionalidade, o qual é utilizado em uma ferramenta de mineração visual chamada de *Projection Explorer* (Paulovich et al., 2007). A representação gerada fornece um mapa do conjunto de dados onde os elementos estão agrupados conforme a sua similaridade, computada segundo alguma medida apropriada. Esse mapa oferece um ponto de partida para uma exploração mais detalhada.

1.2 Objetivos

O principal objetivo deste trabalho é utilizar as referências, especificamente espaço e tempo, associadas aos dados e quase sempre utilizadas como atributos, em processos de análise apoiados por técnicas de projeção multidimensional. Pretende-se investigar e propor modificações em técnicas de projeção multidimensional que permitam ao usuário realizar análises exploratórias em grandes conjuntos de dados multidimensionais considerando explicitamente as referências associadas (temporais e geográficas).

Como objetivos secundários temos:

- Extensão de uma técnica de projeção multidimensional para tratar dados que possuam alguma referência, geográfica ou temporal, associada.
- Investigar o uso de animação como funcionalidade auxiliar para a análise da evolução temporal gerada por projeções.

1.3 Organização da dissertação

Neste documento apresentamos o trabalho desenvolvido, a técnica e as funcionalidades implementadas, sendo este estruturado da seguinte maneira:

- No Capítulo 2 são apresentadas algumas técnicas bem conhecidas para a visualização de dados multidimensionais. A ferramenta *Projection Explorer*, que servirá de base para o desenvolvimento do trabalho, é apresentada em detalhes.
- No Capítulo 3 são descritas técnicas de visualização e demais estratégias aplicadas a dados com algum tipo de referência geográfica ou temporal, e algumas ferramentas que utilizam estas técnicas.
- O Capítulo 4 apresenta a técnica *Least Square Projection* com Pontos de Controle Fixos (LSP_{FCP}), desenvolvida para auxiliar a análise de dados multidimensionais que apresentam referência geográfica ou temporal associada.

- No Capítulo 5 são apresentados os resultados obtidos utilizando as novas funcionalidades em alguns casos de estudo.
- Por fim, no Capítulo 6, a conclusão e sugestões de trabalhos futuros são apresentadas.

Capítulo

2

Visualização de dados multidimensionais

2.1 Considerações iniciais

A análise visual permite extrair informação e conhecimento de grandes volumes de dados, combinando métodos de análise com as facilidades da visualização e interação do usuário (Schreck et al., 2007).

Segundo Mazza (2009), a análise exploratória de dados é uma das aplicações que mais se beneficia de representações visuais e da capacidade de análise pela percepção visual do sistema cognitivo humano.

A análise de dados multidimensionais é um tópico importante na área de visualização e enfrenta dois problemas: apresentar em espaços bidimensionais, ou tridimensionais, dados com múltiplas dimensões, e a dificuldade dos seres humanos em entender estes tipos de dados.

A utilização de gráficos de dispersão (Cleveland, 1993) como representação visual é muito simples e intuitiva quando se tem dois ou três atributos. Os valores são mapeados ao longo dos eixos cartesianos. Como exemplo, a Figura 2.1 ilustra um gráfico de dispersão bidimensional. Cada círculo representa um produto vendido no McDonald's Brasil. No eixo X são mapeados os valores calóricos em quilocalorias, e no eixo Y são mapeados os valores de gordura, em gramas, presentes nos produtos. É fácil observar que o produto menos saudável do cardápio é o *Big Tasty*.

Entretanto o número de situações envolvendo apenas dois ou três atributos é muito limitado, a maioria dos problemas reais resulta em dados com diversos atributos para analisar. Os gráficos de dispersão podem ser estendidos para mapear até sete atributos, adicionando mais um eixo

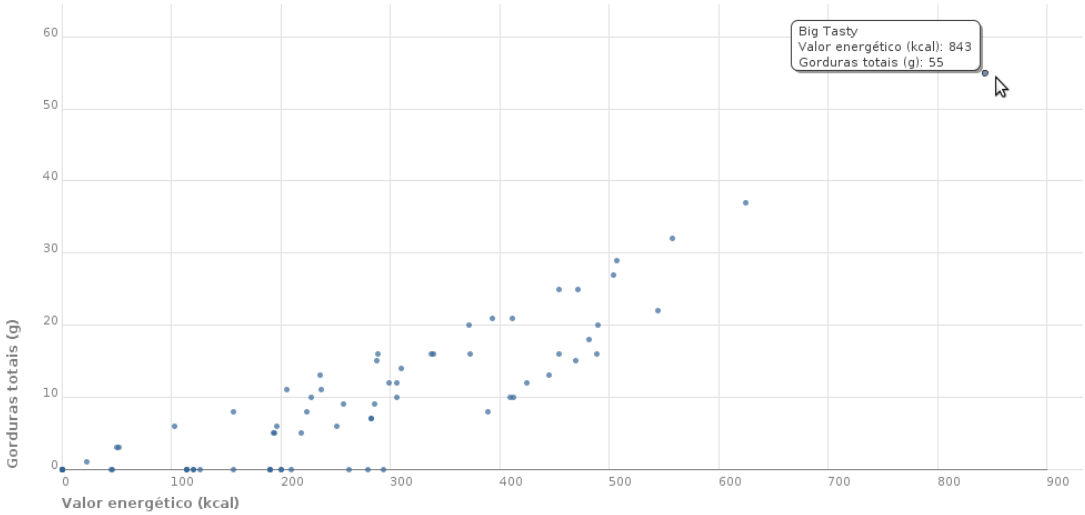


Figura 2.1: Ilustração de um gráfico de dispersão bidimensional.
(Fonte: http://www-958.ibm.com/software/data/cognos/manyeyes/visualizations/calorias-e-gorduras-dos-produtos-m acessado em 20 de setembro de 2011)

representado pela perspectiva da imagem e mapeando atributos para propriedades visuais dos marcadores gráficos como formato, dimensão, cor e textura. A Figura 2.2 ilustra um gráfico de dispersão que representa o conjunto de dados da Figura 2.1, estendido para mapear os valores de gordura trans (em gramas) presentes nos produtos para o diâmetro dos círculos. Porém, mesmo estendendo os gráficos de dispersão, estas representações não são adequadas para todos os tipos de problemas e de dados.

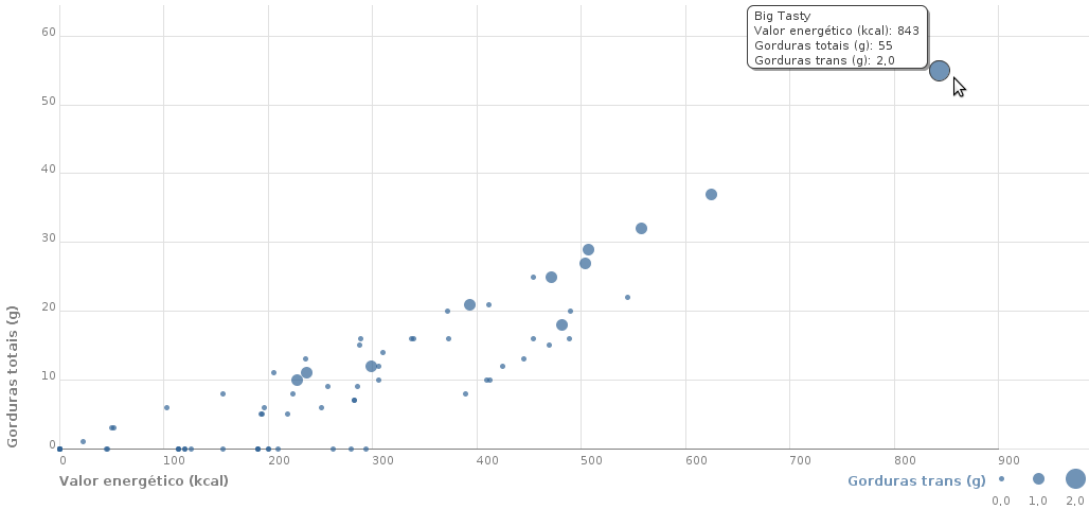


Figura 2.2: Ilustração de um gráfico de dispersão estendido.
(Fonte: http://www-958.ibm.com/software/data/cognos/manyeyes/visualizations/calorias-e-gorduras-dos-produtos-m acessado em 20 de setembro de 2011)

Este capítulo mostra algumas das principais técnicas de visualização de dados multidimensionais. Seguindo a divisão encontrada em Mazza (2009), na Seção 2.2 são apresentadas algumas técnicas classificadas como projeções geométricas. Em seguida, a Seção 2.3 apresenta as técnicas

iconográficas e a Seção 2.4 apresenta as técnicas orientadas a *pixels*. A Seção 2.5 descreve algumas técnicas de projeções multidimensionais e mostra a ferramenta *Projection Explorer*, que serve de base para o desenvolvimento deste projeto.

2.2 Técnicas geométricas

Técnicas geométricas consistem em mapear os valores dos atributos dos dados em espaços geométricos. Gráficos de dispersão fazem parte desta categoria de técnicas. Outras técnicas bem conhecidas em visualização de informação são apresentadas a seguir.

2.2.1 Coordenadas Paralelas

Nesta técnica, cada atributo corresponde a um eixo e os eixos são dispostos de forma paralela e com espaçamento igual entre si (Inselberg e Dimsdale, 1990; Wegman, 1990). Cada registro da base de dados é representado por um conjunto de segmentos de reta que ligam os valores dos atributos em cada eixo.

Como exemplo, a Figura 2.3 ilustra a utilização da técnica de coordenadas paralelas para representar uma base de dados contendo especificações técnicas de 392 modelos de carros produzidos entre 1970 e 1990, em que são observados sete atributos.

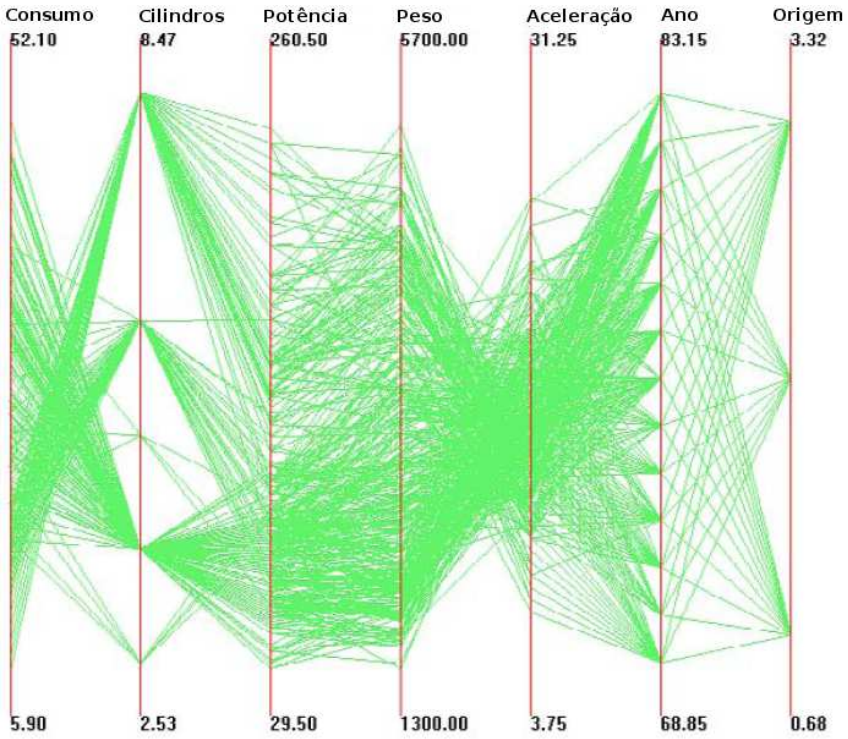


Figura 2.3: Representação da técnica de coordenadas paralelas.
(Fonte: Traduzido de Mazza (2009), página 49)

O exemplo mostra como esta pode ser uma ferramenta poderosa para a análise exploratória de dados pois facilita a observação de correlação entre os atributos. Na Figura 2.3 é possível observar que existe uma relação inversa entre o consumo de combustível e o número de cilindros: quanto maior a quantidade de cilindros, menor é o consumo de combustível.

Porém, esta técnica pode apresentar alguns problemas para grandes bases de dados. Em alguns casos, a representação visual pode estar tão condensada que chega a parecer um polígono de cor uniforme. Este é um problema de qualquer representação visual: o espaço disponível na tela pode ser insuficiente para conter todos os elementos visuais. O arranjo dos eixos também é um fator importante para a análise. A observação de dependência entre os atributos representados por eixos imediatamente adjacentes é óbvia, enquanto que a observação da relação entre atributos representados por eixos distantes pode até não ser possível.

A Figura 2.4 ilustra o problema da oclusão visual causado pela grande quantidade de dados na técnica de coordenadas paralelas. A base utilizada para gerar esta representação corresponde ao censo do Reino Unido relativo ao ano de 1991. Cada elemento do conjunto corresponde a um bairro. As variáveis observadas são as seguintes:

- CROWD: Corresponde a proporção de domicílios com mais de um habitante por dormitório.
- Densidade: Densidade habitacional medida em milhões por metro quadrado.
- LLTI: Porcentagem de pessoas por domicílio que possuem algum tipo de invalidez.
- SC1: Proporção de chefes de família que possuem ocupações gerenciais.
- SPF: Proporção de famílias com pai ou mãe solteiros por área.
- UNEMP: Proporção de homens desempregados por área.

Soluções como a possibilidade de mover os eixos ou colorir os segmentos de reta que representam um ou mais atributos podem ser adotadas para contornar alguns dos problemas mencionados.

2.2.2 Matriz de gráficos de dispersão

Matrizes de gráficos de dispersão representam uma extensão interessante dos gráficos de dispersão bidimensionais (Mazza, 2009). São simples e intuitivas para representar um número genérico de atributos. Esta técnica simples consiste em representar pares de atributos, utilizando gráficos de dispersão bidimensionais, e colocá-los lado a lado compartilhando os mesmo eixos. Desta maneira, uma matriz $M \times M$ é formada, onde M é o número de atributos representados.

A Figura 2.5 ilustra a aplicação da técnica para representar a base de dados Duncan. Esta base é composta pela observação da renda, educação e prestígio de 45 profissões nos Estados Unidos no ano de 1950. Esta figura pode ser obtida com a plataforma R (ver Hornik (2010)) utilizando o comando:

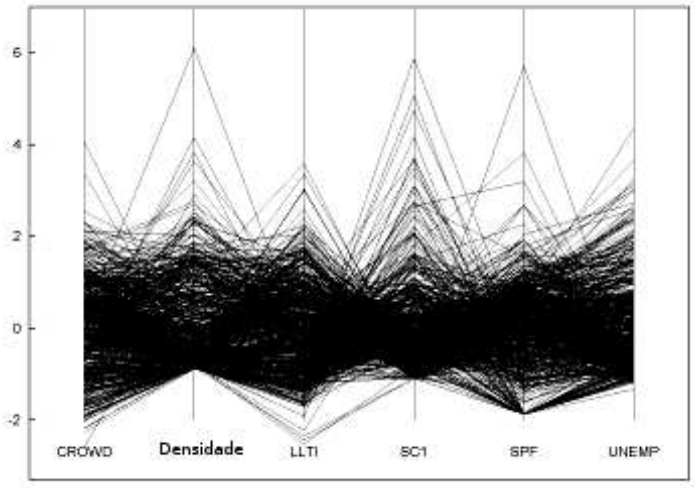


Figura 2.4: Ilustração do problema de oclusão visual da técnica de coordenadas paralelas.
(Fonte: Traduzida de http://www.agocg.ac.uk/reports/visual/casestud/brunsdon/parallel.htm acessado em 20 de setembro de 2011)

```
scatterplotMatrix(~ income + education + prestige, var.labels = c("renda",
                  "educação", "prestígio"), data=Duncan)
```

Poderíamos realizar uma análise destes dados utilizando coordenadas paralelas. Porém, para analisar várias dependências entre pares de atributos poderiam ser necessárias várias interações, como movimentações nos eixos, por exemplo. A utilização de matriz de gráficos de dispersão torna a correlação entre os pares imediatamente visível, sem necessidade de ajustar a representação visual.

Por outro lado, as matrizes de gráficos de dispersão podem apresentar alguns inconvenientes. Por exemplo, para N atributos é necessário formar uma matriz $N \times N$, limitando o espaço disponível para representar os pontos.

2.2.3 TableLens

Esta estrutura foi inspirada nas aplicações de planilha, mas sua principal característica é a representação dos dados utilizando barras horizontais ao invés de valores numéricos. A ferramenta *TableLens* (Rao e Card, 1994) utiliza esta proposta de representação visual para auxiliar a análise de dados.

Os dados são representados em uma matriz, em que os atributos são representados nas colunas e cada elemento da base de dados corresponde a uma linha da matriz. Os valores numéricos de cada atributo são mapeados para o comprimento das barras horizontais. Visualmente, as barras horizontais podem ser representadas em um espaço muito limitado. Desta forma, é possível representar uma grande quantidade de atributos e elementos em uma única tela e permitir que possíveis padrões, tendências e relacionamentos entre os atributos sejam identificados imediatamente.

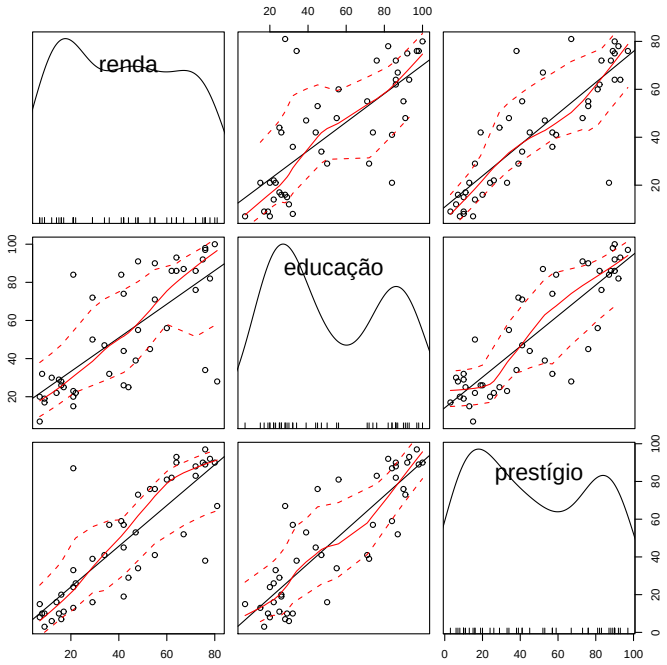


Figura 2.5: Representação da técnica matriz de gráficos de dispersão.

A Figura 2.6 ilustra uma representação visual utilizando a ferramenta *TableLens* de uma base composta por estatísticas individuais dos jogadores de *baseball* no ano de 1986. Especificamente, esta base contem observações de 23 atributos de 323 jogadores.

Uma característica interessante deste tipo de representação é a possibilidade de interação do usuário com a visualização. Essa interação pode ser realizada por exemplo manipulando as colunas, alterando sua ordem, escondendo, exibindo ou ordenando seus valores.

2.3 Técnicas iconográficas

Outro conjunto de técnicas que utilizam as propriedades geométricas de uma imagem para representação de dados multidimensionais são as técnicas iconográficas. O nome se dá pelo fato de que uma figura geométrica (um ícone, que também pode ser chamado de *glyph*) possui um número de características que podem variar: cor, forma, tamanho, orientação, etc. Cada atributo é associado a uma característica na figura geométrica. Mostraremos duas técnicas bastante conhecidas: *Starplot* e *Chernoff faces*.

2.3.1 Starplot

Uma figura geométrica simples e relativamente intuitiva é representada por um polígono em forma de estrela, cujos vértices são definidos por uma coleção de eixos posicionados radialmente (Nascimento e Ferreira, 2006). Cada elemento do conjunto de dados pode ser representado

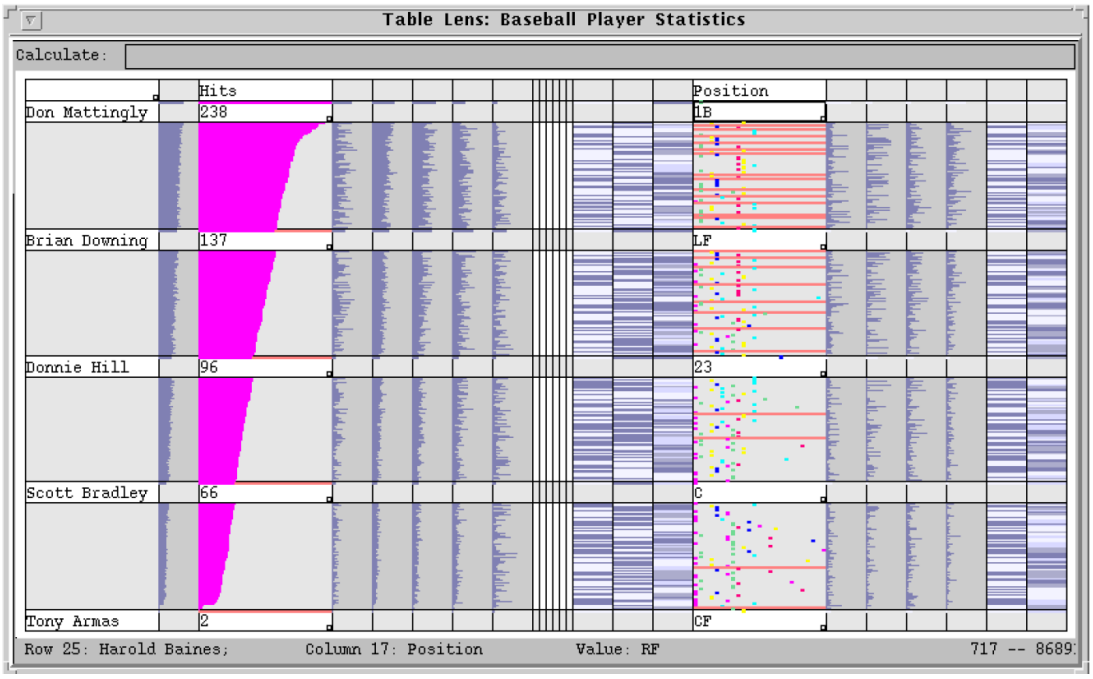


Figura 2.6: Representação visual gerada na ferramenta *TableLens*.
(Fonte: Rao e Card (1994))

por uma estrela com os valores para cada atributo sendo mapeados para o comprimento de um raio. Pela junção dos pontos que correspondem ao comprimento de cada raio, uma figura geométrica é obtida. O formato dessa figura descreve o elemento da base de dados. Esta técnica é uma combinação da técnica *Star Coordinates* (Lee et al., 2003) com *glyphs* e com *Small Multiples* (Tufte, 2001).

A Figura 2.7 ilustra uma representação visual utilizando a técnica *Starplot* para uma base composta pela observação de 9 atributos (preço, consumo, consertos em 1978, consertos em 1977, altura livre acima da cabeça, espaço do banco traseiro, espaço do porta-malas, peso e comprimento) de 16 tipos de carros. Cada ícone representa um carro e cada raio um atributo. É fácil observar que o Cadillac Seville é um dos carros mais caros. Pode-se também comparar os modelos AMC com os modelos Cadillac e observar que os modelos AMC tendem a ser baratos, com baixo consumo e pouca manutenção.

Esta técnica, também conhecida como *star glyph*, pode ser muito útil para comparar diferentes elementos da base de dados pela simples comparação da semelhança dos formatos de cada *glyph*.

Um limite óbvio deste tipo de representação está na escalabilidade. Para um número de elementos que é excessivamente grande, o espaço ocupado na tela torna-se tão denso que não permite a criação de várias formas de ícones. As técnicas iconográficas são aplicáveis quando se quer explorar o aspecto qualitativo dos atributos, e não o quantitativo.

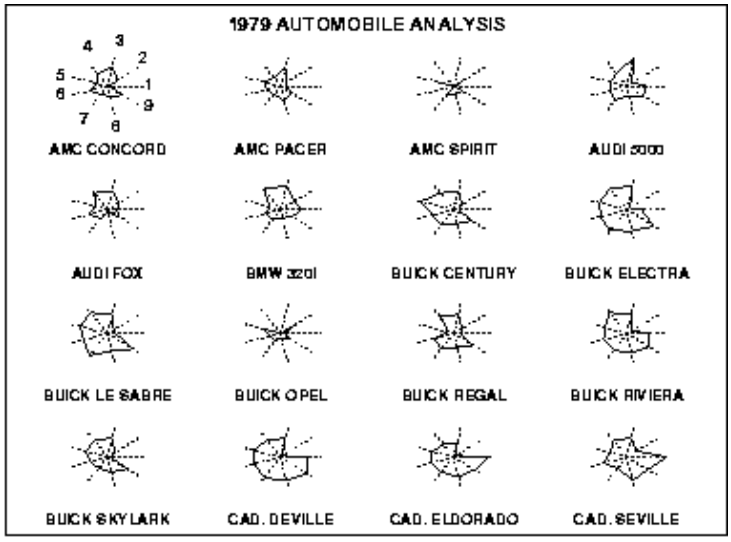


Figura 2.7: Representação visual gerada utilizando *Starplot*.
(Fonte: http://www.itl.nist.gov/div898/handbook/eda/section3/starplot.htm acessado em 20 de setembro de 2011)

2.3.2 Chernoff faces

Como os humanos possuem uma capacidade particular de perceber até mesmo a menor alteração em expressões faciais, Chernoff (1973) propôs que os atributos de uma coleção de dados fossem mapeados para a forma, as dimensões e a orientação das características faciais humanas, como olhos, nariz, boca, orelhas etc. Essa forma de representação dos dados torna essa técnica bastante interessante.

A Figura 2.8 apresenta uma representação visual obtida utilizando a técnica *Chernoff faces* pela ferramenta VisEd (Xavier, 2004). A base utilizada é composta por dados de alunos como sexo, quantidade de acesso ao site de um curso e notas (ver Nascimento e Ferreira (2006)). Cada variável foi mapeada para um atributo visual do rosto: o número de acessos ao site de um curso é representado pela quantidade de cabelo; a nota como a inclinação da boca; e o sexo como a cor da face. Os ícones estão classificados por ordem decrescente de notas. É fácil observar que as notas mais baixas são dos alunos que menos acessaram o site do curso.

Este tipo de mapeamento utiliza características faciais familiares e nos é imediatamente perceptível. Porém, a escolha do mapeamento entre atributos e elementos da face, se mal realizada, pode levar a observações incorretas.

2.4 Técnicas orientadas a pixels

Algumas técnicas utilizam os *pixels* de uma tela como unidades básicas de representação, dessa forma podem maximizar o número de elementos que serão representados. Se uma tela de computador, com uma resolução de 1024×768 *pixels*, fosse utilizada somente para representar elementos de um conjunto de dados unidimensional mapeando o valor do atributo do dado com a cor do *pixel* correspondente ao elemento, seria possível representar até 786.432 elementos dife-

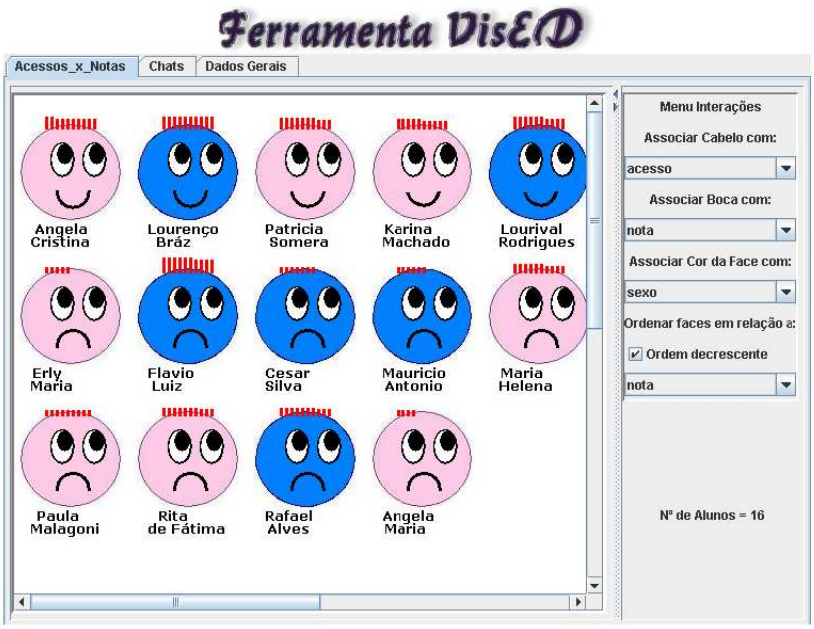


Figura 2.8: Representação visual gerada utilizando *Chernoff faces*.
(Fonte: Nascimento e Ferreira (2006))

rentes. Na prática existem poucas situações em que mapear apenas uma unidade de informação para a cor de um *pixel* é suficiente, pois também é necessário espaço para elementos funcionais e estéticos como botões, bordas, elementos de texto, etc (Mazza, 2009).

Para conjuntos de dados multidimensionais, a ideia é representar a maior quantidade de dados em uma única tela mapeando cada valor para a cor de um *pixel* e agrupar os dados correspondentes a um atributo em uma região específica (ver Figura 2.9).

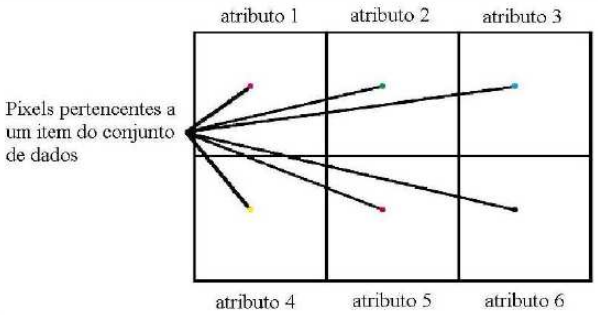


Figura 2.9: Ilustração de um arranjo de janelas para uma base com seis atributos.
(Fonte: Keim (2000))

Alguns fatores devem ser considerados na aplicação deste tipo de técnica (Keim e Kriegel, 1996; Keim, 2000):

- Formato da região: o formato retangular ainda é bastante utilizado por aproveitar o espaço físico da tela. Porém, pode tornar difícil a observação de relacionamentos entre atributos devido a distância entre as regiões. A forma circular é uma alternativa (ver Figura 2.10).

- Arranjo dos *pixels*: este mapeamento deve ser realizado buscando boas propriedades de agrupamento, mas também buscando favorecer a percepção das relações existentes entre os dados. A Figura 2.11 ilustra alguns esquemas de arranjos bem conhecidos, como a curva de Peano-Hilbert (ver Figura 2.11(b)) por exemplo.

- Mapeamento de cor: o uso de cores para mapear os valores dos atributos tem como objetivo evitar que relações entre atributos fiquem ocultas e que elementos que induzem à interpretações equivocadas sejam introduzidos.

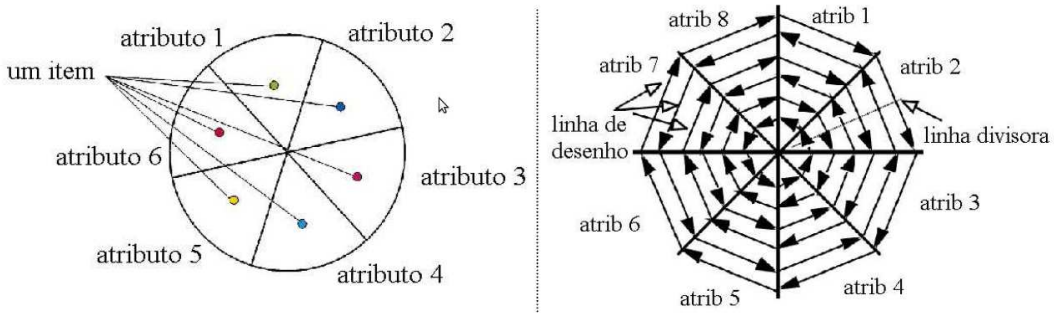
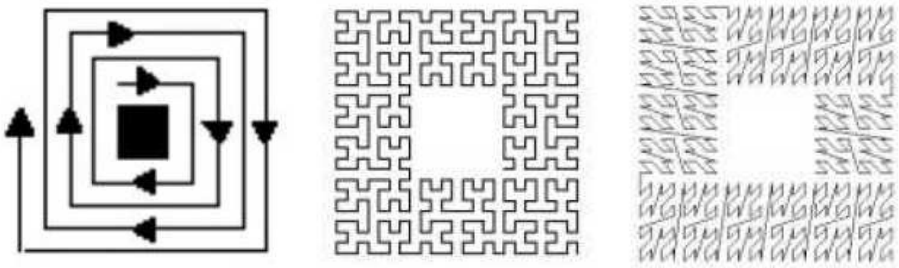


Figura 2.10: Ilustração do formato circular para janelas e ordenação dos *pixels* em cada segmento.
(Fonte: Keim (2000))



(a) Espiral simples (b) Curva de Peano-Hilbert (c) Curva de Morton

Figura 2.11: Ilustração de esquemas de arranjos de *pixels*.
(Fonte: Keim e Kriegel (1996))

Um exemplo de aplicação de técnica orientada a *pixel* é ilustrado pela Figura 2.12, e foi retidado de Mazza (2009). O *Market Radar* é um aplicativo desenvolvido pela SmartMoney [1] e realiza uma análise do histórico das ações cotadas na bolsa de valores dos Estados Unidos. Cada

[1] http://www.smartmoney.com acessado em 20 de setembro de 2011

coluna corresponde a uma semana e cada linha corresponde a uma ação, a cor do *pixel* mapeia o valor da ação. Se o valor da ação esteve em alta a cor do *pixel* é verde, se esteve em baixa a cor é vermelho, a intensidade reflete a variação do preço de uma semana para a outra.

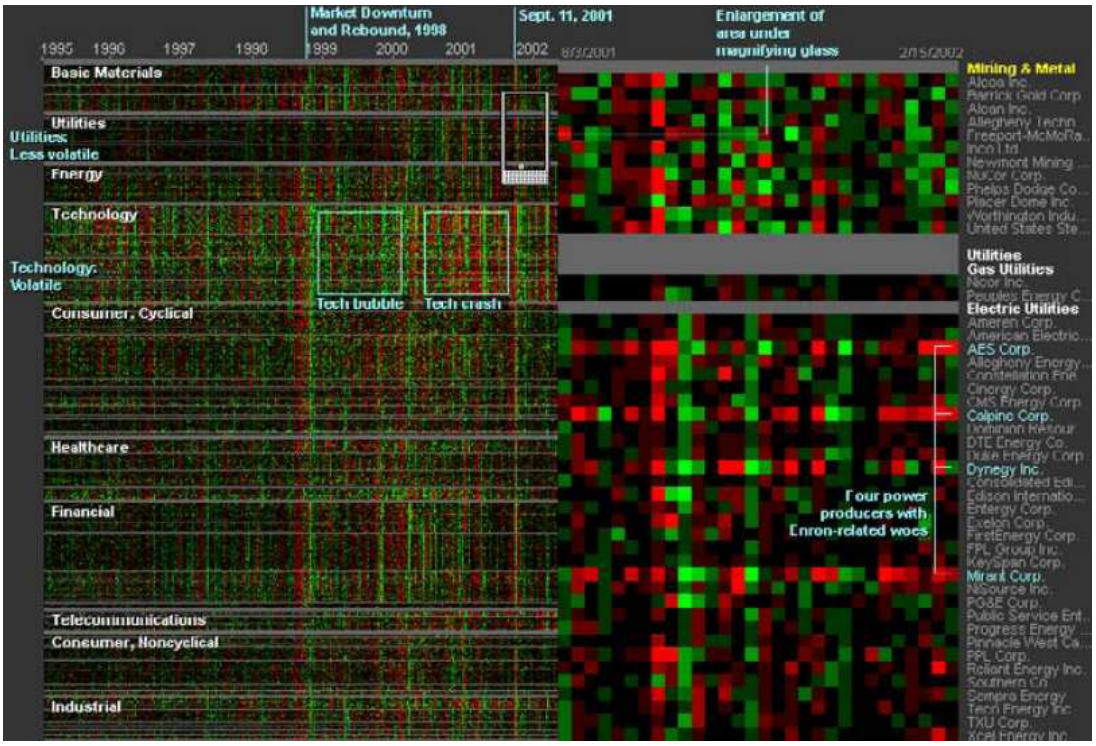


Figura 2.12: Aplicativo *Market Radar*, desenvolvido pela SmartMoney, realiza uma análise do histórico das ações cotadas na bolsa de valores. (Fonte: Mazza (2009), página 61)

2.5 Técnicas de projeção multidimensional

Para grandes conjuntos de dados multidimensionais, as técnicas de projeção proporcionam uma boa representação gráfica da distribuição dos dados no espaço original (Tejada et al., 2003). Dessa forma, são uma boa ferramenta para apoiar a análise de dados estruturados e não estruturados. Estas técnicas projetam dados originalmente definidos em um espaço m-dimensional em um espaço d-dimensional, com $d < m$ e d geralmente igual a 2 ou 3. Assim, as técnicas de projeção multidimensional realizam uma redução de dimensionalidade, partindo de um espaço m-dimensional para um espaço d-dimensional com $d \in \{1, 2, 3\}$ e $d \ll m$. Há uma tentativa de manter no espaço projetado as relações de distância entre os elementos do conjunto observadas no espaço original.

Este tipo de representação gráfica favorece a identificação de padrões globais (Paulovich et al., 2006). Recursos visuais adicionais podem ser associados a projeção para facilitar a análise, como conectar os pontos segundo algum critério ou gerar uma triangulação, colorir os elementos

de acordo com suas classes caso exista alguma classificação prévia, alterar o tamanho dos marcadores gráficos, entre outros.

As técnicas de projeção multidimensionais podem ser classificadas como sendo lineares ou não-lineares, de acordo com a função de projeção utilizada. As técnicas lineares são boas para dados normalmente distribuídos, mas são incapazes de reconhecer características como agrupamentos de formatos arbitrários. As técnicas não-lineares buscam minimizar a perda de informação que ocorre no processo de projeção. Entre as vantagens dessas técnicas pode-se destacar o fato de este tipo de função trabalhar apenas sobre as distâncias calculadas no espaço original, não necessitando trabalhar sobre os vetores originais, e requerer apenas iterações adicionais quando se quer acrescentar novos dados à projeção, sendo desnecessário realizar todos os cálculos novamente.

A seguir serão descritas duas técnicas não-lineares que foram implementadas na ferramenta *Projection Explorer* (Paulovich et al., 2007), a qual serviu de base para o desenvolvimento deste trabalho. A técnica LSP foi estendida neste trabalho, enquanto que a técnica IDMAP é utilizada na obtenção das coordenadas de tela dos pontos de controle, detalhes que serão explicados nas próximas seções. As figuras utilizadas para ilustrar cada uma das técnicas a seguir foram obtidas utilizando esta ferramenta com uma base de dados composta pela observação de quatro atributos relativos a oitenta e dois países, no ano de 1980. Os atributos observados (expectativa de vida, alfabetização de adultos, taxa de escolarização bruta e o produto interno bruto per capita) são utilizados no cálculo do Índice de Desenvolvimento Humano (IDH)[2]. As cores mapeiam a classe a que cada país pertence, segundo o critério do relatório do Programa das Nações Unidas para o Desenvolvimento: países com índice de desenvolvimento baixo (azul), médio (ciano), alto (verde) e muito alto (vermelho).

2.5.1 Interactive Document Map (IDMAP)

A técnica *Interactive Document Map* (IDMAP) (Minghim et al., 2006) é uma combinação de duas outras técnicas, em que primeiramente é gerada uma projeção inicial dos dados utilizando uma técnica de projeção rápida, como as técnicas *Fastmap* (Faloutsos e Lin, 1995) e *Nearest Neighbors Projection* (Tejada et al., 2003). Em seguida, para melhorar o posicionamento dos pontos recuperando informações perdidas durante a projeção, a técnica *Force Scheme* (Tejada et al., 2003) é utilizada. Esta técnica aproxima pontos que estão mais distantes do que deveriam estar e afasta pontos que estão mais próximos do que deveriam estar (em relação ao espaço original), iterativamente. Como exemplo, a Figura 2.13 exibe uma projeção dos atributos do IDH utilizando a técnica IDMAP.

2.5.2 Least Square Projection (LSP)

A técnica *Least Square Projection* (LSP) (Paulovich et al., 2008) busca superar o problema da distorção na projeção que pode surgir quando conjuntos de dados possuem agrupamentos

[2]Dados encontrados em http://hdr.undp.org/en/statistics/data/ acessado em 20 de setembro de 2011

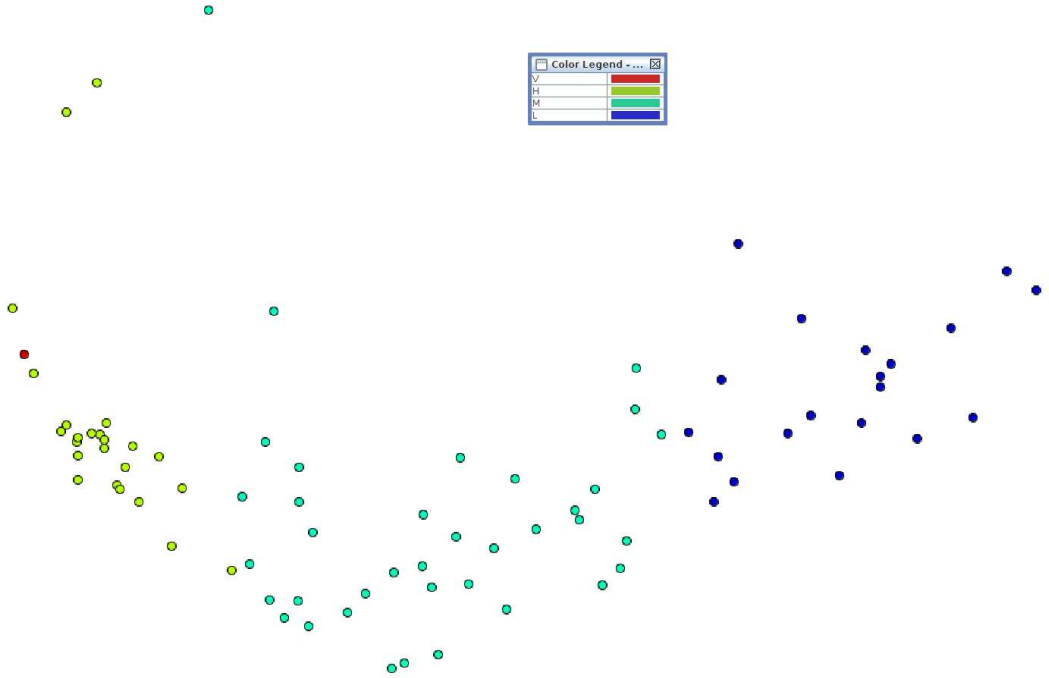


Figura 2.13: Mapa visual projetado com a técnica IDMAP.

inerentes. Esta técnica baseia-se em outra, originalmente empregada na modelagem e reconstrução de malhas de superfícies, conhecida como *least-square meshes* (Sorkine e Cohen-Or, 2004). A Figura 2.14 exibe uma projeção dos atributos do IDH utilizando a técnica LSP.

A LSP busca preservar no espaço projetado relações de vizinhança estabelecidas no espaço m-dimensional, e não relações de distância. Ou seja, busca-se garantir que os vizinhos mais próximos no espaço original também sejam os vizinhos mais próximos no espaço projetado.

Os passos realizados pela LSP podem ser resumidos em três. Dado um conjunto de pontos $S = \{p_1, \ldots, p_n\}$, com S em $\mathbb{R}^m$. Primeiro uma relação de vizinhança entre objetos do espaço original é definida. Em seguida seleciona-se um sub-conjunto de S, $S_c = \{p_{c_1}, \ldots, p_{c_{nc}}\}$, chamado de pontos de controle, e gera-se a projeção destes com uma técnica convencional como a *Fastmap* (Faloutsos e Lin, 1995). Então constrói-se um sistema linear esparso, baseado nas relações de vizinhança dos pontos em seu espaço original e nas coordenadas cartesianas dos pontos de controle projetados no espaço reduzido, para calcular as coordenas dos demais pontos que serão projetados.

A seleção dos pontos de controle pode ser feita de forma aleatória. Porém, idealmente devem ser selecionados elementos que melhor representam a distribuição dos dados no espaço original. Essa seleção pode ser realizada criando nc agrupamentos com um algoritmo de agrupamento, como o k-*means* (Berkhin, 2002) por exemplo. Depois, um objeto representativo de cada agrupamento é escolhido como ponto de controle.

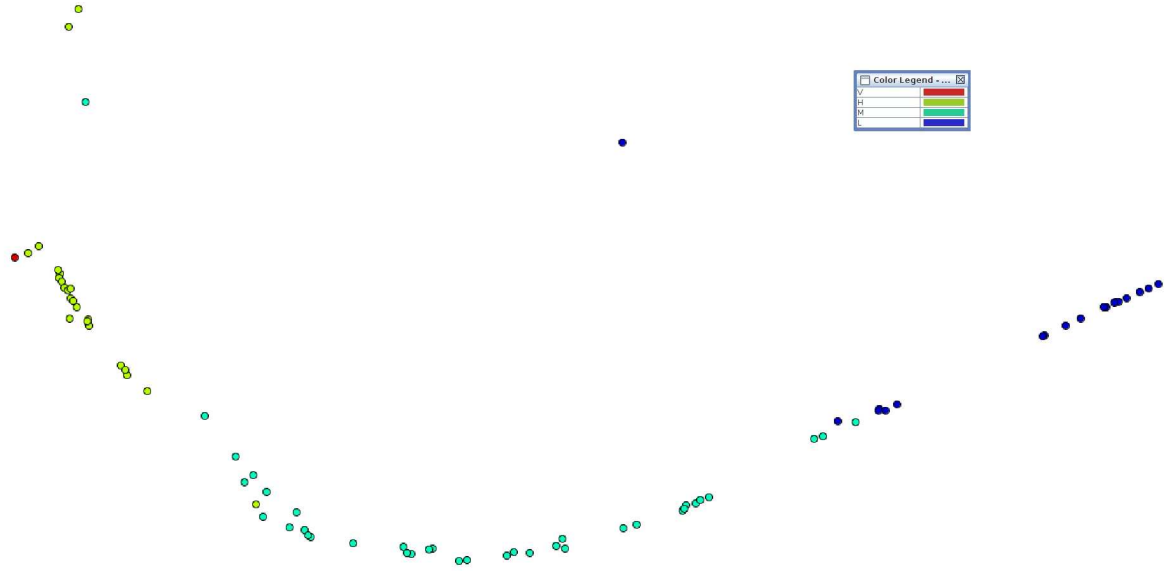


Figura 2.14: Mapa visual projetado com a técnica LSP.

Os fatores mais importantes na escolha dos pontos de controle são a quantidade e a representatividade destes nos grupos, tendo em vista que pequenas mudanças na escolha e posicionamento dos mesmos não afetam a qualidade do *layout* gerado.

Seja $V_i = \{p_{i1}, \ldots, p_{ik_i}\}$ o conjunto de k_i pontos vizinhos de um ponto p_i, e $\tilde{p}_i$ as coordenadas de p_i no espaço projetado ($\mathbb{R}^d$). Então, suponha que as posições $\tilde{p}_i$ sejam dadas pela seguinte equação:

$$\begin{aligned} \tilde{p}_i - \sum_{p_j \in V_i} \alpha_{ij} \tilde{p}_j = 0, \\ 0 \leq \alpha_{ij} \leq 1, \sum_{p_j \in V_i} \alpha_{ij} = 1. \end{aligned} \tag{2.1}$$

Se resolvermos a Equação 2.1 para os pontos em S, cada p_i será posicionado no fecho convexo dos pontos em V_i, ou seja, no menor polígono convexo que contem todos os pontos. Particularmente, quando $\alpha_{ij} = \frac{1}{k_i}$ temos p_i no centróide dos pontos em V_i (Tutte, 1963; Floater, 1997).

A Equação 2.1 resulta em um conjunto de sistemas lineares que torna possível o cálculo das coordenadas $\tilde{p}_i$:

$$L\mathrm{x}_1 = 0, L\mathrm{x}_2 = 0, \ldots, L\mathrm{x}_d = 0 \tag{2.2}$$

onde $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_d$ são vetores contendo as coordenadas cartesianas dos pontos e L é uma matriz $n \times n$, em que n é o número de pontos em S, chamada de matriz Laplaciana e seu *rank* depende da relação de vizinhança entre os pontos, cujas entradas são dadas por:

$$l_{ij} = \begin{cases} 1, & \text{se } i = j \\ -\alpha_{ij}, & \text{se } p_j \in V_i \\ 0 & \text{caso contrário} \end{cases}$$

Os pontos de controle são então utilizados para adicionar informação geométrica ao sistema. Eles são inseridos no sistema linear como novas linhas na matriz. As coordenadas cartesianas dos pontos de controle são adicionadas do lado direito do sistema, levando a um vetor não-nulo. Assim é possível re-escrever a Equação 2.2 da seguinte forma:

$$A\mathrm{x} = b, \tag{2.3}$$

onde A é uma matriz $(n + nc) \times n$ dada por:

$$A = \begin{pmatrix} L \\ C \end{pmatrix}, \quad c_{ij} = \begin{cases} 1 & x_j \text{ é um ponto de controle} \\ 0 & \text{caso contrário} \end{cases}$$

e b é o vetor:

$$b_i = \begin{cases} 0 & i \leq n \\ x_{p_{c_i}} & n < i \leq n + nc \end{cases}$$

onde $x_{p_{c_i}}$ é uma das coordenadas cartesianas do ponto de controle p_{c_i}. A Figura 2.15 apresenta um exemplo de relação de vizinhança para seis pontos, onde os pontos em azul são os pontos de controle. A matriz A para este exemplo é apresentada na Figura 2.16.

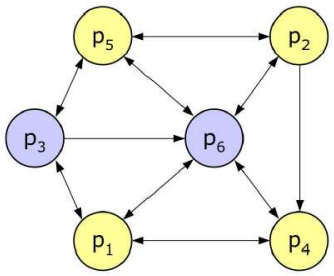


Figura 2.15: Relação de vizinhança para um conjunto de 6 pontos.
(Fonte: Paulovich et al. (2008))

2.5.3 Projection Explorer (PEx)

A utilização das técnicas de visualização juntamente com as técnicas de mineração de dados aumenta a capacidade de associação entre elementos similares. A ferramenta *Projection Explorer* (PEx) (Paulovich et al., 2007) foi desenvolvida com base nessa ideia e utiliza técnicas de projeção como solução para a mineração visual de coleções de dados multidimensionais. A Figura 2.17 mostra uma visão geral desta ferramenta.

$$A = \begin{bmatrix} 1 & 0 & -1/3 & -1/3 & 0 & -1/3 \\ 0 & 1 & 0 & -1/3 & -1/3 & -1/3 \\ -1/3 & 0 & 1 & 0 & -1/3 & -1/3 \\ -1/2 & 0 & 0 & 1 & 0 & -1/2 \\ 0 & -1/3 & -1/3 & 0 & 1 & -1/3 \\ -1/4 & -1/4 & 0 & -1/4 & -1/4 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \leftarrow L \\ \\ \leftarrow C \end{matrix}$$

Figura 2.16: Matriz A resultante (com pontos de controle p_3 e p_6).
(Fonte: Paulovich et al. (2008))

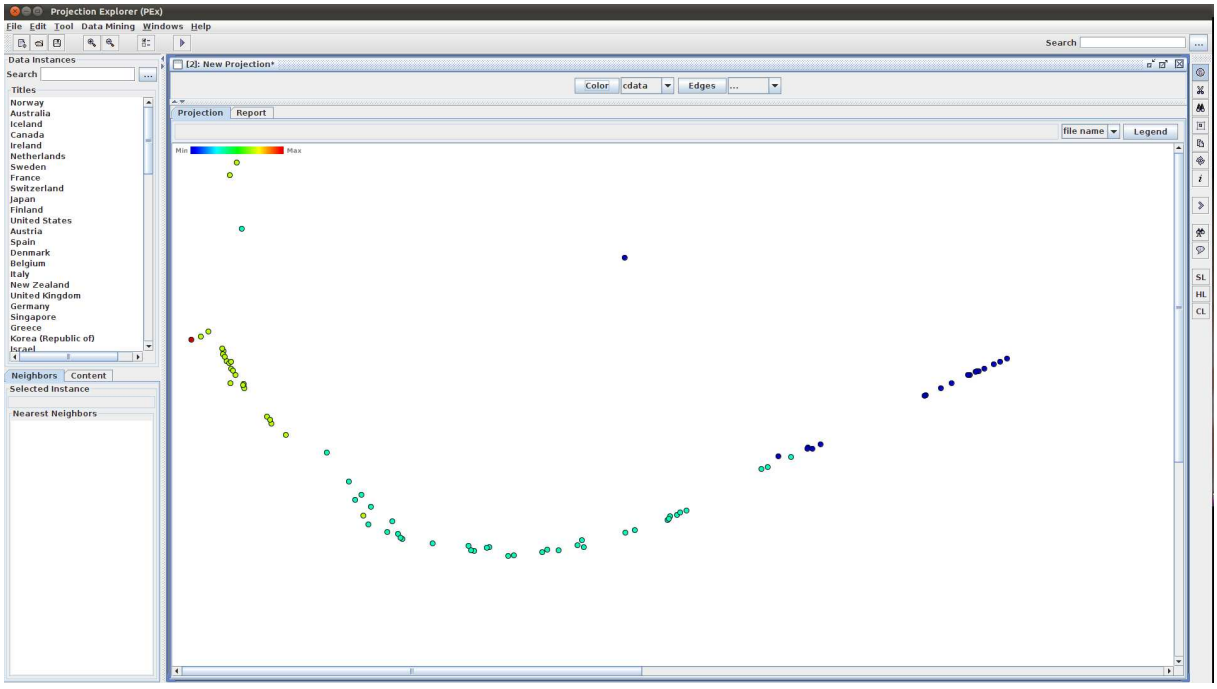


Figura 2.17: Visão geral da ferramenta *Projection Explorer*.

Para criar uma projeção a ferramenta segue um processo geral, ilustrado pela Figura 2.18. O conjunto de dados pode ser submetido a um pré-processamento e logo em seguida o modelo de representação vetorial é construído. O próximo passo é o cálculo das distâncias entre os elementos, estabelecendo assim as relações que serão utilizadas no processo de criação do mapa visual. Ao final, utiliza-se uma técnica de projeção e a representação visual é gerada. O usuário pode agora interagir e iniciar a exploração dos dados. A representação vetorial pode ser salva e utilizada posteriormente como entrada para a geração da representação visual, assim como a matriz de distâncias.

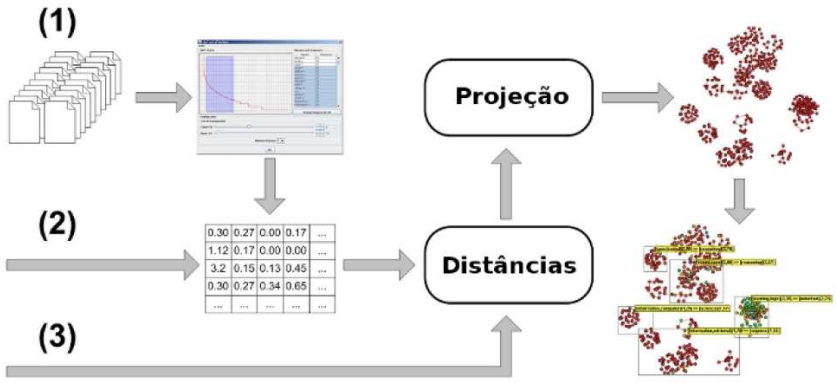


Figura 2.18: Processo de projeção de bases de dados multidimensionais da PEx. (Fonte: Paulovich et al. (2007))

2.6 Considerações finais

Neste capítulo foram apresentadas algumas das técnicas mais comuns para a visualização de dados multidimensionais. O grande desafio é representar dados multidimensionais em um espaço bidimensional, neste caso a tela do computador. Uma maior evidencia foi dada à técnica LSP, uma técnica de projeção multidimensional que é capaz de projetar grandes conjuntos de dados multidimensionais reconstruindo com precisão as relações de similaridade existentes entre os elementos no espaço original. Esta técnica serve de base para o desenvolvimento deste trabalho.

O próximo capítulo aborda a visualização de conjuntos de dados que possuem algum tipo de referência associada, como o tempo ou o par latitude e longitude, por exemplo.

Capítulo

3

Visualização de dados referenciados

3.1 Considerações iniciais

Grande parte dos dados mantidos em formato digital é coletada de forma automática, por sensores e sistemas de monitoramento (Keim et al., 2005). Transações feitas com cartões de crédito e ligações telefônicas são exemplos típicos.

Andrienko e Andrienko (2005) descrevem os dados como sendo resultados de observações ou medições de um fenômeno. Aos dados são atribuídas duas componentes: referências e atributos. Os dados podem conter várias referências, sendo as três mais encontradas: tempo, espaço e população. Por população entende-se um grupo de itens, independentemente de sua natureza. As referências são tomadas como variáveis independentes, enquanto que os atributos são as variáveis dependentes. Ou seja, referências podem ser escolhidas arbitrariamente, enquanto que os valores dos atributos são determinados pelas escolhas feitas. Por exemplo, em um conjunto de dados contendo informações populacionais observadas a cada cinco anos, as referências seriam o tempo e a localidade, e os atributos os dados observados. A seleção dos anos em que os dados são observados pode ser alterada pelo órgão responsável pelo censo, porém cada valor dos dados populacionais estará diretamente ligado a um ano específico e a uma área específica. O valor é completamente determinado pelas componentes temporal e espacial.

Deve-se perceber que variáveis como tempo e espaço nem sempre são tomadas como referências. Em eventos em que ocorrem migrações, como animais que buscam por água, o espaço passa a ser um atributo.

Técnicas de mineração são indispensáveis para analisar grandes conjuntos de dados. Abordagens interativas para mineração unindo aspectos positivos das técnicas automáticas e de técnicas de mineração visual permitem obter bons resultados na tarefa de exploração (Keim et al., 2005).

Segundo Andrienko e Andrienko (2005), apenas a mente humana realiza a tarefa de análise. As ferramentas promovem o suporte, preparando e apresentando o material necessário. A forma apropriada de apresentação é visual, pois a mente trabalha melhor com imagens.

Na Seção 3.2 deste capítulo são apresentadas algumas técnicas de visualização de dados onde a referência é o espaço, a geografia. Em seguida, na Seção 3.3 são descritas técnicas que apoiam a exploração visual de dados cuja referência associada é o tempo. Em ambas as seções podem aparecer exemplos em que os dados possuem os dois tipos de referência.

3.2 Visualização de dados geo-referenciados

Em alguns casos, uma referência geográfica é associada ao dado armazenado. Esta referência pode ser um par latitude e longitude, um código postal, um código de área (quando se trata de ligações telefônicas) ou um endereço, por exemplo. Modelagem do clima global, análise de hábitos de consumo, chamadas telefônicas e dados sobre criminalidade são exemplos de domínios de dados nos quais pode ser importante analisar o relacionamento envolvendo a localização geográfica. Ou seja, o objetivo da análise deste tipo de dados é associar o espaço à análise que se deseja fazer (Druck et al., 2004).

A análise de dados serve como suporte para suprir a necessidade de informação. Para encontrar a informação necessária, o analista planeja uma sequência de operações a ser aplicada aos dados e as executa utilizando as ferramentas disponíveis (Andrienko e Andrienko, 2005). A associação de ferramentas de análise com sistemas de informação geográfica pode favorecer a exploração efetiva do conjunto de dados.

Um conceito importante para a análise de fenômenos espaciais é o de dependência espacial. Por dependência espacial entende-se que eventos mais próximos geograficamente tendem a ser mais similares do que eventos mais distantes (Giannotti e Pedreschi, 2008). Este comportamento contrasta com a independência assumida nas técnicas clássicas de mineração de dados e prejudica o desempenho de algoritmos que ignoram a autocorrelação.

A seguir serão apresentadas alguma técnicas conhecidas para a visualização de dados com alguma referência espacial associada.

3.2.1 Coordenadas Paralelas

A técnica de coordenadas paralelas, apresentada na Seção 2.2.1, também pode ser utilizada como representação visual para conjuntos de dados geo-referenciados. Edsall (2003) apresenta alguns sistemas que utilizam esta técnica. A ideia básica é utilizar dois eixos antecedendo os eixos que correspondem aos atributos, um eixo mapeando a latitude e outro mapeando a longitude. Esta ideia está ilustrada na Figura 3.1.

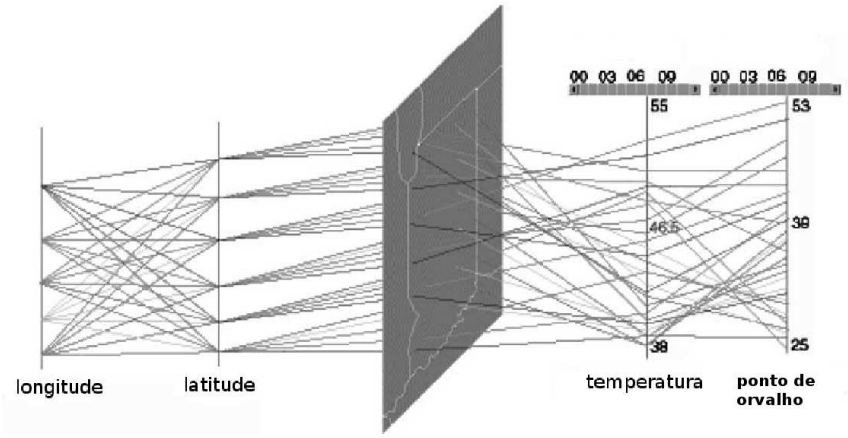


Figura 3.1: Visualização de um conjunto de séries temporais geo-referenciadas por meio da técnica de coordenadas paralelas (Edsall, 2003).

Como explicado na Seção 2.2.1, esta pode ser uma ferramenta poderosa para a análise exploratória de dados. Além dos problemas que são citados na seção, um problema que surge quando utiliza-se a técnica coordenadas paralelas para visualização de dados com referência geográfica, desta maneira, é a dificuldade de associar as coordenadas latitude e longitude aos pontos terrestres. Por exemplo, é difícil saber que a latitude 10° e a longitude 55° correspondem ao Brasil.

3.2.2 Representação cartográfica

O principal método de visualização de dados geo-referenciados é exibir sobre mapas representações gráficas como pontos, setas, polígonos e ícones, tomando a posição de cada objeto a partir de sua informação geográfica. A Figura 3.2 ilustra este método para um conjunto de dados demográficos de dois anos, 1981 e 1991, relacionado a distritos de Portugal.

Dependendo do domínio dos dados, um problema que pode surgir é a sobreposição das representações gráficas no mapa. A disposição destas representações tem sido estudada por diversos autores (Harrie et al., 2005; Edwardes et al., 2005). Entre as soluções pode-se citar: a utilização de legendas e a agregação de dados. Na agregação de dados, um elemento gráfico representa dois ou mais objetos do conjunto que são próximos geograficamente e similares no comportamento.

Uma forma de complementar este método de visualização é oferecer outras técnicas de visualização para os demais atributos por meio de múltiplas janelas de visão. Porém, segundo Maciejewski et al. (2008), um dos problemas mais observados quando se adota esta estratégia, é uma interação limitada entre as diferentes visões. Sendo este o principal método de visualização de dados dessa natureza, a seguir são apresentadas algumas ferramentas que utilizam esta forma de representação visual.

Figura 3.2: Visualização de representações gráficas sobre o mapa (Andrienko e Andrienko, 2005).

3.2.3 Aplicações de visualização de dados geo-referenciados

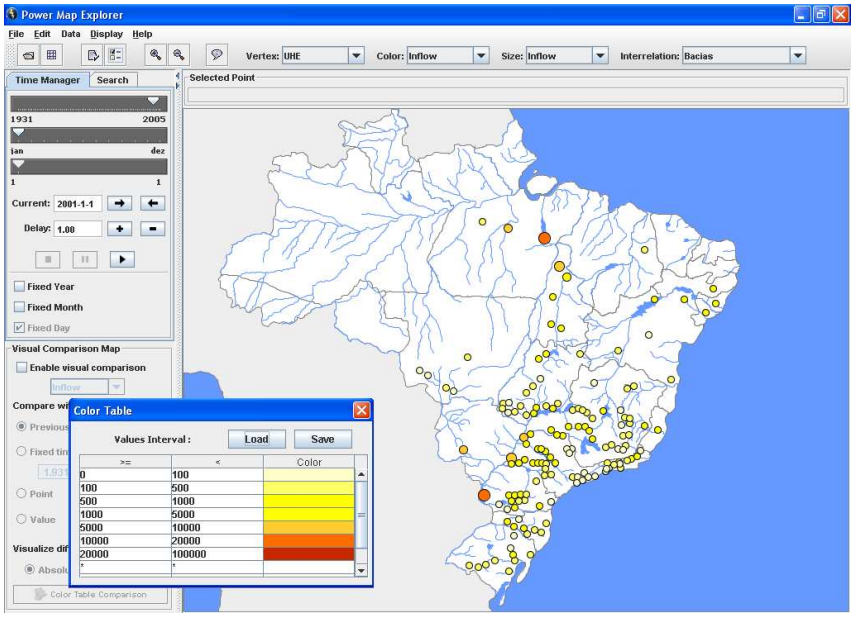
Existem diversas ferramentas que auxiliam a análise exploratória da dados geo-referenciados. A seguir algumas delas são descritas.

3.2.3.1 Power Map Explorer

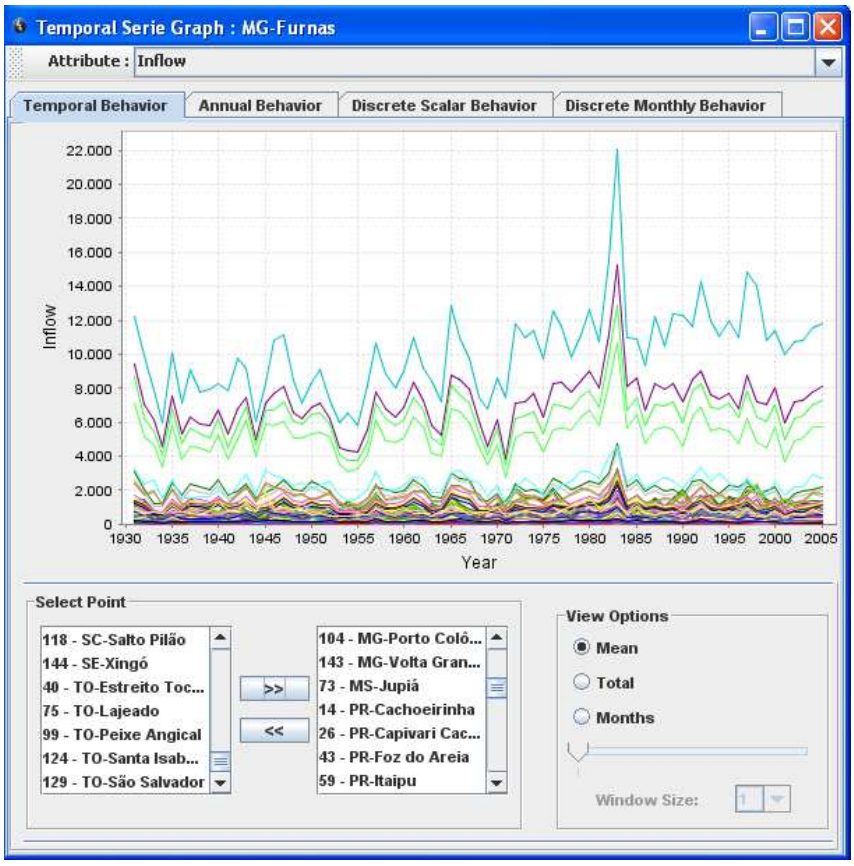
A ferramenta para visualização exploratória *Power Map Explorer* (Silva et al., 2007), tem como finalidade apoiar a análise de séries de vazões de usinas hidrelétricas. Nela as usinas hidrelétricas são representadas por círculos coloridos e posicionadas sobre o mapa geográfico brasileiro de acordo com os atributos espaciais de latitude e longitude, enquanto que a quantidade de vazão mensal é mapeada através dos atributos visuais cor e diâmetro dos círculos (ver Figura 3.3).

Na Figura 3.3(a) pode-se notar que a grande maioria de hidrelétricas do conjunto está concentrada nas regiões Sudeste e Sul do Brasil. Além disso, observa-se uma semelhança existente entre a maioria das usinas em relação à taxa mensal de vazão, que é representada pela cor amarela, para o período analisado. Porém, duas usinas apresentam alta taxa de vazão, representada pela cor vermelho, Itaipu e Tucuruí.

A ferramenta ainda apresenta outras funções que permitem manipulação interativa do tempo, fazendo com que os resultados sejam atualizados dinamicamente no mapa; geração de gráficos de tempo, *bar charts* e *pie charts*; e apresenta também outros componentes visuais como tabelas



(a) Distribuição espacial



(b) Gráfico de tempo

Figura 3.3: Ferramenta *Power Map Explorer*: em 3.3(a) está a representação das usinas hidrelétrica em um mapa e em 3.3(b) a vazão média das usinas (Silva et al., 2007).

de dados, variação uni-escala e projeção por similaridades para o tratamento de dados de vazões. Além de contar com um módulo para previsão das séries.

Esta ferramenta apresenta diversas funções para análise de séries temporais, porém a visualização da referência espacial resume-se ao território brasileiro, devido ao fato de ter sido desenvolvida para um caso específico.

3.2.3.2 LAHVA

A ferramenta LAHVA tem como foco a análise exploratória de conjuntos de dados multivariados, mais especificamente, dados sobre saúde (Maciejewski et al., 2008, 2009). Esta ferramenta combina a visualização de dados sobre o mapa com janelas de visão para análise estatística (ver Figura 3.4).

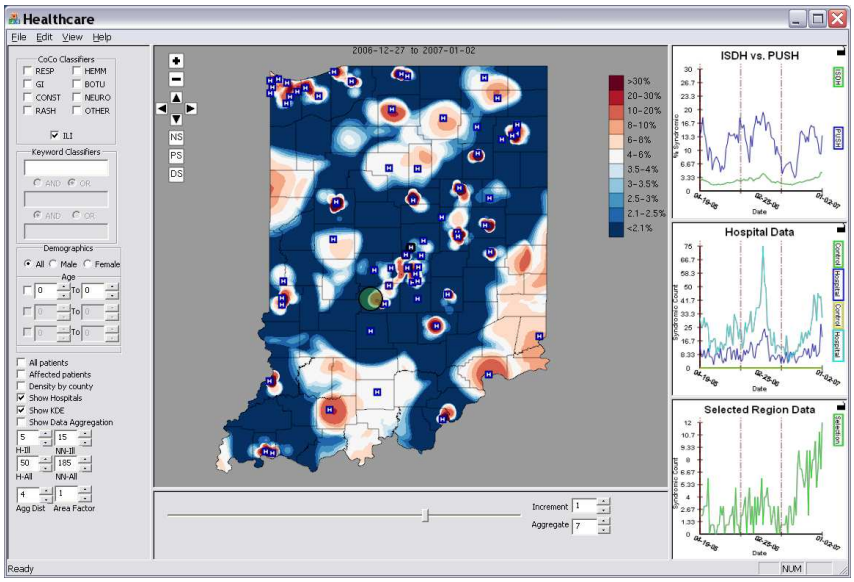
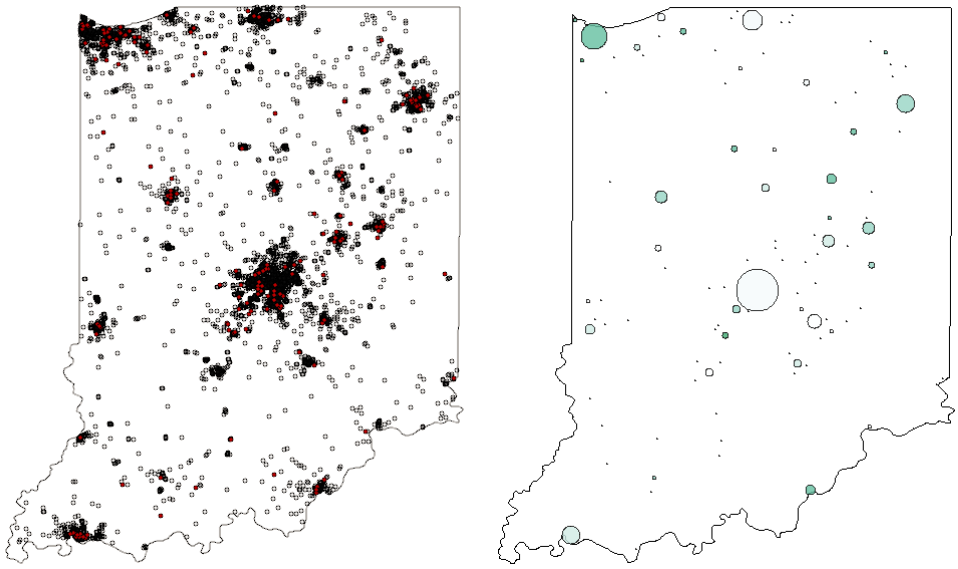


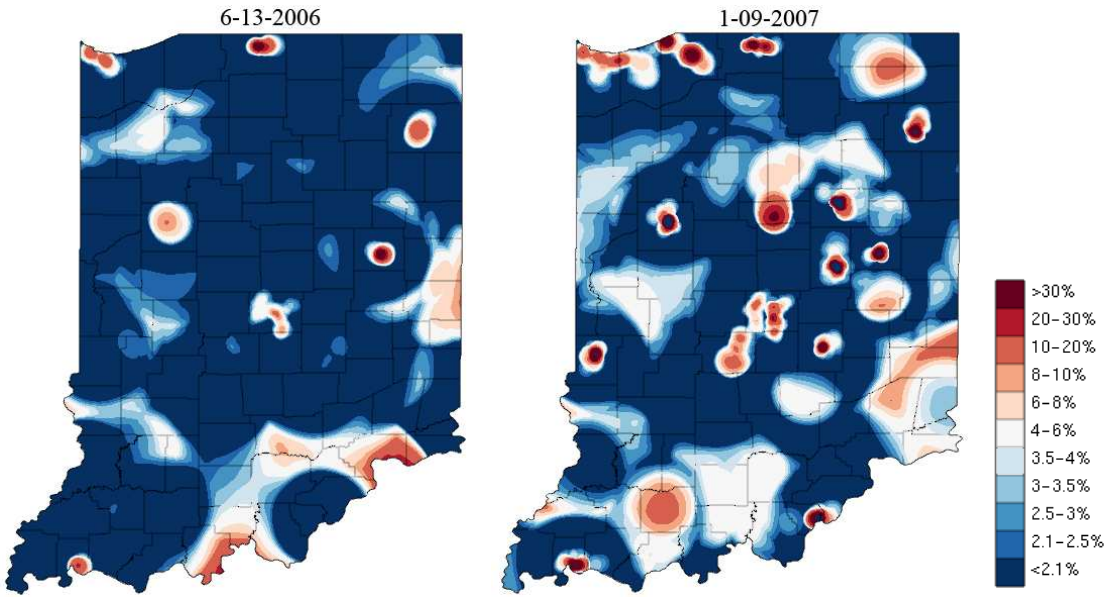
Figura 3.4: Tela principal da ferramenta LAHVA (Maciejewski et al., 2008).

Além de utilizar representações gráficas sobre o mapa e disponibilizar janelas de visão com modelos estatísticos, esta ferramenta oferece funcionalidades como a agregação de dados e mapa de calor geo-espacial. Na Figura 3.5(a) os dados geo-referenciados dos pacientes são exibidos como círculos sobre o mapa. A aplicação da técnica de agregação de dados, ilustrada na Figura 3.5(b), associa dados geograficamente próximos e com características similares. O mapa de calor geo-espacial, ilustrado na Figura 3.6(a), exibe a porcentagem de pacientes com gripe de acordo com a escala ilustrada na Figura 3.6(c). A evolução temporal da doença é ilustrada na Figura 3.6(b), onde é perceptível o aumento ou a diminuição da porcentagem de pacientes infectados em algumas regiões através da observação na mudança das cores.



(a) (b)

Figura 3.5: Representações gráficas da ferramenta LAHVA (Maciejewski et al., 2008).



(a) (b) (c)

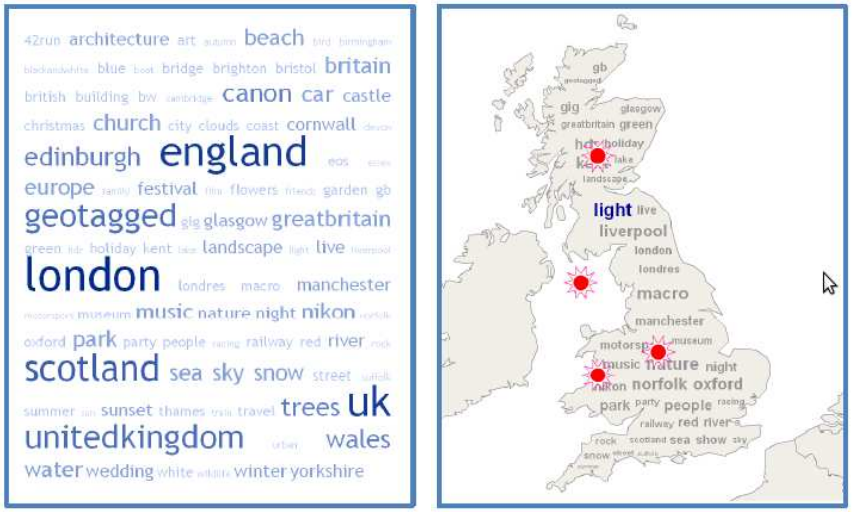
Figura 3.6: Representações gráficas da ferramenta LAHVA (Maciejewski et al., 2008).

3.2.3.3 Taggram

Taggram (Nguyen e Schumann, 2010) é uma ferramenta que combina a técnica *Tag Clouds* com a representação cartográfica, oferecendo suporte a exploração de dados geo-anotados. A ideia é posicionar e movimentar as *tags* sobre os mapas.

O objetivo é organizar e representar as *tags* associando-as com sua localização geográfica. Assim, a ferramenta torna possível a análise e observação de tendências ou padrões em bases de dados geo-referenciados. As *tags* podem ser posicionadas livremente, porém devem estar dentro da região geográfica associada a elas.

A Figura 3.7 ilustra um exemplo de uso da ferramenta. Na Figura 3.7(a) são apresentadas as *tags* mais populares para o termo "*United Kingdom*" utilizando a API do Flickr[1]. Em seguida, na Figura 3.7(b), o conjunto de *tags* é posicionado sobre o mapa do Reino Unido. Alguns locais são marcados quando a *tag light* é selecionada.



(a) (b)

Figura 3.7: Representações gráficas da ferramenta *Taggram* (Nguyen e Schumann, 2010).

3.3 Visualização de dados variantes no tempo

Diversas técnicas de visualização de dados com referência temporal associada podem ser encontradas na literatura. São dados que podem estar em diversas áreas como finanças, meteorologia, biologia, entre outras. O tempo é uma importante dimensão com características distintas em diversos domínios de aplicação. Laudos médicos, que descrevem o histórico de consultas de cada paciente, são um exemplo de conjunto de dados com referência temporal associada, neste caso uma coleção de documentos.

[1]http://www.flickr.com/services/api acessado em 20 de setembro de 2011

Séries temporais são um caso específico de dados multidimensionais variantes no tempo. Uma série temporal é um conjunto de observações X_t, onde t representa o instante de tempo em que a observação foi feita. A diferença em relação a outros dados é que cada instante de tempo é considerado como uma dimensão, enquanto que em outros conjuntos de dados os atributos são as dimensões. Uma das técnicas de visualização de séries mais simples que existe é o gráfico de linha, obtido unindo os pontos que representam os valores por segmentos de reta. Uma das características deste tipo de visualização é a facilidade de observar picos; porém, comportamentos cíclicos não são observados com tanta facilidade. Um exemplo é mostrado na Figura 3.8.

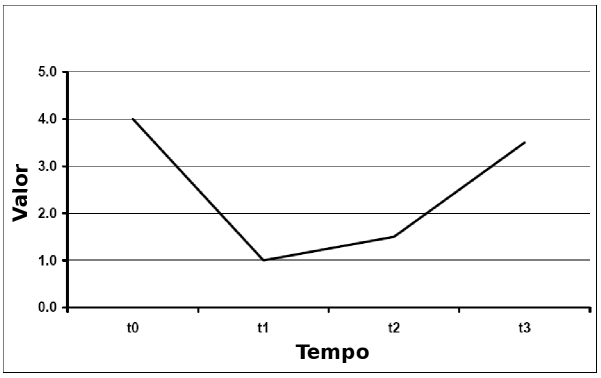


Figura 3.8: Exemplo de um gráfico de linha.

Além de não evidenciar comportamentos periódicos, outro problema em potencial quando se utiliza gráficos de linha para visualização de séries temporais é a oclusão visual. Dependendo da quantidade de séries pode haver uma sobreposição, inviabilizando a interpretação. Este é um problema comum em diversas técnicas de visualização.

Para solucionar dificuldades e proporcionar melhores maneiras de visualizar e analisar conjuntos de dados com esta característica, vários trabalhos foram ou estão sendo desenvolvidos. Estes trabalhos buscam apresentar solução para os seguintes problemas: multidimensionalidade, presença de múltiplos atributos, identificação de padrões periódicos, exploração e interatividade com a visualização gerada. Porém, muitos outros métodos são encontrados na literatura, os quais podem ser utilizados em diferentes etapas e tarefas do processo de análise exploratória.

3.3.1 Coordenadas Paralelas

Como foi descrito na Seção 2.2.1, a técnica de coordenadas paralelas é uma das mais conhecidas e favorece a identificação de correlação entre atributos.

Para a visualização de dados com referência temporal, cada eixo pode mapear um instante de tempo. Este tipo de visualização pode exigir uma grande quantidade de eixos, fato que dificulta a aplicação deste tipo de técnica para dados desta natureza.

A Figura 3.9 ilustra a técnica de coordenadas paralelas sendo aplicada a um conjunto de observações de temperatura em diferentes locais a cada três horas. Cada linha poligonal representa um local.

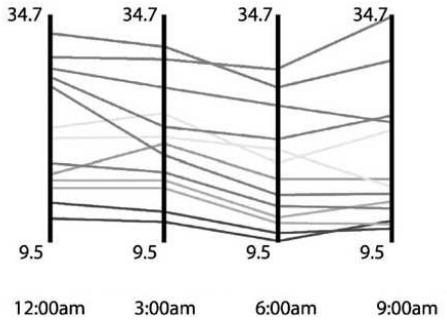


Figura 3.9: Técnica de coordenadas paralelas aplicada a um conjunto de séries temporais representando observações de temperatura em diferentes locais a cada três horas (Edsall, 2003).

A vantagem da utilização da técnica de coordenadas paralelas em relação a utilização do gráfico de linhas é que, como cada instante de tempo é considerado como uma dimensão e será mapeado em um eixo, é possível manipular os eixos de forma que se consiga identificar correlação entre instantes de tempo distantes. No gráfico de linha o tempo segue sua ordem natural, podendo ser difícil identificar uma correlação deste tipo.

3.3.2 Técnica Cluster and Calendar

Identificar tendências e padrões em múltiplas escalas de tempo, como dias, meses e anos é um problema que a técnica *Cluster and Calendar View* (Wijk e Selow, 1999) busca solucionar. Esta técnica combina o método de agrupamento, agrupando os dados diários similares, com uma metáfora de calendário para representar a evolução temporal.

A Figura 3.10 ilustra a utilização da técnica para representar um conjunto de séries temporais referentes ao número de funcionários trabalhando na ECN (*Netherlands Energy Research Foundation*). As observações foram feitas a cada hora, entre as 06:00 e 20:00 horas, em todos os dias do ano de 1997. Os agrupamentos representam um determinado dia ou conjuntos de dias; por exemplo, o grupo (*cluster*) 721 representa as datas relativas a um dia após um feriado, enquanto o grupo 5/12/1997 representa apenas o dia 5 de dezembro. Na parte à direita, as médias de valores de cada grupos são apresentadas em um gráfico de linhas, enquanto que na parte a esquerda cada dia recebe a cor de acordo com o grupo a que pertence. Observando o grupo 721 percebe-se que algumas pessoas tiram folga após um feriado, por exemplo.

Apesar de utilizar uma boa forma de representar os dados, pois a metáfora de calendário permite identificar de forma fácil se a similaridade do grupo é por conta da estação climática, do dia da semana ou qualquer outra correlação deste tipo, este método é limitado a dados que estejam em uma escala compatível com o modelo do calendário, apresentando valores diários por exemplo. Além disso, para determinar o número de *clusters* a serem formados, um conhecimento prévio dos padrões presentes nos dados também é necessário.

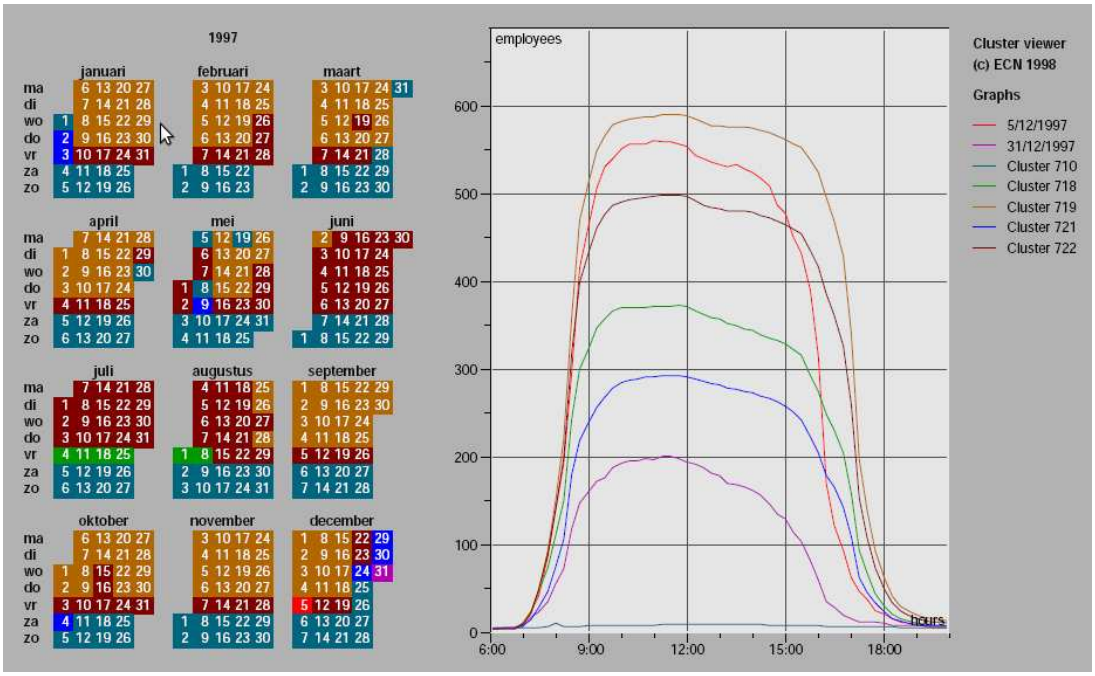


Figura 3.10: Exemplo da técnica *Cluster and Calendar View* (Wijk e Selow, 1999).

3.3.3 Técnicas orientadas a pixels

As técnicas orientadas a *pixels*, descritas na Seção 2.4, também podem ser utilizadas para gerar representações visuais de dados variantes no tempo.

Lammarsch et al. (2009) apresenta o método chamado *Granulary Overview Overlay* (GROOVE), que utiliza a representação baseada em *pixels* para a exploração de dados com referência temporal. A ferramenta utiliza as linhas e colunas da representação para mapear as estruturas de tempo.

A Figura 3.11 ilustra a utilização da técnica GROOVE para um mês de observação de atividades policiais. Cada bloco representa um dia e as linhas representam as horas. As semanas correspondem as linhas de blocos.

Uma dificuldade encontrada na utilização desta técnica é a observação do valor exato que é mapeado por uma determinada cor, o que é fácil observar quando se utiliza gráficos de linha, por exemplo.

3.3.4 Projeções multidimensionais

Projeções multidimensionais, aplicadas a conjuntos de séries temporais, partem do pressuposto que uma série temporal pode ser representada como sendo um ponto definido no espaço m-dimensional.

A Figura 3.12 mostra a representação gráfica resultante de uma projeção multidimensional de um conjunto de séries temporais que descrevem vazões de 76 usinas hidrelétricas da bacia do Paraná, registradas ou estimadas ao longo de vários anos.

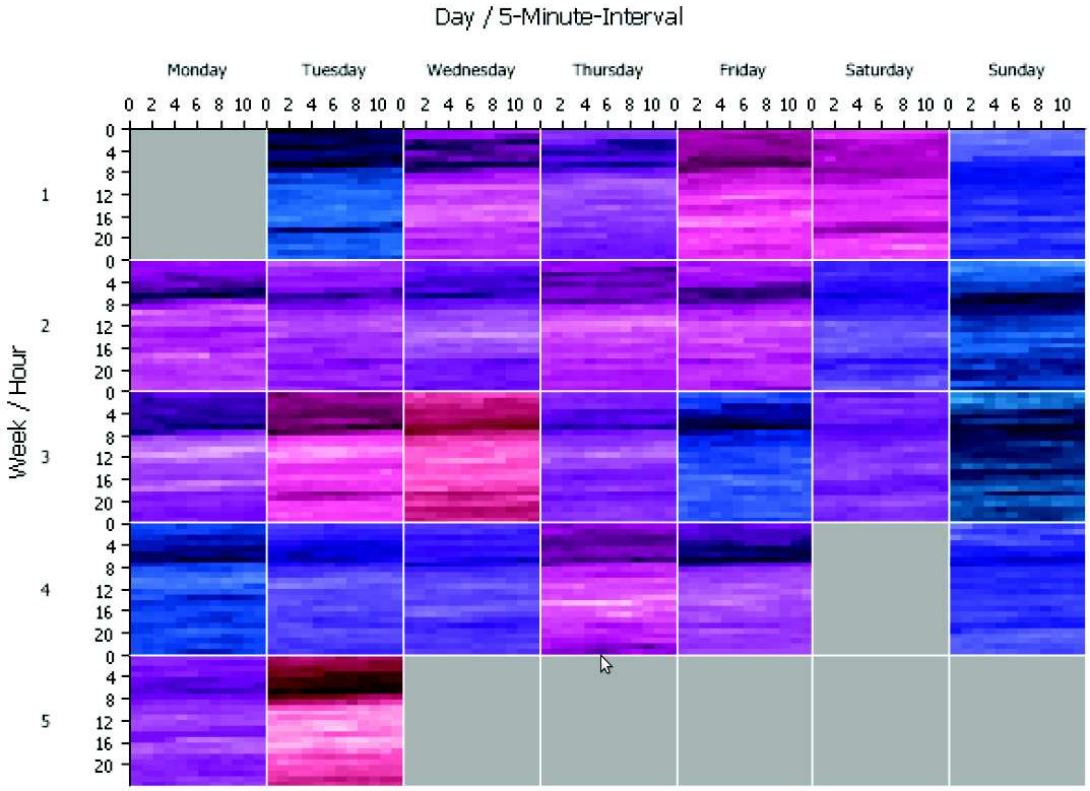


Figura 3.11: Exemplo da técnica GROOVE (Lammarsch et al., 2009).

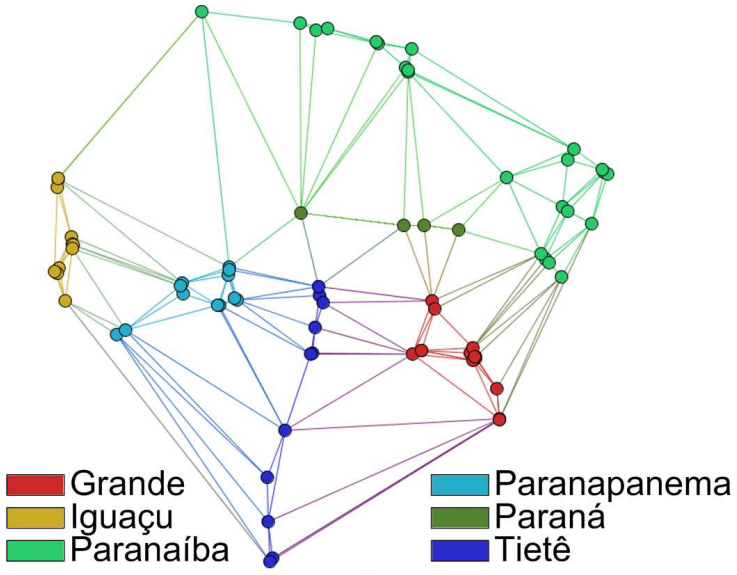


Figura 3.12: Exemplo de representação gráfica resultante da técnica de projeção multidimensional (Alencar et al., 2007).

A utilização das técnicas de visualização juntamente com as técnicas de mineração de dados aumenta a capacidade de associação entre séries similares. A ferramenta *Temporal-PEx* (Alencar et al., 2007) foi desenvolvida com base nessa ideia e utiliza técnicas de projeção como solução para a mineração visual de coleções de dados multidimensionais. Esta ferramenta permite que os usuários tenham uma visão geral do comportamento da coleção de séries temporais e facilita a observação de grupos de séries similares (ver a Figura 3.13).

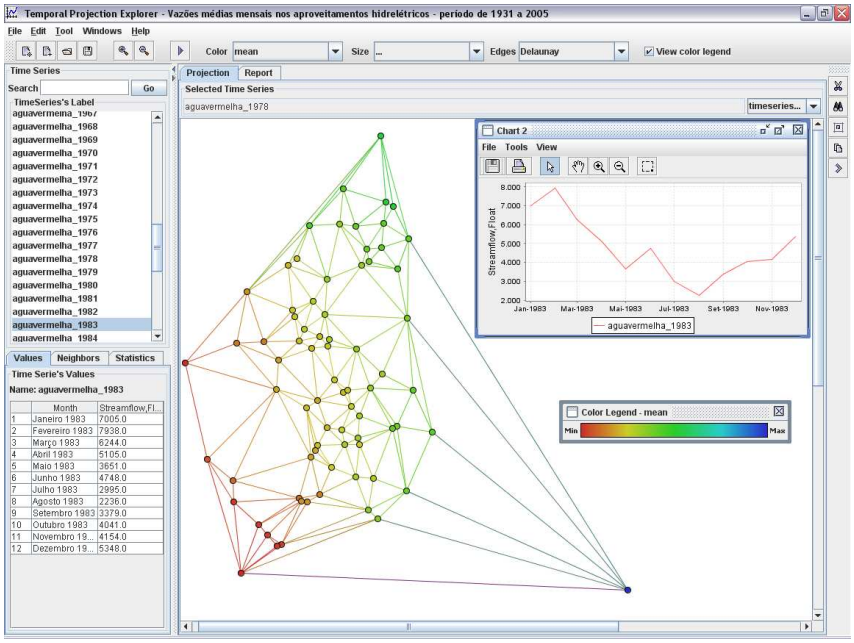


Figura 3.13: Tela da ferramenta *Temporal-PEx* (Alencar et al., 2007).

A ferramenta *Temporal-PEx* mostrou-se bastante útil e eficiente no suporte a tarefa de análise exploratória de conjuntos de séries temporais.

3.3.5 Animação apoiando a visualização

A forma mais comum para a observação de tendências nos dados é exibição das mudanças das variáveis ao longo do tempo, utilizando gráfico de dispersão (ver Seção 2.1) por exemplo. Animação é uma sequência de imagens, que mudam ao longo do tempo, criando a ilusão de movimento.

Cleveland e McGill (1985) mostra que mudanças de posicionamento são mais perceptíveis do que mudanças na área. Ou seja, durante a animação, o deslocamento de pontos será mais eficaz para o entendimento dos dados do que o aumento ou diminuição do diâmetro, por exemplo.

Segundo Robertson et al. (2008), animações tem sido utilizadas para exibir tendências em dados multidimensionais. Um exemplo bastante interessante é a ferramenta *Gapminder Trendalyzer*[2] que utiliza animação e uma apresentação atraente para ajudar os usuários a entender os resultados da análise dos dados. Esta ferramenta foi desenvolvida para exibir tendências ao

[2]http://www.gapminder.org acessado em 20 de setembro de 2011

longo do tempo utilizando animação, mapeando os valores dos atributos em eixos cartesianos, no diâmetro dos círculos e nas cores.

A Figura 3.14 ilustra a ferramenta *Gapminder Trendalyzer* mostrando quanto tempo as pessoas vivem (expectativa de vida em anos mapeada no eixo vertical) e quanto elas ganham (produto interno bruto per capita mapeado no eixo horizontal) em diferentes países. O diâmetro das circunferências, que representam os países, é utilizado para mapear a população de cada país e as cores representam as regiões geográficas. A evolução temporal pode ser visualizada utilizando a barra deslizante, localizada abaixo do eixo horizontal, ou a animação, acionada pelo botão *play*.

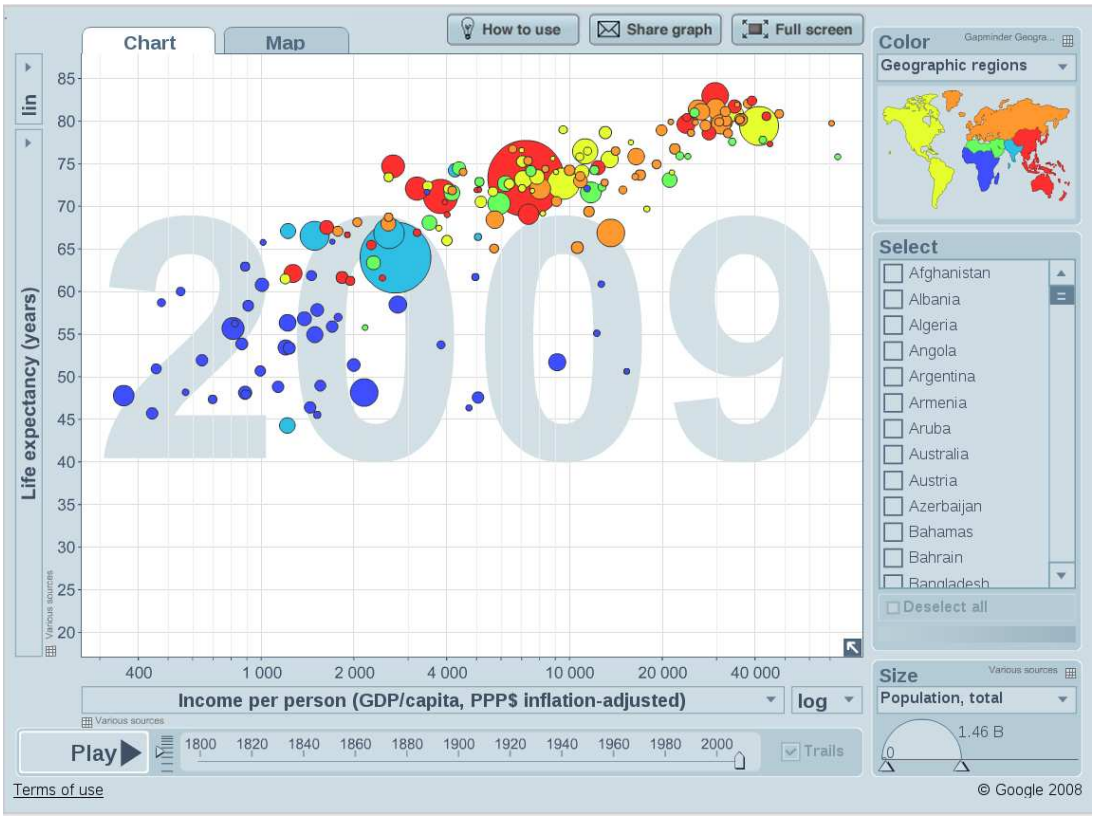


Figura 3.14: Ferramenta *Gapminder Trendalyzer*.

Outra forma de observar a evolução temporal é selecionando um país e analisar o rastro deixado pela animação ao longo do tempo. A Figura 3.15 exibe a evolução temporal do Brasil para os mesmos dados observados na Figura 3.14. É fácil observar que a expectativa de vida e a renda dos brasileiros cresceram ao longo dos anos.

Um aspecto interessante da aplicação *Gapminder Trendalyzer* é a possibilidade do usuário manipular os indicadores que são mapeados nos eixos e observar o gráfico mudando ao longo do tempo, conforme os anos de observação vão passando.

Robertson et al. (2008) alertam que a utilização de animação pode ser pouco eficaz para a análise, se comparada com representações estáticas, devido ao número de repetições que podem

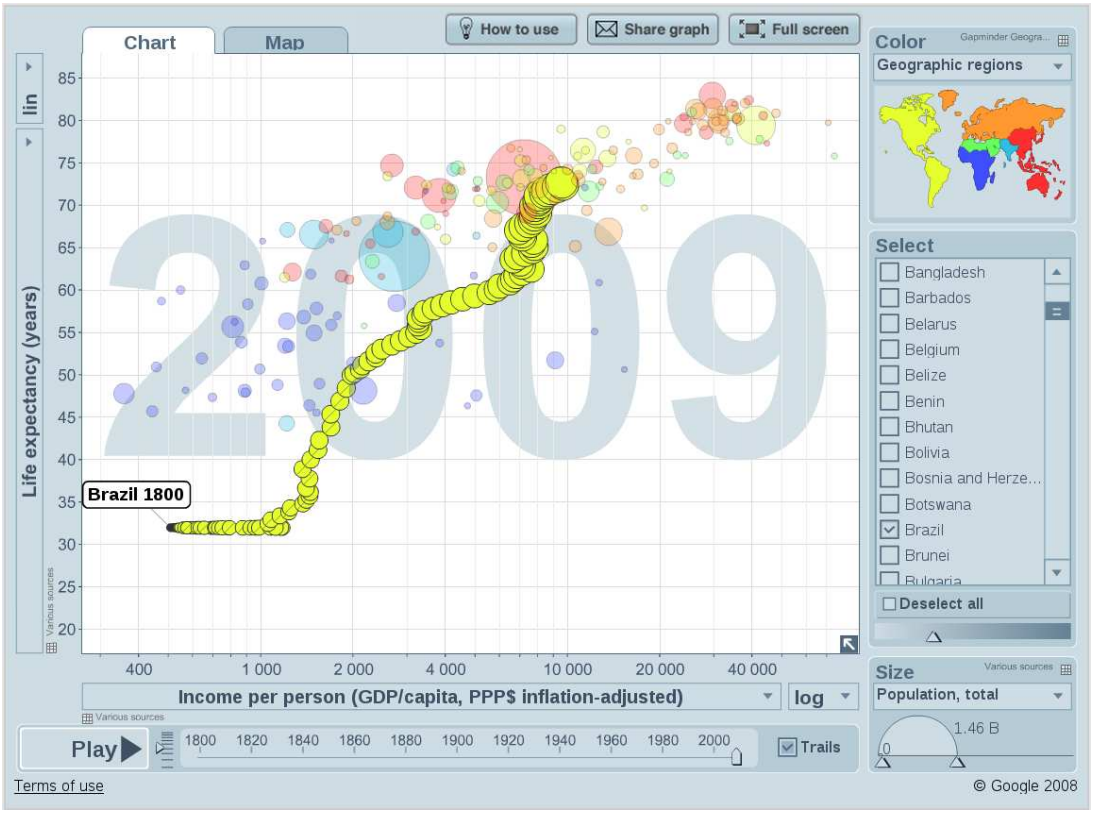


Figura 3.15: Evolução temporal do Brasil utilizando a ferramenta *Gapminder Trendalyzer*.

vir a ser necessárias para o entendimento dos dados e observação de tendências ou padrões. Os autores também citam alguns problemas das técnicas estáticas, como oclusão visual por exemplo, que podem ser minimizados com o auxílio da animação. Dessa forma, este tipo de técnica pode ser uma ferramenta valiosa para tarefas de análise exploratória dos dados, quando utilizada de forma conjunta com outras técnicas.

3.4 Considerações finais

Neste capítulo foram apresentadas algumas técnicas de visualização de dados geo-referenciados e dados variantes no tempo. A representação cartográfica ainda é a forma mais utilizada para visualizar a referência geográfica e, para visualização da referência temporal, o principal método é a utilização de um eixo cartesiano para representar o tempo. Como complemento, algumas ferramentas utilizadas para realizar estas tarefas também foram apresentadas.

A maioria das técnicas encontradas são muito específicas para um tipo de referência: ou tratam a referência temporal ou tratam a referência espacial. Quando buscam tratar as duas referências em conjunto, apresentam limitação com relação à quantidade de dimensões que podem ser observadas.

O próximo capítulo descreve a técnica implementada para auxiliar um usuário na análise exploratória de dados multidimensionais referenciados utilizando projeção multidimensional.

Capítulo

4

Least Square Projection com pontos de controle fixos

4.1 Considerações iniciais

Como visto na Seção 2.5, técnicas de projeção multidimensional permitem gerar representações visuais de grandes coleções de dados. Nestas representações, os elementos do espaço multidimensional são projetados em um espaço bidimensional, sendo exibidos como pontos.

Diversas técnicas de projeção multidimensional foram investigadas pelo grupo de pesquisa em Visualização, Imagens e Computação Gráfica do ICMC. Entre elas está a técnica *Least Square Projection* (LSP) (Paulovich et al., 2008), descrita na Seção 2.5.2. A LSP busca preservar relações de vizinhança, definidas no espaço multidimensional, entre os objetos no espaço projetado. Esta abordagem favorece a identificação de agrupamentos inerentes ao conjunto de dados.

No entanto, a técnica LSP em sua formulação original não favorece a análise de aspectos temporais ou geográficos presentes nos dados, de modo que o usuário não consegue associar o espaço geográfico ou perceber comportamentos que acontecem ao longo do tempo de forma intuitiva.

A técnica LSP utiliza o posicionamento dos chamados pontos de controle para inferir posição geométrica dos demais pontos. Nossa proposta é utilizar os pontos de controle para controlar alguns elementos da projeção de modo a tratar essas limitações. No caso da referência temporal, a ideia é criar uma projeção para cada instante de tempo e utilizar os mesmos pontos de controle em cada uma delas, preservando a coerência temporal. Dessa forma, busca-se uma correlação

entre as projeções. É possível também definir a posição dos pontos de controle com as suas coordenadas geográficas, e assim incorporar esta referência à análise, além de buscar uma melhor distribuição de todos os pontos no espaço. A análise da informação geográfica pode favorecer a observação de tendências regionais ao longo do tempo, por exemplo, países de continentes diferentes que tendem a se comportar de forma similar.

A utilização conjunta das duas abordagens permite a análise de conjuntos de dados geo-referenciados variantes no tempo. De forma a lidar com dados que apresentam estas características estendeu-se a técnica LSP. A nova abordagem é chamada de *Least Square Projection with Fixed Control Points*, ou simplesmente LSP_{FCP}, e busca auxiliar a observação do comportamento global de conjuntos de dados que mudam ao longo do tempo e associar a informação geográfica ao processo de exploração dos dados por meio dos pontos de controle, quando esta referência existir.

A maioria das figuras utilizadas neste capítulo para ilustrar as funcionalidades implementadas foi obtida utilizando a ferramenta PEx com uma base de dados composta pela observação de quatro atributos entre os anos de 1980 e 2005, expectativa de vida, alfabetização de adultos, taxa de escolarização bruta e o produto interno bruto per capita, os quais são utilizados no cálculo do Índice de Desenvolvimento Humano (IDH)[1]. Nas visualizações as cores mapeiam a classe de cada país, segundo o critério do relatório do Programa das Nações Unidas para o Desenvolvimento: países com índice de desenvolvimento baixo (azul), médio (ciano), alto (verde) e muito alto (vermelho).

Os primeiros tipos de dados manipulados no início do trabalho foram conjuntos de séries temporais. Dessa maneira optou-se por estender a ferramenta *Temporal-PEx*, descrita na Seção 3.3.4. Ao longo do desenvolvimento, percebeu-se que esta abordagem poderia ser utilizada para outros tipos de dados e decidiu-se estender a ferramenta PEx, na versão 1.6.3, descrita na Seção 2.5.3.

A seguir os detalhes da LSP_{FCP} são apresentados. Em seguida, na Seção 4.2.3, o processo de criação das projeções é descrito para a ferramenta PEx e, na Seção 4.3, é descrito para a ferramenta *Temporal-PEx*.

4.2 Least Square Projection com pontos de controle fixos

O novo processo de projeção é executado em três etapas, como mostra a Figura 4.1. A primeira etapa é a obtenção das coordenadas dos pontos de controle. Em seguida, utiliza-se estas coordenadas para gerar uma projeção para cada *timestamp* do conjunto de dados. Ou seja, todas as projeções terão os mesmos pontos de controle com as mesmas coordenadas. Ao final, as projeções criadas na segunda etapa são reunidas em uma única projeção, e pode-se definir algumas funcionalidades para ajudar na observação do comportamento dos elementos do conjunto, ao longo do tempo.

[1]Dados encontrados em http://hdr.undp.org/en/statistics/data/ acessado em 20 de setembro de 2011

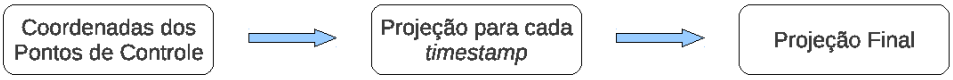


Figura 4.1: Etapas da técnica LSP_{FCP}.

Os pontos de controle podem ser informados pelo usuário ou ser obtidos pelos mecanismos implementados na PEx, já apresentados na Seção 2.5.2. No caso da escolha dos pontos de controle ser feita pelo procedimento convencional, as coordenadas de tela serão obtidas projetando-se os mesmos. A técnica de projeção adotada é a IDMAP que combina a técnica rápida *Fastmap* (Faloutsos e Lin, 1995) com a técnica *Force Scheme* (Tejada et al., 2003), como descrito na Seção 2.5.1.

Alternativamente, os pontos de controle podem ser informados pelo usuário. Este pode, por exemplo, utilizar as referências geográficas como coordenadas cartesianas dos pontos de controle para resolver o sistema linear que resulta nas coordenadas dos pontos projetados. Geralmente, utiliza-se o par latitude e longitude como referência geográfica, o qual indica com precisão a posição de qualquer ponto sobre a superfície terrestre. Primeiramente estes valores são normalizados. Os valores para latitude vão de -90º à 90º, enquanto que os valores para longitude vão de -180º à 180º (ver Figura 4.2). Assim, a normalização é feita utilizando as seguintes equações:

$$x = \frac{longitude + 180}{360} \tag{4.1}$$

$$y = \frac{latitude + 90}{180} \tag{4.2}$$

Figura 4.2: Mapa cartesiano ilustrando as coordenadas geográficas.
(Fonte: http://geoconceicao.blogspot.com/2010/03/coordenadas-geograficas.html acessado em 20 de setembro de 2011)

De posse destas coordenadas, o sistema linear é construído de maneira análoga à técnica LSP original (ver Seção 2.5.2), correspondendo à segunda etapa da nova estratégia. A resolução deste sistema dá as coordenadas dos pontos projetados.

Considerando os conjuntos representando as coordenadas cartesianas dos pontos de controle $X = x_{pc_1}, x_{pc_2}, \ldots, x_{pc_w}$ e $Y = y_{pc_1}, y_{pc_2}, \ldots, y_{pc_w}$, com w sendo o número de pontos de controle escolhidos. Para cada *timestamp* dois sistemas lineares são resolvidos, um para o conjunto X e outro para o conjunto Y, com o resultado sendo as coordenadas no espaço projetado de cada elemento do conjunto de pontos.

Cada sistema tem como base a Equação 2.3, escrita na Seção 2.5.2 da seguinte forma:

$$A\mathbf{x} = b, \tag{4.3}$$

onde A é uma matriz $(n + nc) \times n$ dada por:

$$A = \begin{pmatrix} L \\ C \end{pmatrix}, \quad c_{ij} = \begin{cases} 1 & x_j \text{ é um ponto de controle} \\ 0 & \text{caso contrário} \end{cases}$$

e b é o vetor:

$$b_i = \begin{cases} 0 & i \leq n \\ x_{p_{c_i}} & n < i \leq n + nc \end{cases}$$

onde $x_{p_{c_i}}$ é uma das coordenadas cartesianas do ponto de controle p_{c_i}. Os valores para $x_{p_{c_i}}$ em um sistema serão dados pelo conjunto X e no outro pelo conjunto Y. A matriz L foi descrita na Seção 2.5.2 e seus valores são dados por:

$$l_{ij} = \begin{cases} 1, & i = j \\ -\alpha_{ij}, & p_j \in V_i \\ 0 & \text{caso contrário} \end{cases}$$

com $\alpha_{ij} = \frac{1}{k_i}$ e k_i sendo a quantidade de vizinhos do ponto p_i. O vetor $\mathbf{x}$ contém as coordenadas cartesianas dos pontos, as quais serão utilizadas para o posicionamento na projeção, ou seja, existe um vetor $\mathbf{x}$ para cada sistema. Os valores dos vetores $\mathbf{x}$ são então mapeados para as coordenadas de tela e a projeção, correspondente a cada *timestamp*, é exibida.

Em seguida, na terceira etapa, as projeções geradas (ou um sub-conjunto selecionado) são justapostas em uma única projeção (ver Figura 4.3) e a transição entre elas é evidenciada utilizando as funcionalidades descritas adiante, na Seção 4.2.2.

Como forma de ilustrar as diferentes abordagens, foram geradas diversas sequências temporais de projeções dos dados que compõem o IDH. Em todas utilizou-se oito pontos de controle, variando a forma como estes são escolhidos. A Tabela 4.1 exibe as diferenças entre cada sequência de projeções. O seguinte critério de escolha foi adotado pelo usuário: dois pontos de cada grupo de classificação, segundo os critérios da ONU, que não mudaram de classe ao longo do tempo ou mudaram apenas uma vez.

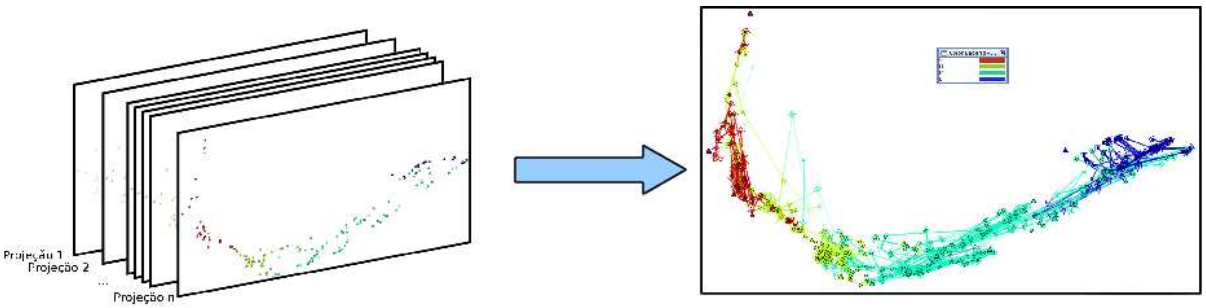


Figura 4.3: Ilustração da justaposição da sequência temporal de projeções.

Tabela 4.1: Escolha dos pontos de controle para gerar as sequências temporais.

	Técnica de Projeção	**Escolha dos pontos de controle**
Sequência 1	LSP	k-*means*
Sequência 2	LSP_{FCP}	k-*means* para o ano de 1980 e coordenadas projetadas obtidas pela técnica IDMAP
Sequência 3	LSP_{FCP}	k-*means* para o ano de 1980 e coordenadas geográficas como coordenadas projetadas
Sequência 4	LSP_{FCP}	Pontos de controle escolhidos pelo usuário e coordenadas projetadas obtidas pela técnica IDMAP
Sequência 5	LSP_{FCP}	Pontos de controle escolhidos pelo usuário e coordenadas geográficas como coordenadas projetadas

A Figura 4.4 exibe as projeções do *timestamp* correspondente ao ano de 2005 para cada sequência encontrada na Tabela 4.1. As projeções ilustradas nas Figuras 4.4(a) e 4.4(b) apresentam a mesma estrutura global. A diferença está no fato de que na Figura 4.4(a) os países que apresentam melhor IDH estão a direita enquanto os que apresentam pior IDH ficam a esquerda; na Figura 4.4(b) acontece o oposto. As demais projeções apresentam estruturas diferentes, porém, nas projeções que utilizam coordenadas geográficas como coordenadas cartesianas os países que apresentam melhor IDH tendem a ficar na parte superior enquanto que os países que apresentam pior IDH estão na parte inferior.

Os pontos de controle, selecionados para os dados de 1980, não são necessariamente os pontos mais representativos nos outros anos. Esse fato pode ser um problema na abordagem da técnica LSP_{FCP}, e também no caso da escolha dos pontos de controle pelo usuário, pois a precisão de mapeamento tende a diminuir.

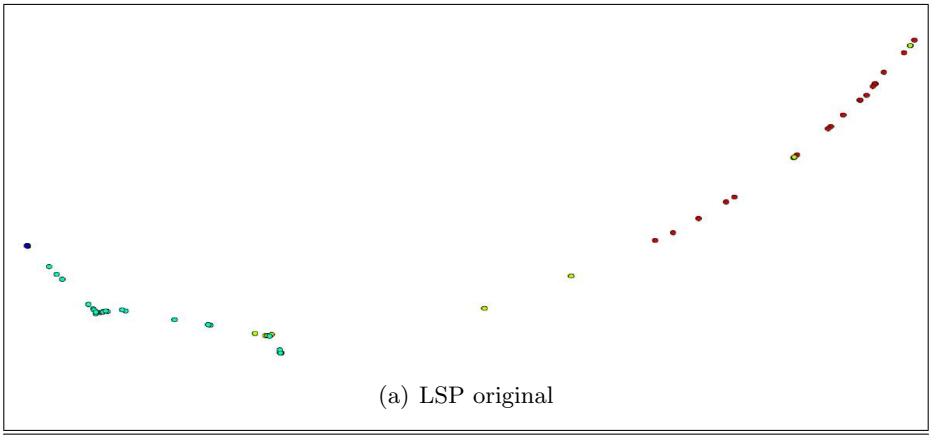
A Figura 4.5 apresenta o gráfico *Neighborhood Hit*, que representa a proporção de vizinhos de um ponto p, no espaço projetado, que pertencem à mesma classe de p, e na Figura 4.6 é apresentado o gráfico *Neighborhood Preservation*, que representa a proporção de vizinhos do espaço original que são mantidos como vizinhos no espaço projetado. Ambos os gráficos comparam as projeções das sequências apresentadas na Tabela 4.1 e ilustradas pela Figura 4.4.

Como esperado, no geral há perda de precisão e preservação de vizinhos quando se utiliza os pontos de controle escolhidos no ano de 1980 para gerar a projeção dos dados de 2005, assim como há perda de precisão e preservação quando os pontos de controle são escolhidos pelo usuário. Porém, utilizando-se as coordenadas dadas pela técnica IDMAP, curvas em azul e rosa, a perda de precisão é pequena se comparada com a técnica LSP em sua formulação original, curva em ciano. Isto para este conjunto de dados em especial.

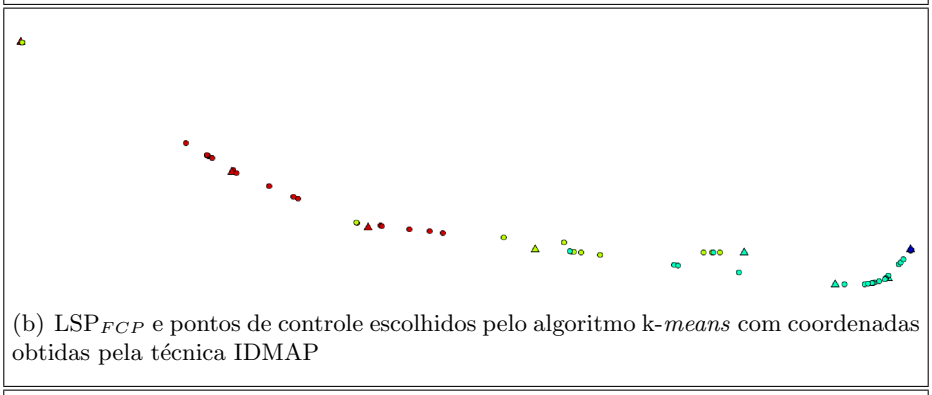
4.2.1 Animação dos pontos

Na sequência temporal a movimentação dos pontos pode indicar ao especialista possíveis mudanças no comportamento dos dados, pois reflete as mudanças de similaridade entre os elementos que ocorrem entre sucessivos *timestamps*. A fixação dos pontos de controle, garantindo a coerência temporal, evita problemas como o ilustrado pela Figura 4.7 que exibe projeções geradas utilizando a LSP em sua abordagem original. Nas projeções dos dados do IDH, utilizando a LSP original, observa-se uma inversão no posicionamento dos pontos em sua estrutura global, o que não acontece nas projeções utilizando a LSP_{FCP} (ver Figura 4.8).

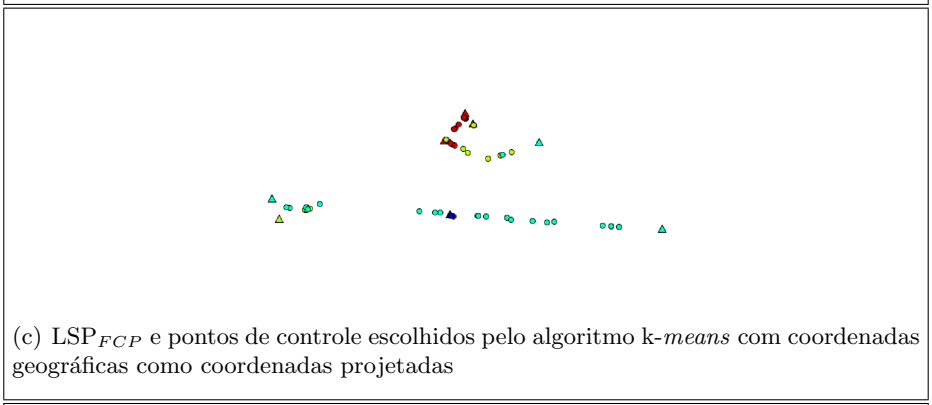
O usuário pode escolher entre três opções para a animação dos pontos: pontos se movimentando, setas crescendo ou pontos deixando rastro com transparência. A ilustração da animação, com os pontos se movendo, pode ser observada na Figura 4.9. A Figura 4.9(a) exibe o início da animação, correspondente à primeira projeção. Na Figura 4.9(b) um instante de tempo intermediário é exibido e, na Figura 4.9(c), a animação chega ao final. Durante a evolução da animação o ponteiro da barra de deslizamento se movimenta, indicando o instante de tempo corrente. Os valores exibidos na barra de deslizamento identificam as projeções. A transparência



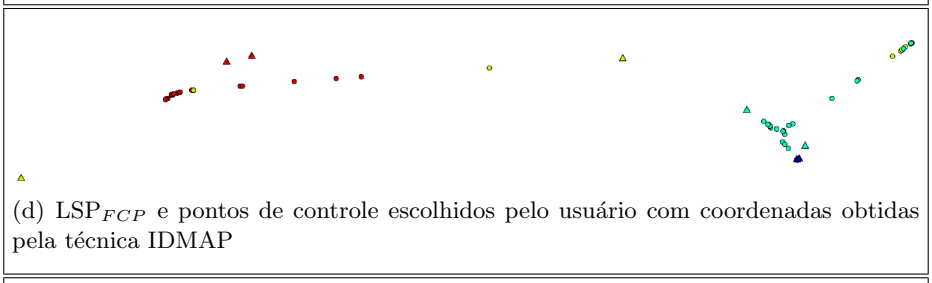
(a) LSP original



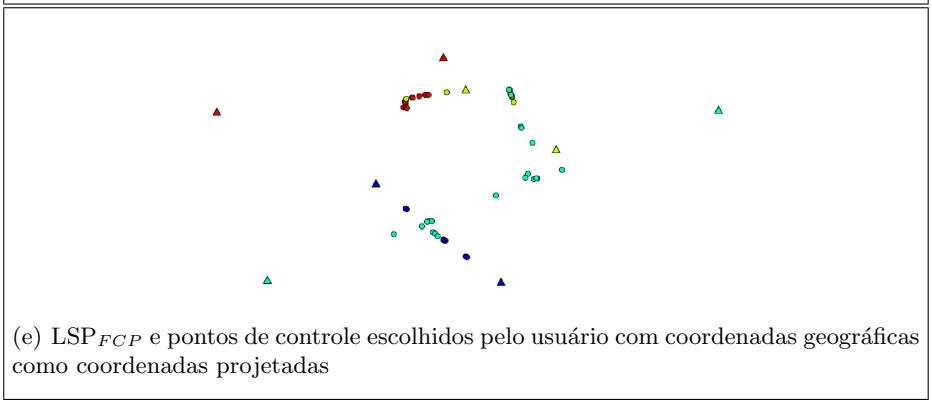
(b) LSP$_{FCP}$ e pontos de controle escolhidos pelo algoritmo k-*means* com coordenadas obtidas pela técnica IDMAP



(c) LSP$_{FCP}$ e pontos de controle escolhidos pelo algoritmo k-*means* com coordenadas geográficas como coordenadas projetadas



(d) LSP$_{FCP}$ e pontos de controle escolhidos pelo usuário com coordenadas obtidas pela técnica IDMAP



(e) LSP$_{FCP}$ e pontos de controle escolhidos pelo usuário com coordenadas geográficas como coordenadas projetadas

Figura 4.4: Diferentes abordagens na geração de sequências temporais de projeções.

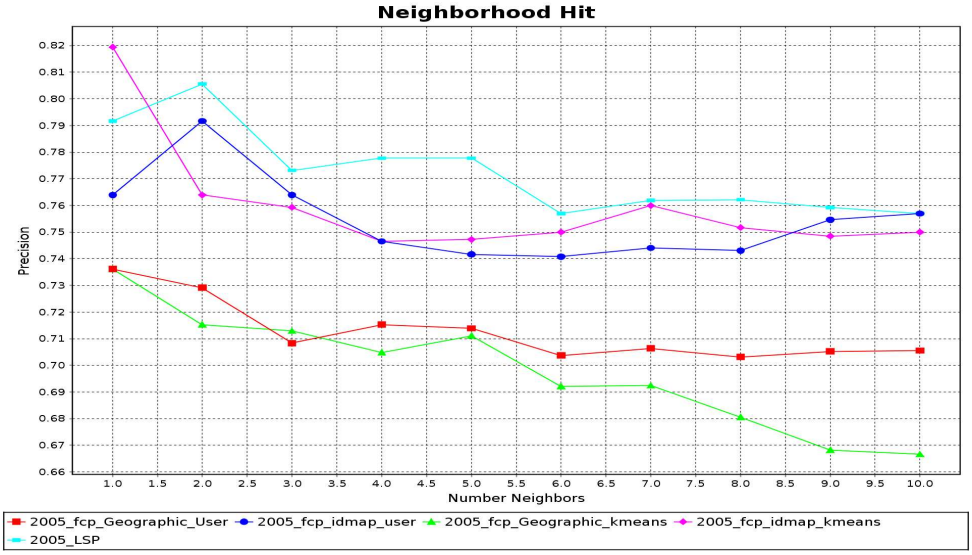


Figura 4.5: Precisão das projeções utilizando diferentes técnicas sobre o conjunto de dados do IDH no ano de 2005.

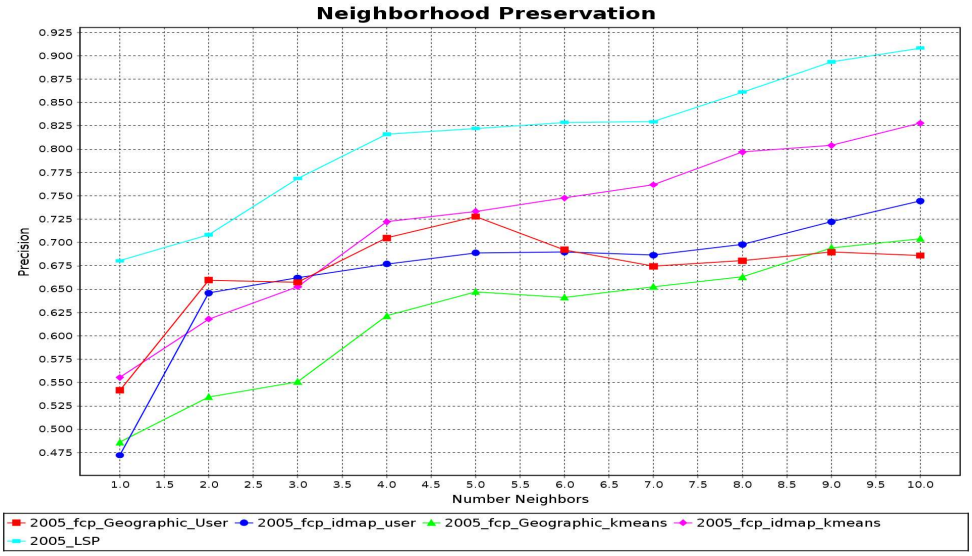


Figura 4.6: Conservação dos vizinhos do espaço original no espaço projetado utilizando diferentes técnicas e configurações sobre o conjunto de dados do IDH no ano de 2005.

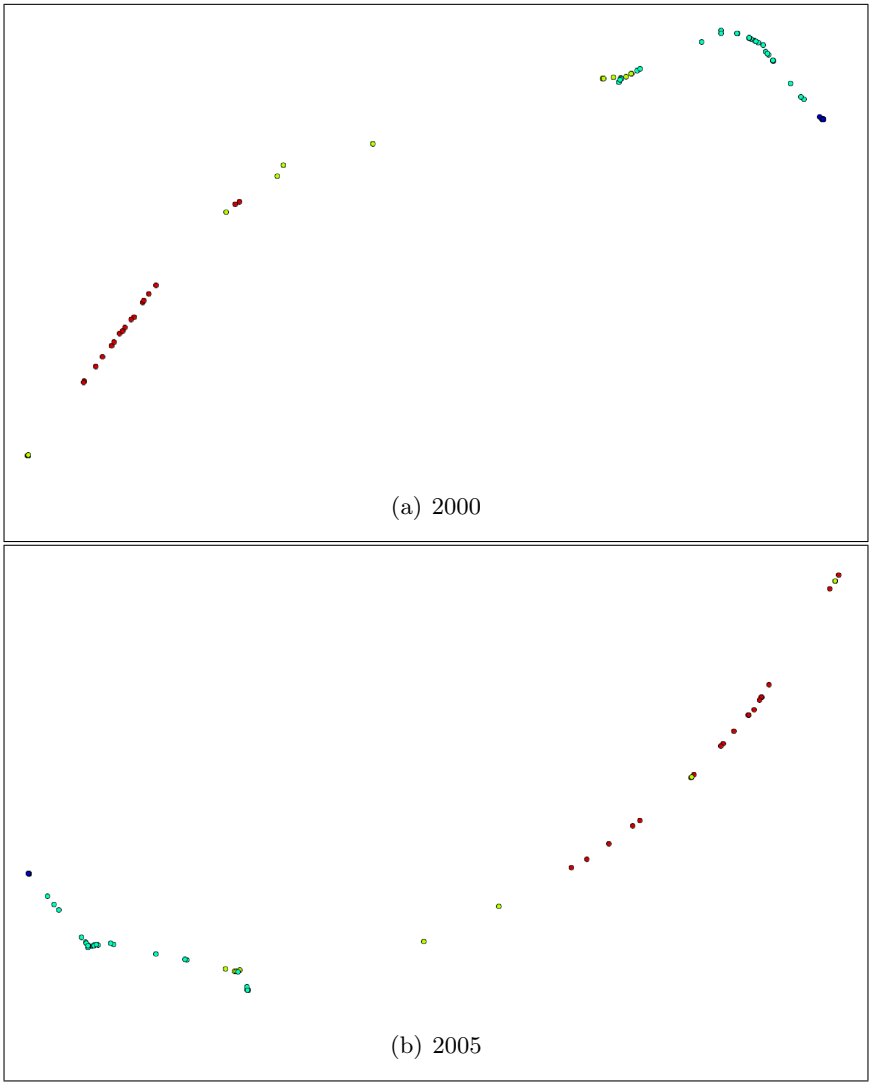


Figura 4.7: Projeções dos dados do IDH em 2000 (a) e 2005 (b) utilizando a LSP em sua abordagem original.

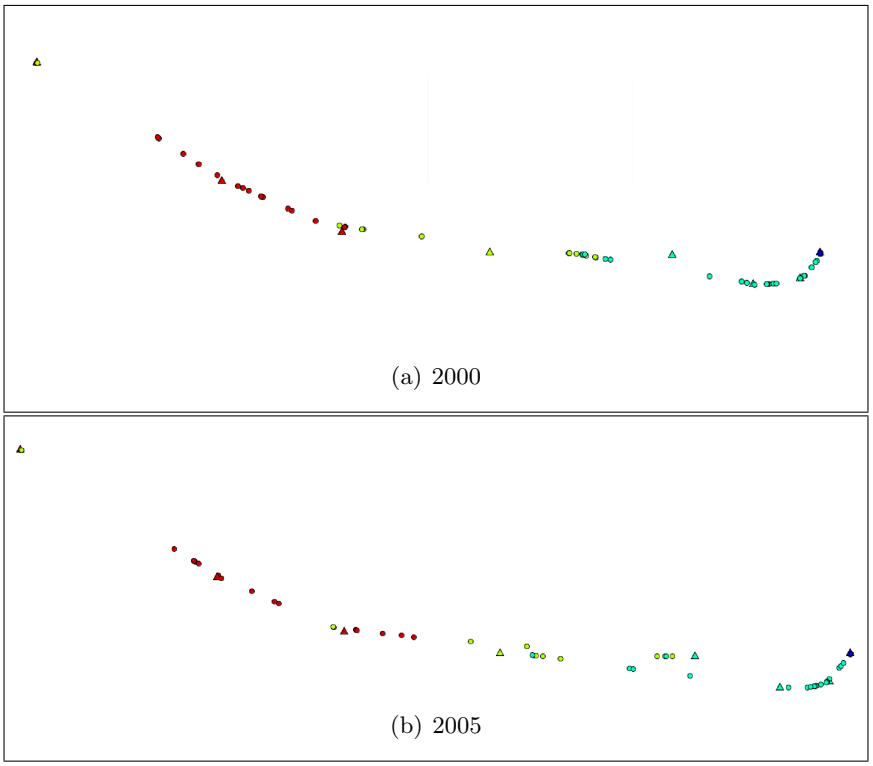


Figura 4.8: Projeções dos dados do IDH em 2000 (a) e 2005 (b) utilizando a técnica LSP$_{FCP}$.

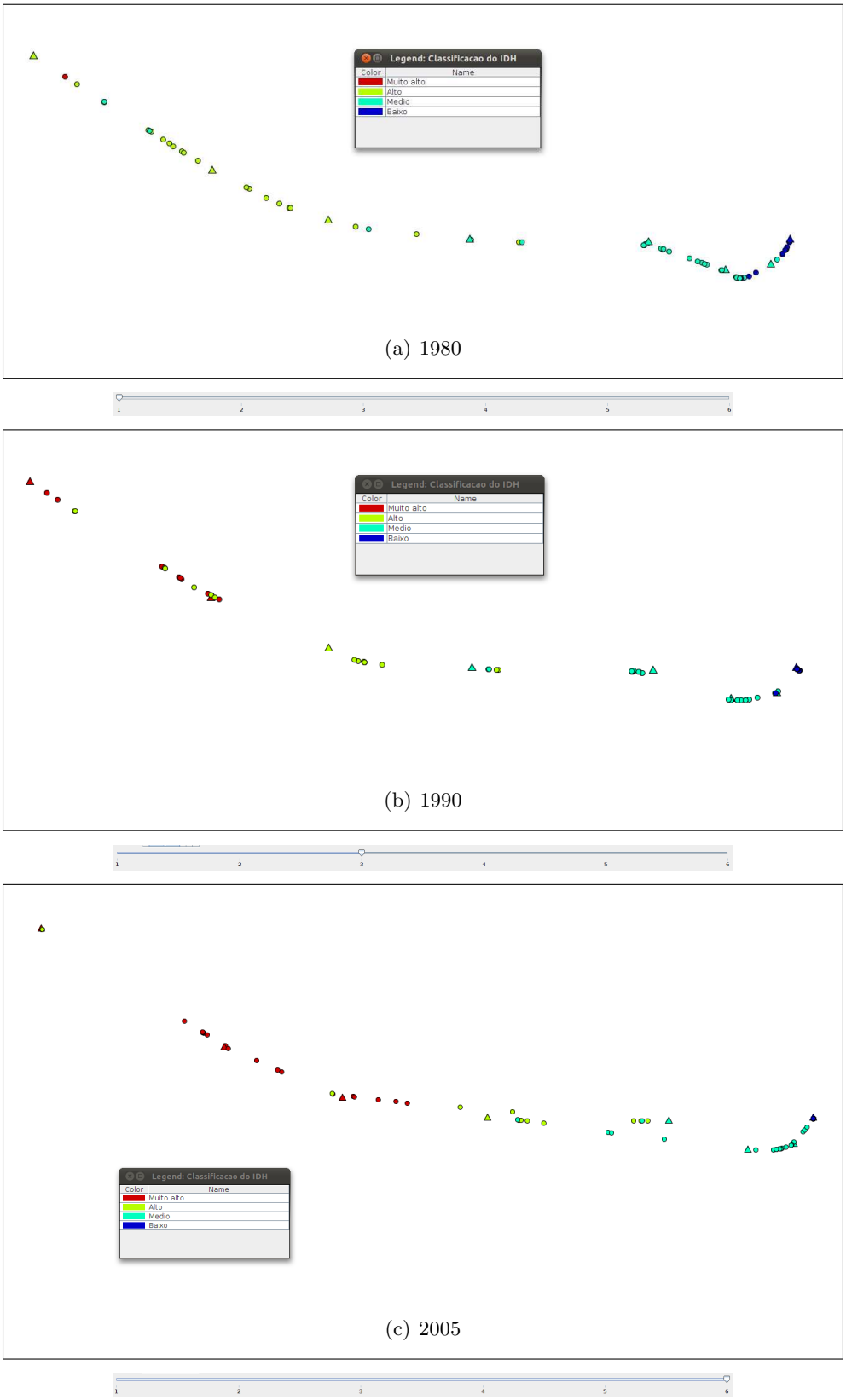
dos pontos vai mudando, dependendo da projeção que está sendo exibida, assim como a cor, caso algum elemento mude de classe. A classe dos elementos é informada em cada *timestamp*, podendo mudar de um para o outro.

O usuário pode controlar a seu critério a velocidade da animação. A qualquer momento a animação pode ser interrompida e retomada, permitindo que o usuário analise o comportamento do conjunto de dados em qualquer instante de tempo.

A animação pode ser um recurso auxiliar no processo de análise, pois quando todas as projeções são exibidas simultaneamente o problema de oclusão visual tende a se agravar devido à grande quantidade de pontos. Um exemplo do problema de oclusão é ilustrado na Figura 4.10: algumas regiões desta visualização apresentam alta densidade de pontos, tornando difícil observar a origem de alguns deles. Além disso, a repetição de elementos na mesma projeção pode prejudicar a interpretação, levando o usuário a pensar que um único elemento pode ser um conjunto de elementos. Para evitar esse problema o usuário deve estar atento a utilização de transparência para cada *timestamp*.

4.2.2 Visualização dos rastros

É possível selecionar um ou mais elementos de interesse para que a animação mostre o rastro, as posições por onde estes elementos passaram, a medida em que a animação avança. Esta é uma característica que pode ajudar bastante na observação de aparentes anomalias, como



(a) 1980

(b) 1990

(c) 2005

Figura 4.9: Ilustração da animação dos pontos: em (a) primeiro instante de tempo, (b) um instante de tempo intermediário e (c) instante de tempo final da animação.

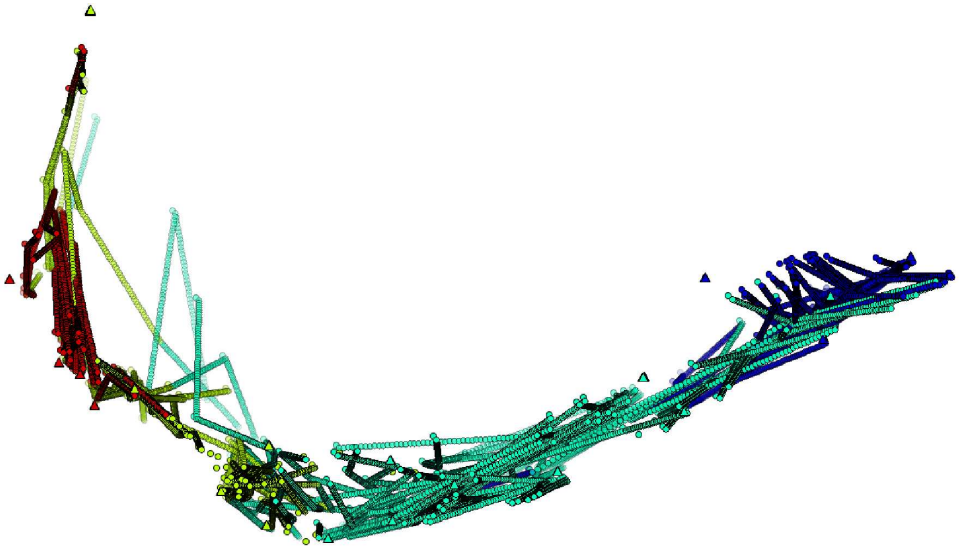


Figura 4.10: Ilustração do problema de oclusão visual.

contra-tendências, durante a animação, como elementos que seguem um fluxo diferente do padrão observado pelos outros elementos. Como exemplo, a Figura 4.11 exibe o rastro correspondente ao país Marrocos (em inglês *Morocco*). A sequência é obtida adotando maior transparência para a origem do ponto e tornando-o mais opaco a medida em que o mesmo avança ao destino. Neste exemplo observa-se que o elemento muda do grupo de países com índice de desenvolvimento baixo (azul) para o grupo de países com índice de desenvolvimento médio (ciano).

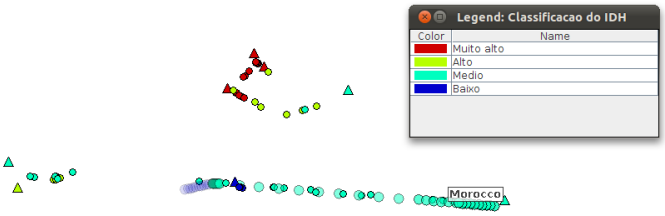


Figura 4.11: Ilustração da utilização de transparência para indicar a direção de movimento entre os anos de 1980 e 2005.

Robertson et al. (2008) citam como problema da visualização de rastro a dificuldade em observar a direção. Para tratar esse problema, o rastro dos pontos pode ser exibido utilizando transparência ou setas, que ajudam a observar a direção. A Figura 4.12 exibe a visualização das projeções dos conjuntos de observações dos índices que compõem o IDH para os anos de 1980 e 1985 utilizando setas para representar os rastros dos pontos. É possível observar que Israel tende, ao longo dos anos, a se aproximar dos países com índice de desenvolvimento alto (verde).

A utilização de setas para indicar a tendência de movimentação dos pontos, quando aplicada a todos os *timestamps*, tornou difícil a observação global do comportamento dos dados devido

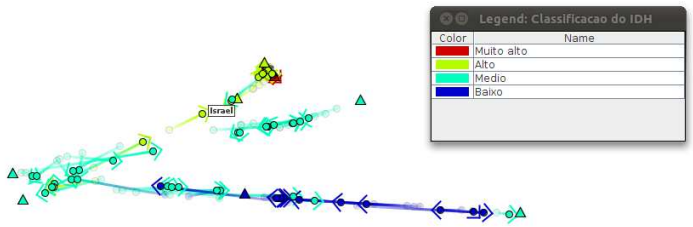


Figura 4.12: Ilustração da utilização de setas para indicar a direção de movimento entre os anos de 1980 e 1985.

à oclusão visual (ver Figura 4.13). Porém, esta funcionalidade mostrou-se bastante útil quando aplicada a um número menor de *timestamps*, dois ou três, como ilustrado na Figura 4.12, ou quando aplicada a um ponto ou um conjunto pequeno de pontos durante todo o período.

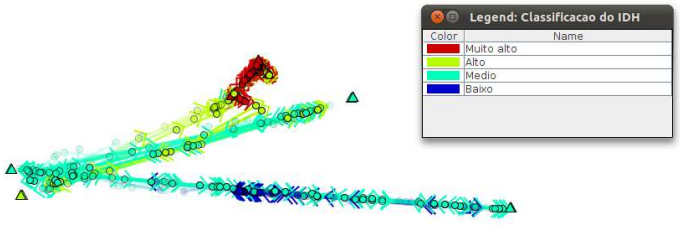


Figura 4.13: Projeções dos dados do IDH utilizando a técnica LSP_{FCP} ilustrando o problema da oclusão visual causada pela utilização de setas em todos os *timestamps*.

A oclusão visual também acontece quando se aplica a exibição do rastro a todos os elementos da projeção e a todos os *timestamps*. A Figura 4.14 ilustra o problema, quando a técnica é aplicada aos mesmos dados da Figura 4.13.

Entretanto, analogamente às setas, esta funcionalidade pode ser bastante útil para observar o rastro de um ponto ou de um conjunto pequeno de pontos. A diferença entre a aplicação desta funcionalidade e das setas é o fato de que esta pode ser aplicada a todos os *timestamps*, como ilustrado na Figura 4.11, que ilustra a exibição do rastro para o país Marrocos entre todos os *timestamps*.

4.2.3 LSP_{FCP} na Projection Explorer

O processo de obter a projeção começa com o usuário informando o local da base de dados. Como tem-se um arquivo de entrada para cada instante de tempo, a aquisição dos dados para a projeção está associada a um diretório que o usuário seleciona. Dessa forma, foi incluída uma

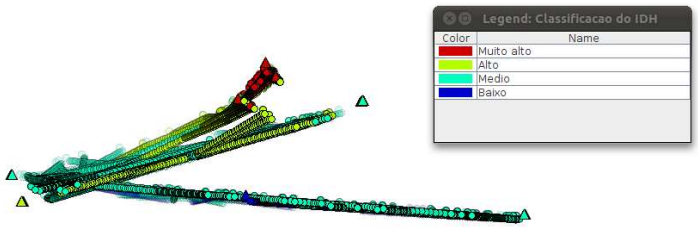


Figura 4.14: Projeções dos dados do IDH utilizando a técnica LSP_{FCP} ilustrando o problema da oclusão visual causada pela a utilização de rastros em todos os elementos do conjunto de dados.

opção no *wizard* da ferramenta para que o usuário informe esse diretório e, de forma automática, é possível gerar as projeções para cada arquivo existente no mesmo. A Figura 4.15 exibe a opção de escolha do diretório.

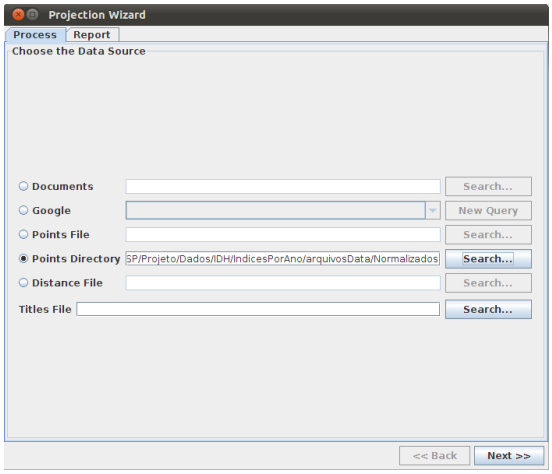


Figura 4.15: *Projection wizard*: escolha da base de dados.

Quando o usuário seleciona a técnica de projeção LSP_{FCP} (ver Figura 4.16), o próximo passo é a definição dos parâmetros a serem utilizados pela técnica. A Figura 4.17 exibe a tela de escolha de parâmetros. Caso o usuário escolha gerar as projeções utilizando pontos de controle lidos de arquivo, opção *File* em *Control Points Choice*, não é necessário alterar a quantidade de pontos de controle em *Number Control Points*, pois este valor deve estar no arquivo. Também é possível gravar um arquivo com os pontos de controle utilizados na projeção para a construção de outras projeções, em *Save Control Points*, armazenando as coordenadas ou os valores dos pontos de controle.

O arquivo de pontos de controle deve seguir o modelo exemplificado pela Tabela 4.2. A primeira linha é utilizada para informar se os pontos de controle foram armazenados no arquivo com as coordenadas cartesianas ou com os valores para cada atributo, para que a PEx realize o

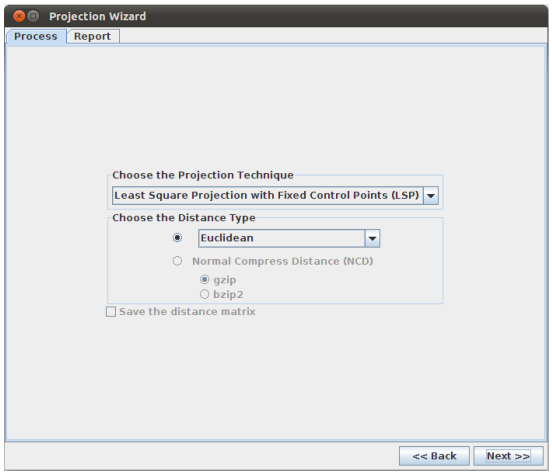


Figura 4.16: *Projection wizard*: escolha da técnica e da métrica de distância.

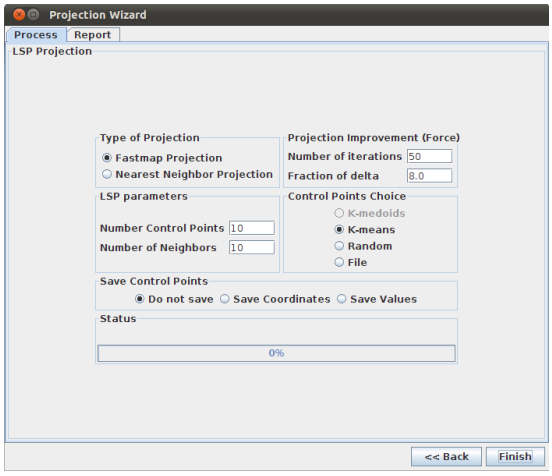


Figura 4.17: Escolha dos parâmetros da técnica LSP_{FCP}.

cálculo das coordenadas utilizando a técnica IDMAP. Na segunda linha a quantidade de pontos de controle é informada e, em seguida, os pontos de controle são apresentados.

Tabela 4.2: Exemplo de formato do arquivo LCP.

```
#Coordenadas ou valores
C
#Quantidade de pontos de controle
14
#Pontos de controle
Kuwait;1373.0559;300.0
Norway;300.0;3857.486
Greece;1954.2689;5301.9604
Burundi;12202.5;3557.1377
Finland;1429.018;5521.0396
Paraguay;5351.5503;7044.893
Australia;1235.5352;5116.392
Sweden;959.3637;4959.935
Togo;9479.148;3834.6921
Jordan;6169.9644;6562.163
Bangladesh;11048.3;4653.635
India;9201.834;5601.312
Brazil;5336.2397;6245.42
Argentina;3099.4749;6647.5977
```

Ao final do processo as projeções são exibidas. Para exemplificar o resultado da projeção, foi utilizado o conjunto de dados de observações dos índices que compõem o IDH. Para cada instante de tempo, neste caso a cada 5 anos, uma projeção é gerada. Os pontos de controle foram escolhidos utilizando-se o método k-*means* implementado na PEx e são exibidos como triângulos, ficam fixos na mesma posição de tela em todas as projeções criadas.

A Figura 4.18 mostra a visualização com todas as projeções justapostas em uma única janela. As setas ligam os pontos em cada projeção e indicam a direção de seu movimento. Cada projeção é exibida com um grau de transparência, sendo a primeira a mais transparente e a última a mais opaca. A transição entre as projeções pode ser exibida utilizando linhas com gradiente de cor (ver Figura 4.19(a)) ou transparência de pontos (ver Figura 4.19(b)); em ambos os casos a representação que mapeia a origem dos pontos recebe maior transparência e vai se tornando mais opaca ao se aproximar do destino. Movimentando o ponteiro da barra de deslizamento é possível controlar a observação da evolução temporal.

4.3 Visualizando séries temporais geo-referenciadas

A ferramenta *Temporal-PEx* (Alencar et al., 2007) foi estendida neste trabalho de forma a permitir a análise de conjuntos de séries temporais com referência geográfica. Como essa

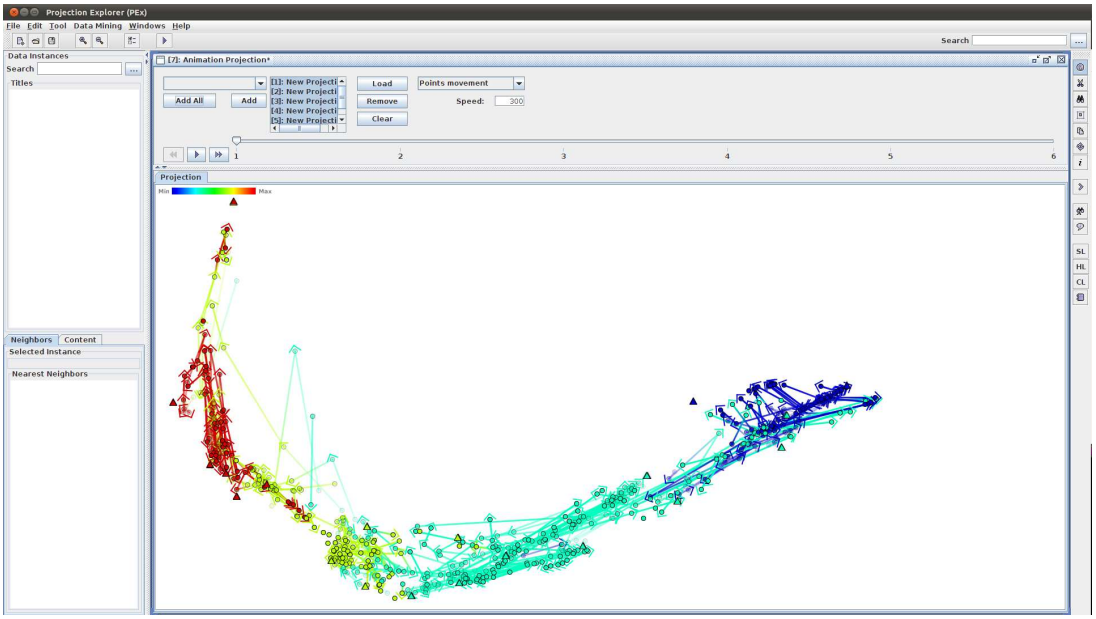


Figura 4.18: Tela exibindo todas as projeções geradas utilizando a técnica LSP_{FCP} para o conjunto de observações dos índices do IDH entre os anos de 1980 e 2005.

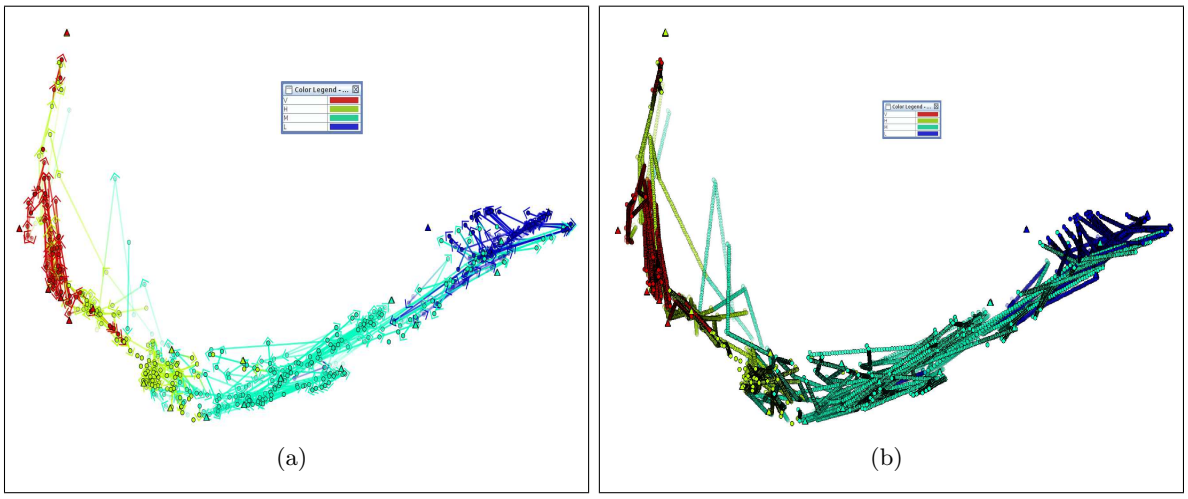


Figura 4.19: Exemplo ilustrando a transição dos pontos entre as projeções: em (a) utilizando linhas com gradiente de cor e (b) utilizando transparência no rastro.

ferramenta já incorpora diversos recursos úteis para a análise de séries temporais, optou-se por estendê-la com essa funcionalidade.

Além de implementar o conjunto de funcionalidades desenvolvidas para a PEx, descritas na Seção 4.2, foi implementada uma abordagem para a visualização das séries com referência geográfica sobre mapas.

Utilizou-se o SDK NASA *World Wind* Java (WWJ)[2] para criar a representação visual dos dados sobre o mapa, pois ela utiliza a linguagem Java, que é a mesma da ferramenta *Temporal-PEx*, e é de código aberto, o que permite que alterações no código sejam feitas, incorporadas a outras ferramentas e redistribuídas.

A ideia é apresentar uma representação geo-referenciada das séries, posicionada em uma representação do globo terrestre. Esta visualização pode ser apresentada simultaneamente à projeção multidimensional, e ambas são coordenadas, ou seja, as interações do usuário em uma das representações (por exemplo, seleção de um elemento) são imediatamente refletidas na outra.

Para gerar uma visualização de um conjunto de séries temporais geo-referenciadas, primeiramente o usuário informa o arquivo que contém o conjunto de dados. A Figura 4.20 exibe a opção de escolha do arquivo de entrada.

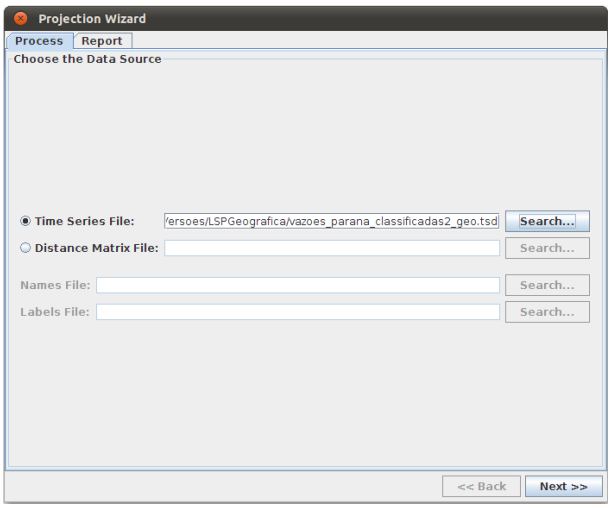


Figura 4.20: *Projection Wizard*: escolha da base de dados.

O próximo passo importante é a escolha da técnica de projeção a ser utilizada, neste caso a técnica LSP_{FCP}, e a métrica de distância desejada (ver Figura 4.21). As seguintes métricas de distância estão implementadas na ferramenta: Distância Euclidiana, *Dynamic Time-Warping* (DTW) (Berndt e Clifford, 1994) e *Compression-based Dissimilarity Measure* (CDM) (Keogh et al., 2007).

Outro passo importante é a escolha dos parâmetros a serem utilizados pela técnica. A Figura 4.22 exibe a tela de escolha de parâmetros. Também na *Temporal-PEx* o usuário conta

[2]http://worldwind.arc.nasa.gov/java/ acessado em 20 de setembro de 2011

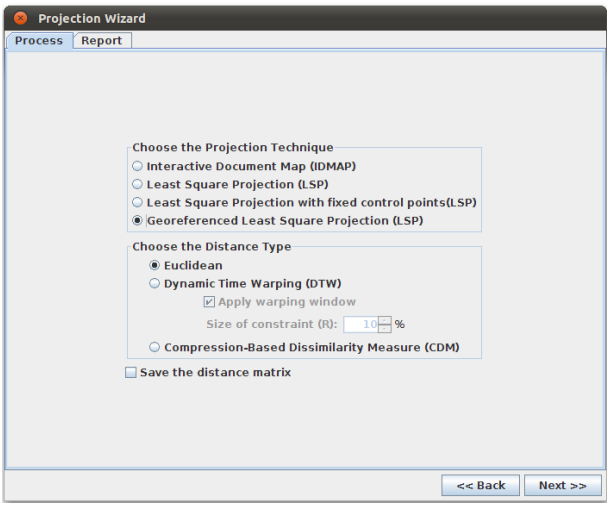


Figura 4.21: *Projection Wizard*: escolha da técnica e da métrica de distância.

com a opção de gerar as projeções utilizando pontos de controle lidos de um arquivo, assim como na PEx.

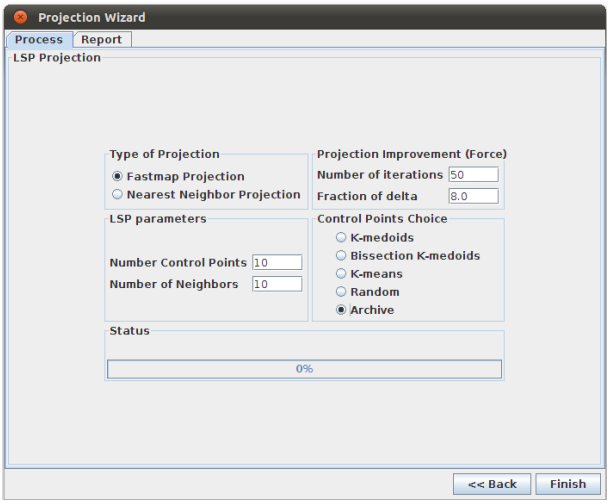


Figura 4.22: *Projection Wizard*: escolha dos parâmetros da técnica LSP_{FCP}.

O conjunto de dados analisado nesta seção é um dos resultados do projeto "Vazões Médias Mensais nos Aproveitamentos Hidrelétricos - Período 1931 a 2005" (ONS, 2006). Para esta análise foram extraídas do conjunto de dados as séries de vazões das usinas localizadas na região hidrográfica do Paraná, que são 76 no total. Cada série é formada por 900 observações mensais correspondentes a uma usina, o que corresponde a 75 anos de observação.

A Figura 4.23 ilustra a tela principal da *Temporal-PEx* já incorporando a visualização do globo terrestre. Na projeção (à esquerda) os pontos amarelos são os pontos de controle e estão fixos em suas posições geográficas. Para os demais pontos utiliza-se transparência para indicar sua movimentação. Os círculos mais opacos estão em sua posição no mapa por similaridade, enquanto que os círculos mais transparentes indicam as posições geográficas. Na visualização

no globo os pontos representam as usinas, posicionados com base na informação geográfica associada a cada série histórica (ver Figura 4.24). As mesmas arestas exibidas na projeção multidimensional são representadas na visualização geográfica. Desta forma, é possível verificar, por exemplo, se os vizinhos mais próximos na projeção por similaridade também são os vizinhos geograficamente mais próximos.

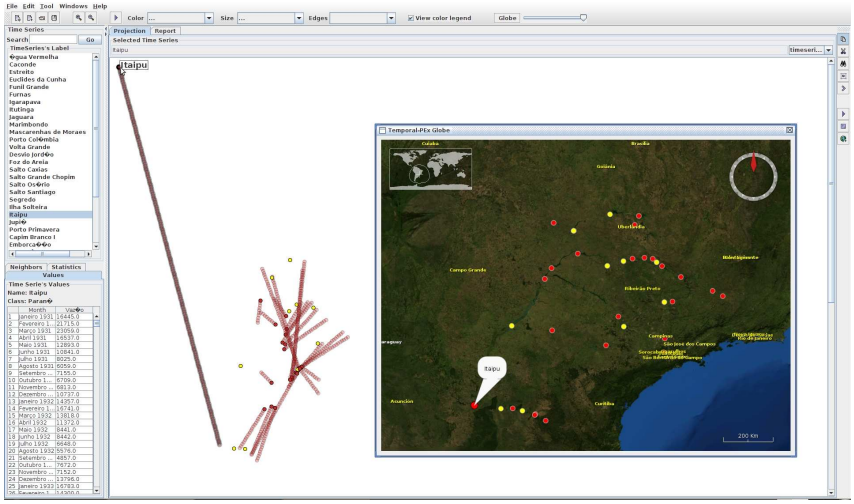
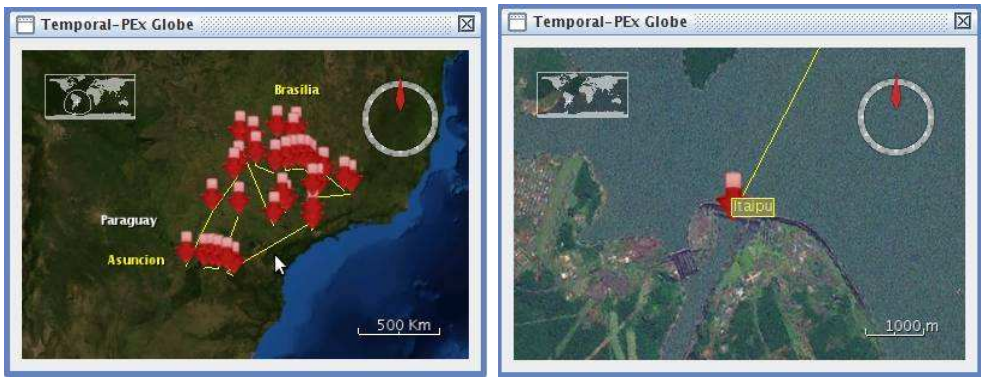


Figura 4.23: Janela principal da ferramenta *Temporal-PEx* exibindo uma visualização gerada utilizando a técnica LSP_{FCP} e a visualização da referência geográfica sobre o mapa.



(a) Local geográfico do conjunto de séries

(b) Local geográfico da série que representa a usina de Itaipu

Figura 4.24: Visualização da informação geográfica.

A ideia, nesta solução, é evidenciar visualmente a relação entre comportamento similar (dos dados multidimensionais) e sua posição geográfica. O pressuposto é que, em muitas situações, espera-se comportamento similar de dados espacialmente próximos, e se isso não ocorre tende a ser uma informação relevante para um analista.

O método proposto utiliza diretamente a referência geográfica para posicionar os pontos de controle, e mantém o posicionamento dos demais pontos em sua vizinhança por meio de uma interpolação - analogamente à LSP original.

Se na projeção multidimensional os vizinhos do objeto p forem os mesmos vizinhos observados no espaço original, o conceito de dependência espacial é válido – ou seja, o posicionamento geográfico é determinante no comportamento. Por outro lado, caso a vizinhança mude, infere-se que tal dependência não ocorre.

Um botão deslizante, localizado na barra de ferramentas da parte superior, permite ao usuário explorar a movimentação dos pontos, enquanto botões localizados na barra de ferramentas à direita controlam uma animação dessa movimentação. Utilizando essas funções o usuário pode interagir melhor com a visualização, observando a aproximação ou o afastamento dos pontos que representam os objetos do conjunto.

A primeira opção da animação (ver Figura 4.23) projeta diversos pontos no plano e apenas altera a sua opacidade, em intervalos de tempo pré-definidos. A projeção destes pontos utilizando transparência, de forma estática, descreve a trajetória de cada elemento da coleção. Outra forma de animação é a utilização de linhas com gradiente de transparência, como mostra a Figura 4.25 que ilustra uma projeção do mesmo conjunto de dados da Figura 4.23. As cores mapeiam as classes, que correspondem às bacias hidrográficas a que cada usina hidrelétrica pertence.

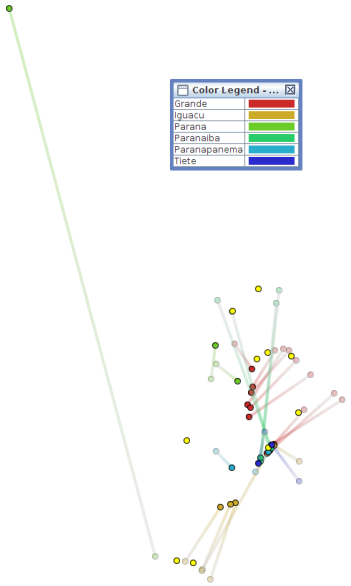


Figura 4.25: Projeção do conjunto de séries que descrevem vazões históricas de algumas usinas hidrelétricas da bacia do Paraná utilizando linhas com gradiente para o traçado do caminhos dos pontos.

Esta abordagem foi apresentada como poster na 13ª. Escola de Séries Temporais e Econometria (Tiburtino e Oliveira, 2009).

4.4 Considerações finais

Este capítulo apresentou o conjunto de funcionalidades desenvolvidas com o propósito de auxiliar a exploração interativa de conjuntos de dados com referências, temporal ou geográfica. Dessa forma, são oferecidas aos usuários diferentes visões dos conjuntos de dados.

A principal contribuição é a técnica *Least Square Projection* com Pontos de Controle Fixos (LSP_{FCP}), que permite utilizar as referências presentes nos dados de forma diferenciada na geração das projeções.

No caso da referência temporal, para cada instante de tempo é gerada uma projeção e os pontos de controle são os mesmos em todas elas. Em seguida essas projeções são justapostas e a transição entre elas pode ser exibida utilizando linhas com gradiente ou transparência nos pontos. O uso de animação como forma de representar a evolução temporal dos conjuntos de dados também é uma opção que pode ser de bastante utilidade.

Como os pontos de controle são escolhidos em um determinado *timestamp*, podem não ser os pontos mais representativos nos demais *timestamps*. Esse fato acarreta alguma perda na precisão das projeções.

As referências geográficas podem ser utilizadas como coordenadas dos pontos de controle, para resolver o sistema linear, ou podem ser projetadas sobre um mapa.

O próximo capítulo descreve os resultados de alguns estudos de casos como forma de validação da estratégia proposta e das funcionalidades auxiliares implementadas.

Capítulo

5

Resultados

5.1 Considerações iniciais

Neste capítulo são apresentados alguns estudos realizados como forma de validar a abordagem proposta, e as funcionalidades auxiliares implementadas. Os estudos de caso realizados têm como objetivo mostrar a utilidade e potenciais limitações da abordagem na observação do comportamento temporal dos dados, auxiliada pela animação e pelas funcionalidades de exibição do rastro, integradas ao uso de Coordenadas Paralelas.

Para facilitar o trabalho de extração de dados para os estudos, partindo de dados existentes em outras bases, foi implementado um módulo chamado de *PExParser*. Além de formatar as bases existentes em seu repositório, obtidas do relatório das Nações Unidas e do *Gapminder*, este módulo também é capaz de concatenar essas diferentes bases, permitindo que o usuário escolha quais indicadores irão compor os arquivos de entrada para a ferramenta PEx. Também é possível criar arquivos contendo pontos de controle, no formato LCP descrito na Seção 4.2.3.

A Figura 5.1 exibe a tela principal do módulo. A árvore de diretórios no lado esquerdo da tela permite a escolha dos indicadores das bases encontradas no repositório do módulo, que serão utilizados como atributos no arquivo de entrada para a ferramenta PEx. No lado direito é possível observar os indicadores escolhidos. Também é possível determinar se os dados recebem uma pré-classificação (por continente por exemplo). Em seguida basta escolher o diretório e a nova base de dados é gerada.

A Figura 5.2 ilustra a tela da função de criação do arquivo de pontos de controle. O usuário seleciona os países na lista a direita, a esquerda são exibidos aqueles que já foram selecionados.

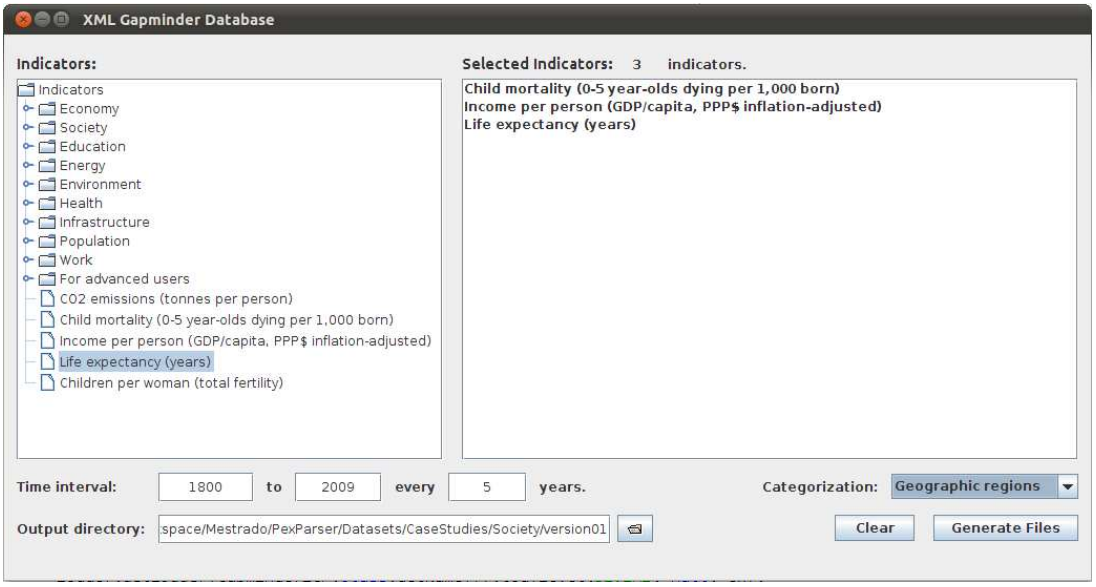


Figura 5.1: Módulo *PExParser*.

É possível carregar os países a partir de outro arquivo de pontos de controle. O usuário pode escolher utilizar o par latitude e longitude como coordenadas cartesianas, utilizar os valores das observações do conjunto de dados ou aplicar a técnica IDMAP e utilizar as coordenadas de tela retornadas. O formato do arquivo de pontos de controle foi descrito na Seção 4.2.

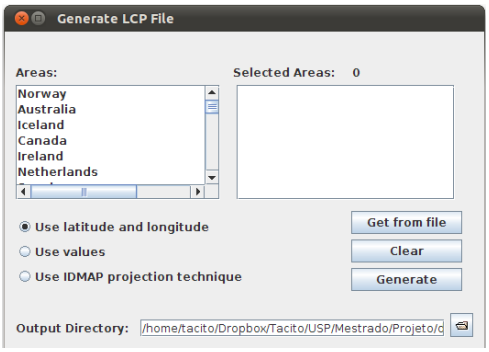


Figura 5.2: Módulo PExParser: criação do arquivo de pontos de controle.

Esta funcionalidade pode ser de grande utilidade para os trabalhos realizados pelo grupo de pesquisa em Visualização de Informação e Computação Gráfica do ICMC, permitindo compor diferentes bases de dados multidimensionais variantes no tempo para serem utilizadas na validação dos trabalhos.

A seguir três estudos de casos são apresentados: um deles com dados do Relatório de Desenvolvimento Humano e outros dois com dados obtidos da base do *Gapminder*.

5.2 Índice de Desenvolvimento Humano

O Índice de Desenvolvimento Humano (IDH) busca avaliar o desenvolvimento de uma sociedade considerando não apenas a questão econômica, mas também características sociais que influenciam na qualidade da vida humana (Sen, 2000). Os dados utilizados no cálculo do IDH, apresentado no Relatório de Desenvolvimento Humano do Programa das Nações Unidas para o Desenvolvimento (PNUD), são fornecidos por várias fontes públicas internacionais [1]. Anteriormente ao ano de 2010, o cálculo do índice era obtido pela média aritmética de três medidas relativas à renda, educação e longevidade. As mudanças na metodologia a partir de 2010 impedem a comparação com índices dos anos anteriores. Entre as mudanças, destacam-se: a substituição do PIB (Produto Interno Bruto) pela Renda Nacional; e a substituição do índice de alfabetização da população de um país pelos anos médios de estudo.

- **Renda** - A renda é calculada utilizando o Produto Interno Bruto (PIB) por pessoa do país.
- **Educação** - Para avaliar a educação são utilizados dois indicadores: a taxa de alfabetização de pessoas com 15 anos ou mais; e o somatório das pessoas, independentemente de idade, matriculadas em algum curso dividido pela quantidade de pessoas entre 7 e 22 anos da localidade.
- **Longevidade** - Esta medida indica a expectativa de vida de uma pessoa nascida em uma localidade.

O IDH varia entre zero e um, e segundo o Relatório de Desenvolvimento Humano nos anos anteriores à 2010, os países são classificados da seguinte maneira:

- IDH entre 0 e 0,499 - países de desenvolvimento baixo;
- IDH entre 0,500 e 0,799 - países de desenvolvimento médio;
- IDH entre 0,800 e 0,899 - países de desenvolvimento alto;
- IDH entre 0,900 e 1 - países de desenvolvimento muito alto.

O relatório do Programa das Nações Unidas para o Desenvolvimento, publicado no ano de 2008, avalia 182 países. Porém, para este estudo de caso foram excluídos os países para os quais os indicadores não estão disponíveis para todo o intervalo de tempo entre 1980 e 2005. Dessa forma, este conjunto de dados é formado pela observação de quatro índices que compõem o IDH de 72 países nos anos de 1980, 1985, 1990, 1995, 2000 e 2005. Os países são classificados seguindo o mesmo critério do relatório do Programa das Nações Unidas para o Desenvolvimento.

[1] Dados encontrados em http://hdr.undp.org/en/statistics/data acessado em 20 de setembro de 2011

Utilizando esta base de dados, gerou-se a sequência temporal de projeções com a técnica LSP_{FCP} (ver Figura 5.3). Os pontos de controle foram escolhidos segundo o seguinte critério: dois elementos de cada grupo de classificação (ver Figura 5.4) que não mudaram de classe ao longo do tempo ou mudaram apenas uma vez, em um total de 8 pontos:

- Noruega (em Inglês, *Norway*);
- Estados Unidos (em Inglês, *United States*);
- Emirados Árabes Unidos (em Inglês, *United Arab Emirates*);
- Hungria (em Inglês, *Hungary*);
- China;
- Paraguai (em Inglês, *Paraguay*);
- Moçambique (em Inglês, *Mozambique*);
- Guiné-Bissau (em Inglês, *Guinea-Bissau*).

Analisando a sequência temporal é possível observar a tendência de formação de dois grupos: um grupo com os países que apresentam desenvolvimento muito alto e alto, a esquerda, e outro grupo com os países que apresentam desenvolvimento médio e baixo, a direita. Esses grupos estão destacados na Figura 5.5.

Utilizando a funcionalidade de animação dos pontos com exibição do rastro, foi possível observar dois países (Portugal e Coreia do Sul) que começam com desenvolvimento médio, em 1980, e evoluem até o desenvolvimento muito alto, em 2005. O comportamento desses países é ilustrado na Figura 5.6. Em um primeiro momento, entre 1980 e 1985, os dois países tendem a se juntar aos países do grupo de IDH baixo e médio. Porém, gradualmente, eles se deslocam para o grupo dos países de IDH muito alto e alto.

Observando a Figura 5.7, entre os anos de 2000 e 2005 Portugal segue a tendência de movimentação dos países do grupo. Porém, a Coreia do Sul não segue o mesmo comportamento. Analisando os dados utilizando a técnica de Coordenadas Paralelas, ilustrada na Figura 5.8, observa-se um salto nos índices da Coreia do Sul, principalmente naqueles referentes à educação, quando comparados aos de Portugal. Na Figura 5.8(a), Portugal é representado pela linha de cima e a Coreia do Sul pela linha de baixo, enquanto que, na Figura 5.8(b), Portugal é representado pela linha de baixo e a Coreia do Sul pela linha cima.

Para analisar a evolução do IDH nos continentes utilizou-se o par de coordenadas latitude e longitude de cada país para posicioná-los. Comparando a situação ilustrada na Figura 5.9(a), IDH no ano de 1980, e a ilustrada na Figura 5.9(b), IDH no ano de 2005, observa-se que a Europa deu um salto de qualidade, passando da maioria de países com desenvolvimento médio para países com desenvolvimento muito alto, enquanto que o continente Africano praticamente não apresentou evolução. A Oceania, apesar de ser representada por apenas dois países nessa base (Austrália e Nova Zelândia), apresentou evolução similar à evolução da Europa.

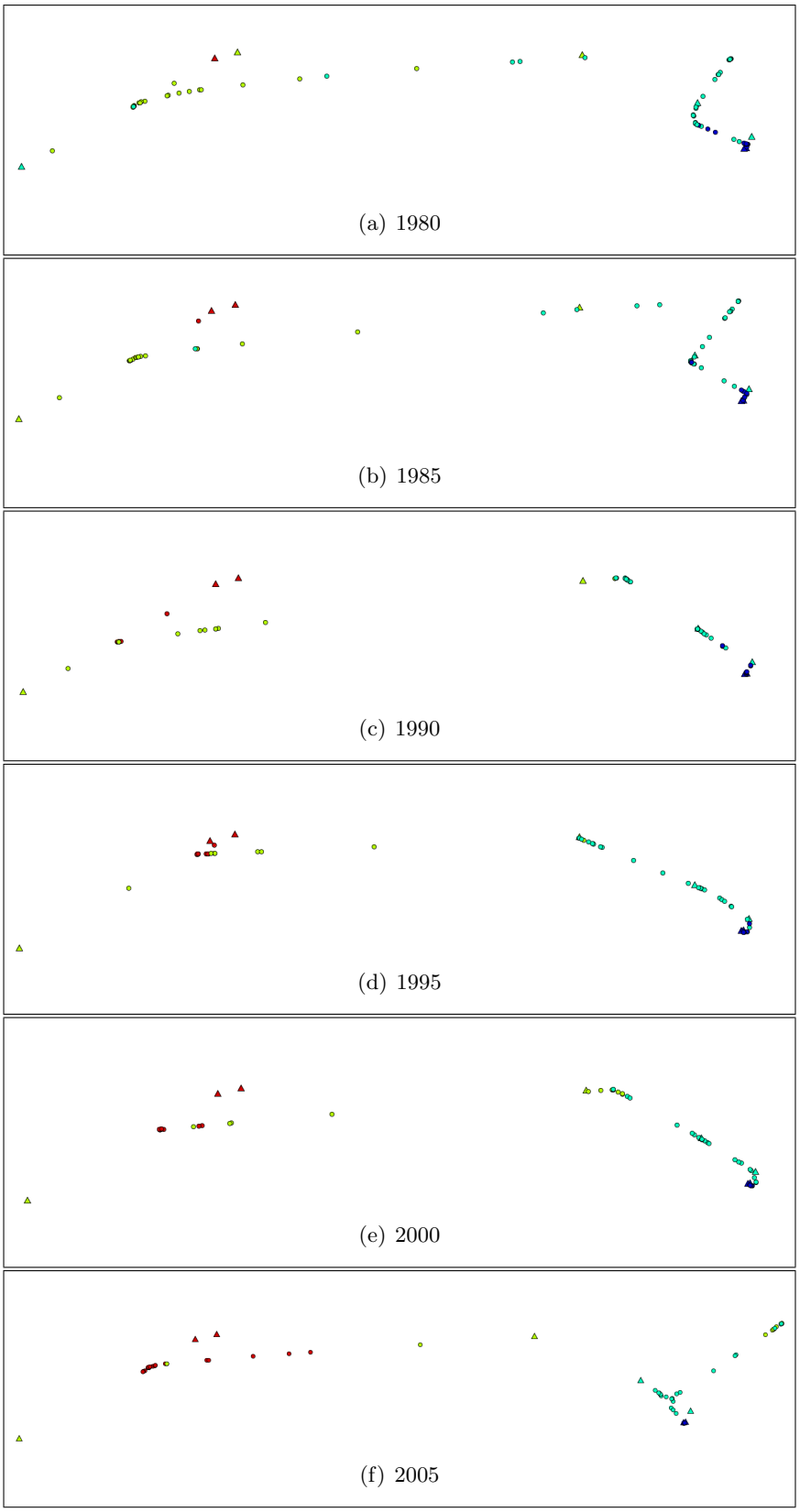


Figura 5.3: Sequência temporal de projeções para os dados do IDH entre os anos de 1980 e 2005.

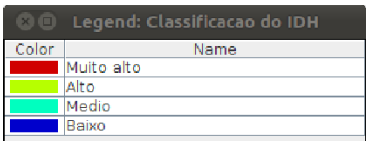


Figura 5.4: Legenda das cores nos mapas da sequência temporal.

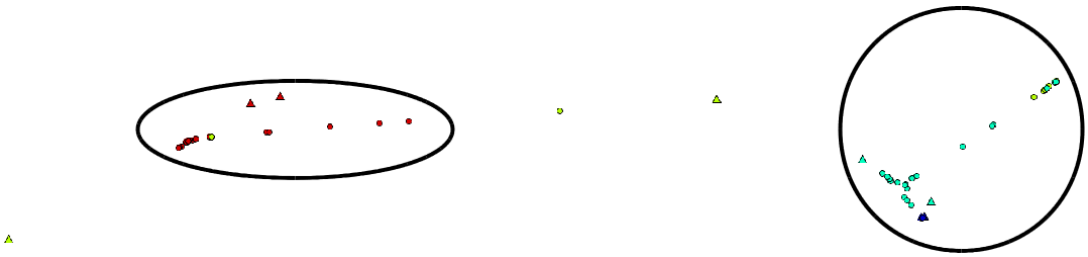
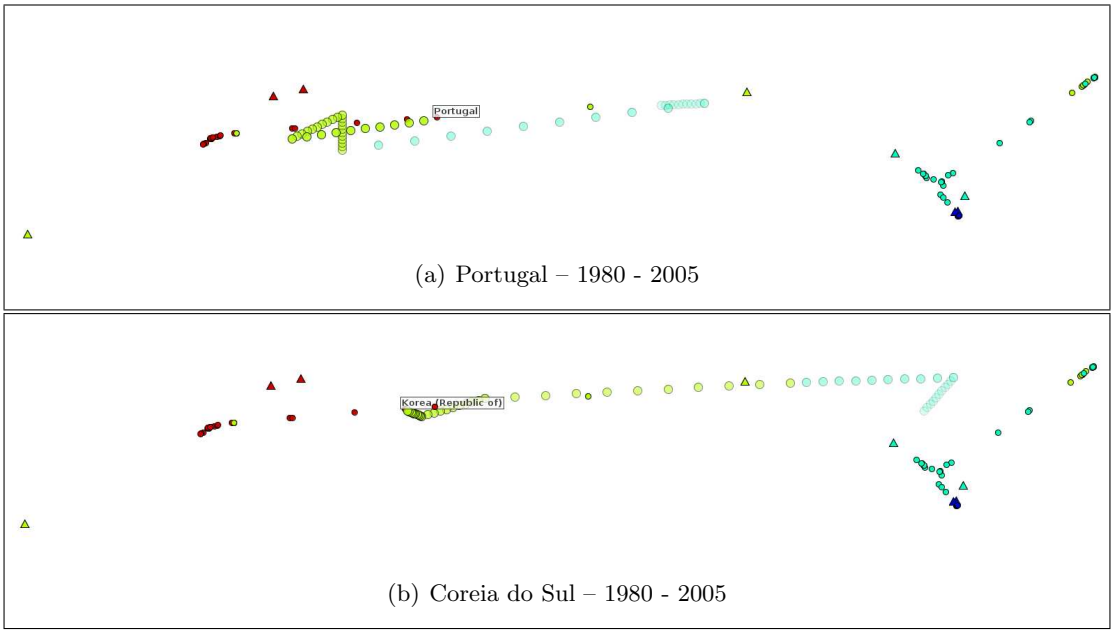


Figura 5.5: Formação de dois grupos de países no ano de 2005: um grupo de países com desenvolvimento muito alto e alto (a esquerda) e outro grupo de países com desenvolvimento médio e baixo (a direita).



(a) Portugal – 1980 - 2005

(b) Coreia do Sul – 1980 - 2005

Figura 5.6: Observação do comportamento de Portugal e da Coreia do Sul ao longo do tempo.

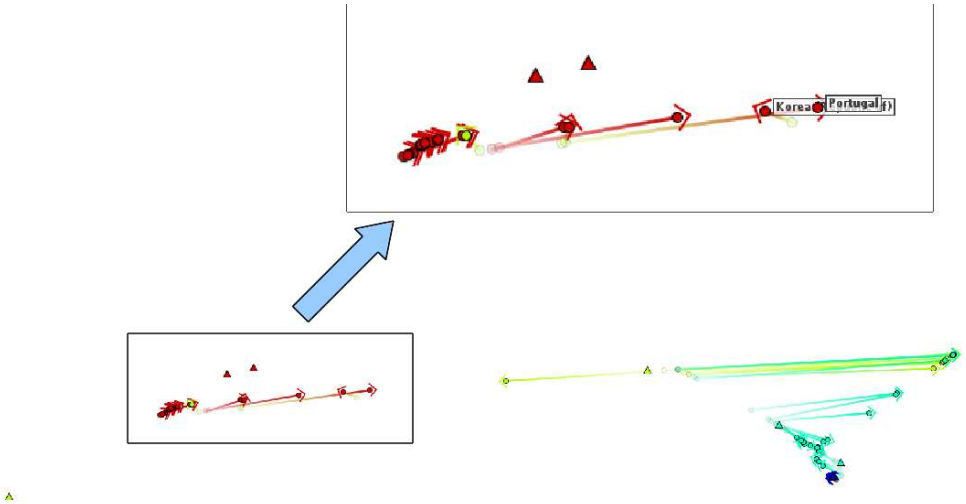
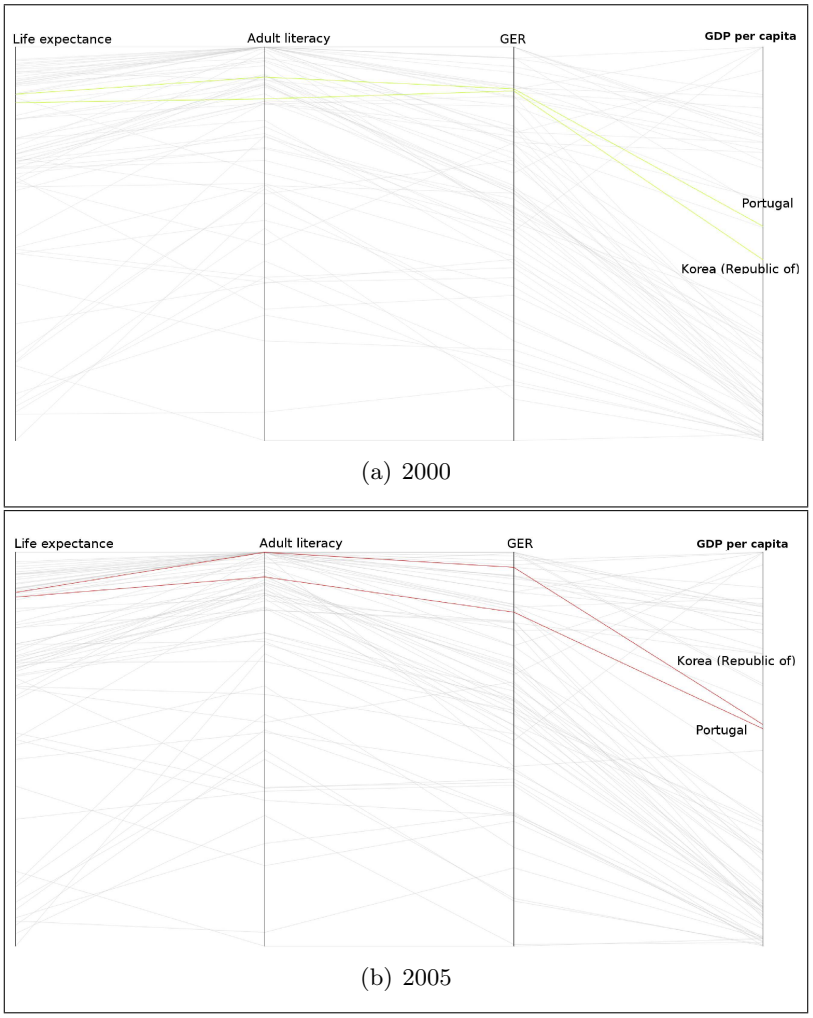


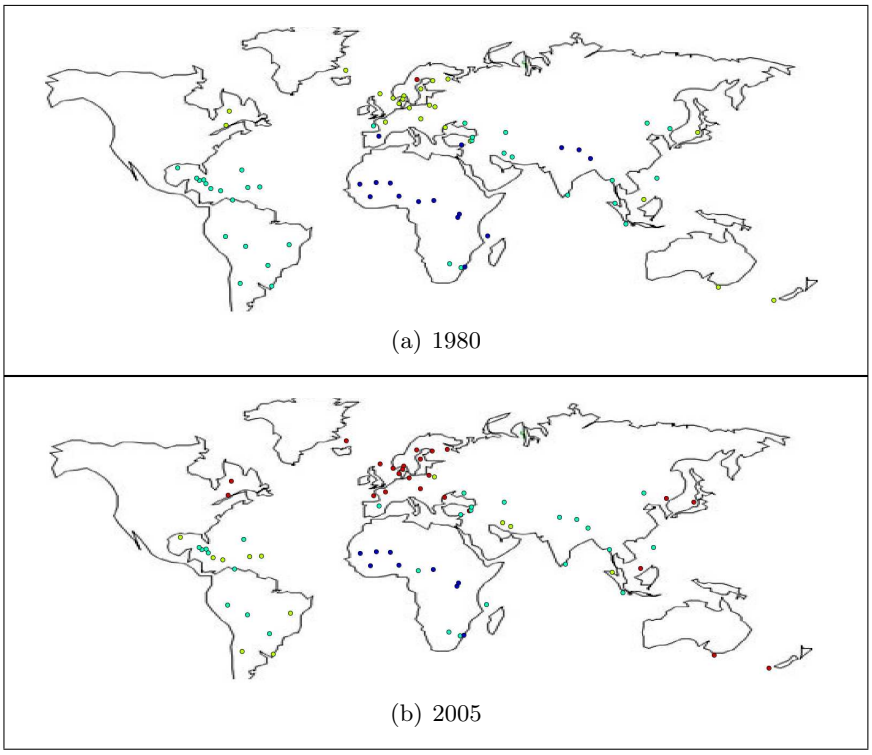
Figura 5.7: Entre os anos de 2000 e 2005, Portugal parece seguir a tendência de movimento dos países do grupo, porém a Coreia do Sul apresenta outro comportamento.



(a) 2000

(b) 2005

Figura 5.8: Análise dos dados de 2000 e 2005 de Portugal e Coreia do Sul utilizando Coordenadas Paralelas.



(a) 1980

(b) 2005

(c) Legenda

Figura 5.9: Comparação do IDH em 1980 e 2005 utilizando a informação geográfica.

5.3 Dados do Gapminder

O projeto *Gapminder*[2] surgiu com o desenvolvimento de uma ferramenta para visualização de dados estatísticos, chamada de *Trendalyzer* e ficou disponível na forma de *web-service* com o nome de *Gapminder World*. Este projeto agrega em sua coleção dados de diferentes categorias coletados por diversas entidades reconhecidas pela Organização das Nações Unidas (ONU) referentes a aproximadamente 250 localidades. A Fundação *Gapminder* não é a provedora dos dados, porém, em alguns casos, os dados provenientes de diversas fontes passam por uma compilação. Nos arquivos com os dados coletados as fontes são citadas, os que passam por compilação têm como fonte o termo "*various sources*". Mesmo assim, a Fundação *Gapminder* disponibiliza os dados para serem baixados e utilizados.

As categorias dos dados são: economia, sociedade, educação, energia, meio ambiente, saúde, infraestrutura, população e trabalho. Entretanto, alguns dados não são categorizados.

Para a construção dos arquivos de entrada, utilizando o módulo *PExParser*, o intervalo de tempo a ser considerado é definido pela combinação de indicadores escolhidos. O intervalo é dado pelo conjunto de anos comum a todos os indicadores.

Assim sendo, dois estudos são apresentados a seguir.

[2]http://www.gapminder.org acessado em 20 de setembro de 2011

5.3.1 Dados do grupo economia

Para realizar um estudo com dados do grupo economia, foram escolhidos 6 indicadores: taxa de juros do banco central (em porcentagem anual), exportação (em porcentagem do Produto Interno Bruto (PIB)), PIB per capita, importação (em porcentagem do PIB), inflação (em porcentagem anual) e a balança comercial (em porcentagem do PIB). Estes indicadores foram escolhidos por serem conhecidos e representam renda, crescimento e comércio das localidades.

Para a composição do intervalo de tempo foram escolhidos anos que apresentavam 100 ou mais localidades com observação dos indicadores. Selecionando-se apenas as localidades que apresentaram observação em todos os *timestamps* da sequência temporal, o total de localidades que compõem a base de dados foi de 160 e o intervalo de tempo foi de 1995 a 2005.

Para a criar a sequência temporal, utilizou-se a técnica LSP_{FCP} com 16 pontos de controle escolhidos pelo algoritmo k-*means* para os dados do ano de 2005. As localidades são classificadas de acordo com a renda nacional bruta per capita, calculada de acordo com o método Atlas[3], pelos grupos de renda: localidades de renda alta pertencentes ao grupo *Organisation for Economic Co-operation and Development* (OECD), localidades de renda alta não pertencentes ao grupo OECD, localidades de renda média alta, localidades de renda média baixa e localidades de renda baixa (ver Figura 5.10). A faixa de valores de cada grupo é a seguinte:

- Renda de $995 ou menos - localidade de renda baixa;
- Renda entre $996 e $3,945 - localidades de renda média baixa;
- Renda entre $3,946 e $12,195 - localidades de renda média alta;
- Renda de $12,196 ou mais - localidades de renda alta.

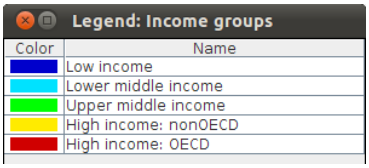


Figura 5.10: Legenda das cores nos mapas da sequência temporal.

Na Figura 5.11 são exibidas três projeções da sequência temporal, em 5.11(a) é ilustrada a projeção para os dados do ano de 1995, em 5.11(b) para o ano de 2000 e em 5.11(c) para o ano de 2005.

De acordo com a Figura 5.11 verifica-se que ao longo do tempo, a maioria das localidades que apresentam renda alta fica mais a esquerda do mapa, enquanto que as localidades de renda baixa ficam mais à direita. Analisando-se a distribuição dos pontos no mapa, com o auxílio da

[3]disponível em http://data.worldbank.org/about/country-classifications/world-bank-atlas-method acessado em 20 de setembro de 2011

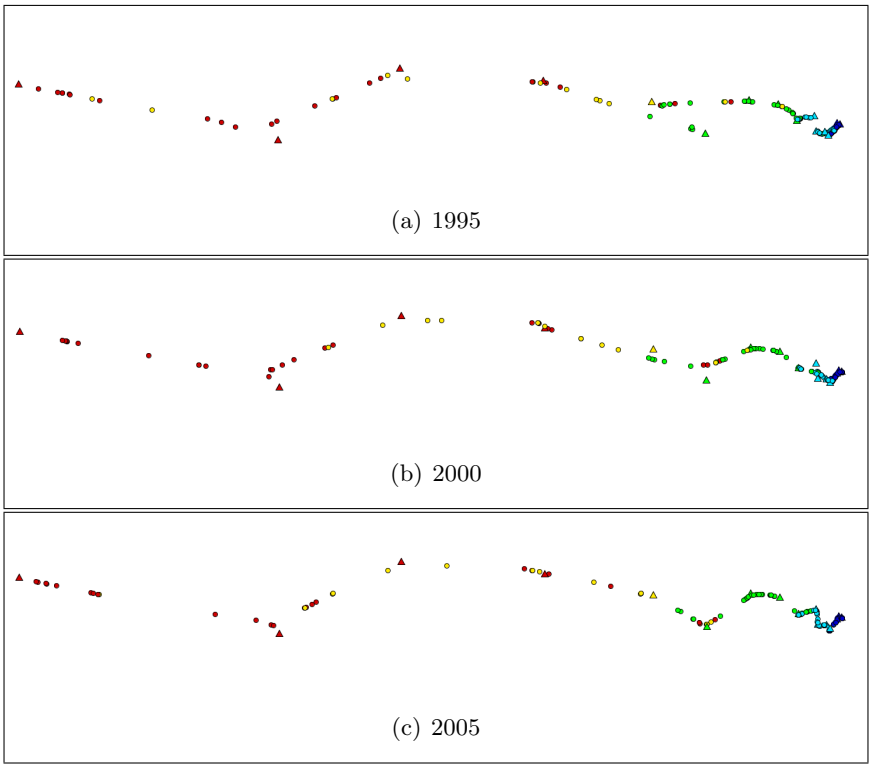


Figura 5.11: Sequência temporal de projeções para os dados do grupo economia entre os anos de 1995 e 2005.

técnica de Coordenadas Paralelas, é possível observar que o indicador referente ao PIB per capita determina o posicionamento, como ilustrado para o ano de 2005 na Figura 5.12. Nas imagens ilustrando a técnica de Coordenadas Paralelas, os eixos mapeiam os indicadores da seguinte forma: **A** representa a taxa de juros do banco central (em porcentagem anual), **B** representa a exportação (em porcentagem do PIB), **C** representa o PIB per capita, **D** representa a importação (em porcentagem do PIB), **E** representa a inflação (em porcentagem anual) e **F** representa a balança comercial (em porcentagem do PIB).

Entretanto, no último ano de observação, algumas localidades de renda alta (Eslováquia, República Checa, Croácia, Hungria e Estônia) são posicionadas próximas a localidades de renda média alta (Argentina, St. Kitts e Nevis, Uruguai, Palau e Chile). Esse grupo é destacado na Figura 5.13(a), sendo os pontos selecionados. Observando os indicadores utilizando Coordenadas Paralelas, ilustrada na Figura 5.13(b), observa-se que o PIB per capita desse grupo de localidades apresenta valores próximos, mesmo recebendo uma classificação diferente segundo os critérios do *Gapminder*.

Utilizando-se a animação dos pontos também observou-se que as localidades se movimentam muito pouco. Porém, a movimentação de Macau (China) destaca-se das demais. A Figura 5.14 ilustra a movimentação de Macau utilizando o rastro em todos os *timestamps*. Em um primeiro momento, entre 1995 e 2000, sua movimentação é da esquerda para a direita de forma acentuada. Observando os dados, notamos uma queda no PIB per capita, que em 1995 era de $15,144.83

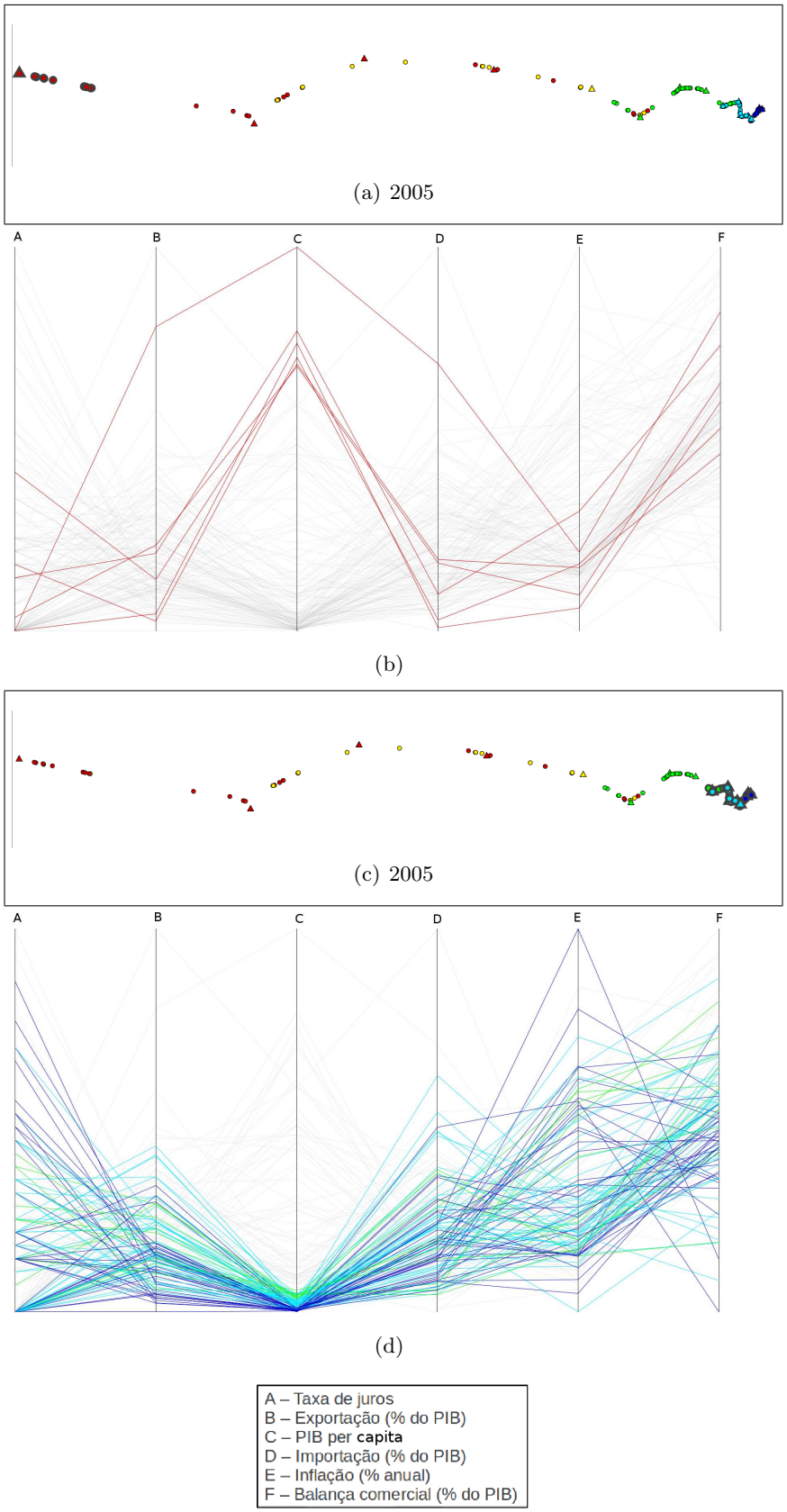


Figura 5.12: Ilustração do efeito do PIB per capita no posicionamento dos pontos em 2005: em (a) os pontos mais a esquerda são selecionados e em (b) observa-se os mesmos destacados, assim como em (c) os pontos mais a direita são selecionados e em (d) observa-se os mesmos destacados.

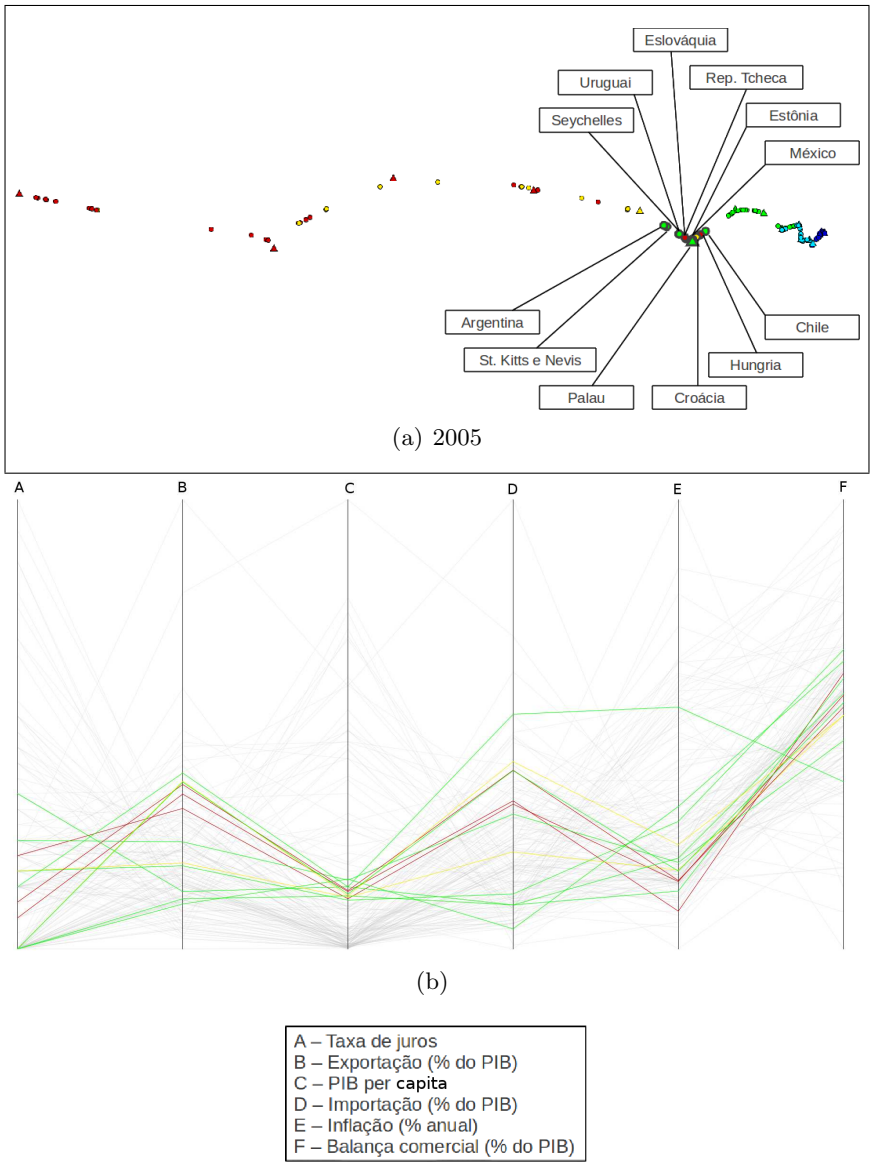


Figura 5.13: Ilustração do efeito do PIB per capita no posicionamento dos pontos em 2005: em (a) os pontos selecionados representam localidades de classificação diferente, segundo o *Gapminder*, porém em (b) pode-se observar que elas possuem valores para o PIB per capita similares.

e passou para $13,839.47 em 2000. A partir de 2001 o PIB per capita de Macau começa a subir, chegando a $22,032.09 em 2005, e nesse período observa-se uma movimentação no sentido contrário. A Figura 5.15 ilustra a variação do PIB per capita por Coordenadas Paralelas junto com as projeções para cada ano.

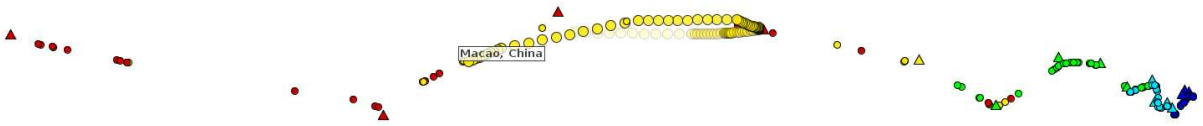


Figura 5.14: Ilustração do rastro de Macau (China) entre os anos de 1995 e 2005.

5.3.2 Componentes do IDH

Em sua coleção de bases de dados, o projeto *Gapminder* conta com alguns indicadores que fazem parte do cálculo do IDH. Três desses indicadores foram escolhidos para um novo estudo: renda per capita, expectativa de vida e educação de adultos. Essa base de dados se diferencia da utilizada no estudo descrito na Seção 5.2, principalmente, pelo intervalo de tempo, que agora vai do ano 2000 ao ano 2005 com observações anuais.

A sequência temporal foi criada utilizando-se a técnica LSP_{FCP} com 16 pontos de controle escolhidos pelo algoritmo k-*means* no ano de 2005. Assim como no estudo apresentado na Seção 5.3.1, as localidades são classificadas pelo grupo de renda (ver Figura 5.10). As projeções que compõem esta sequência temporal são ilustradas na Figura 5.16.

A disposição dos pontos nas projeções é similar ao estudo anterior, e apresenta uma boa separação segundo a classificação: países de renda alta mais a esquerda da projeção, sendo seguidos pelos países de renda média alta, depois pelos países de renda média baixa e, mais a direita, os países de renda baixa. Neste caso, observamos que o principal indicador responsável pela disposição dos pontos é renda per capita, como ilustrado para o ano de 2005 na Figura 5.17.

Na Figura 5.18(a), é possível observar que Trinidade e Tobago, Estônia e Hungria, apesar de terem recebido a classificação de localidades de renda alta pelo *Gapminder*, estão próximas às localidades de renda média alta. Esse comportamento se repete ao longo do tempo, porém nos últimos anos estas localidades apresentam uma tendência de se aproximarem das localidades de renda alta. Esta tendência pode ser observada na ilustração da Figura 5.18(b).

5.4 Considerações finais

Este capítulo apresentou alguns estudos de casos em que a técnica LSP_{FCP} foi utilizada como auxílio à análise exploratória de dados multidimensionais referenciados.

O estudo de caso com os indicadores que compõem o IDH, provenientes do Relatório de Desenvolvimento Humano do PNUD, mostrou que a técnica ajuda na observação do comportamento global dos elementos da base de dados, sendo possível verificar a separação, ao

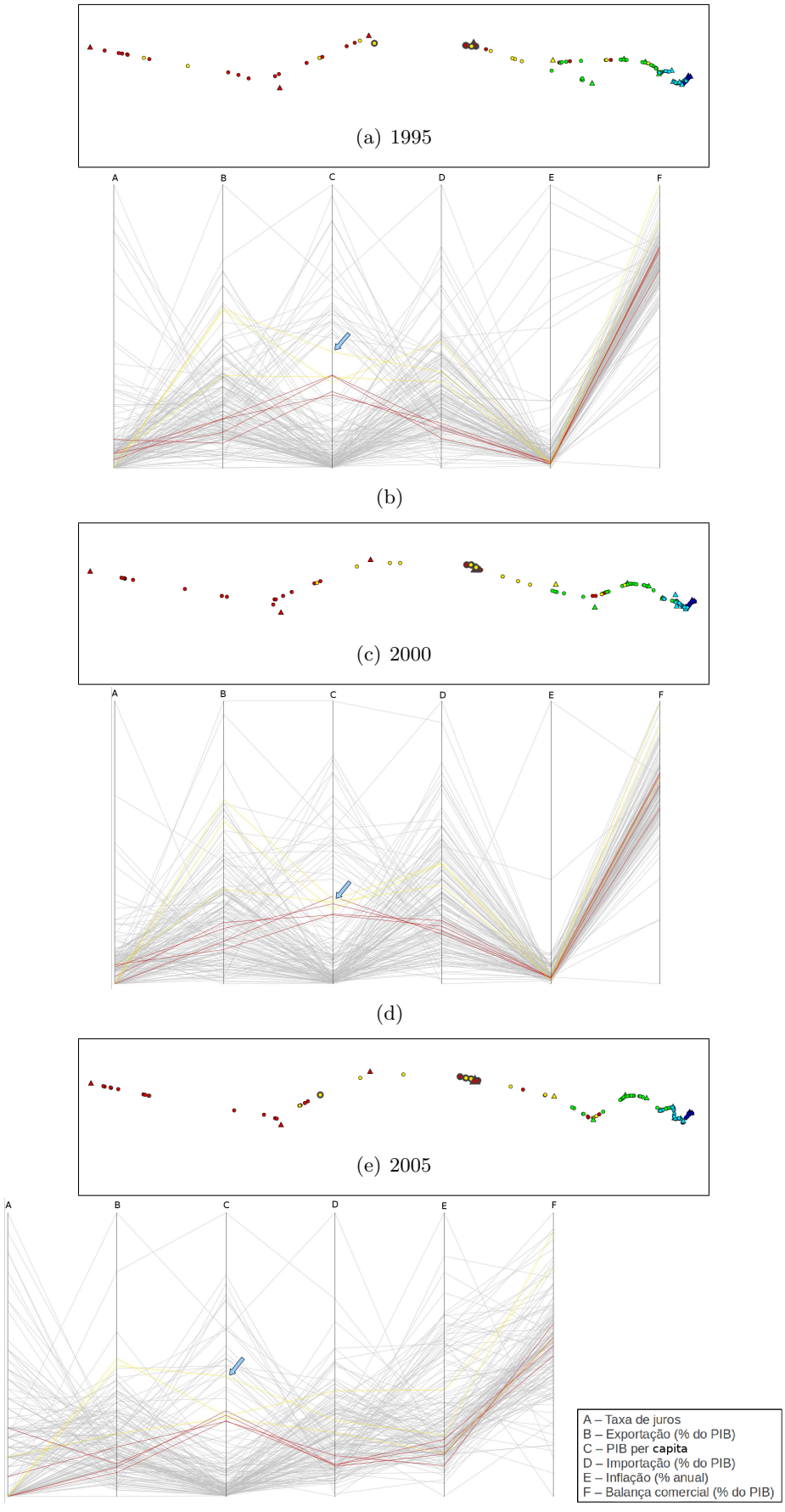
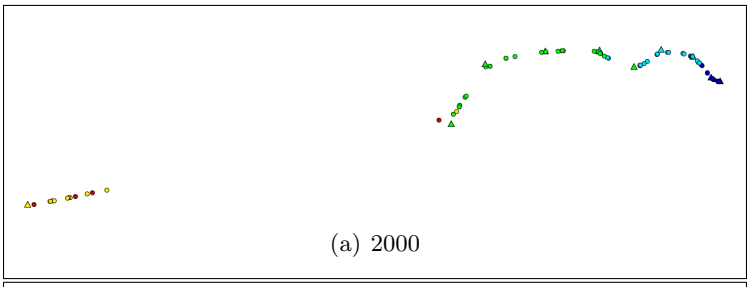
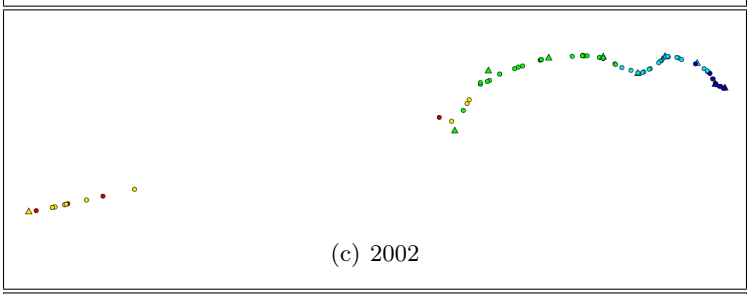


Figura 5.15: Ilustração da variação do PIB per capita de Macau entre 1995 e 2005: entre 1995 e 2000 houve uma diminuição, enquanto que entre 2000 e 2005 houve um aumento.

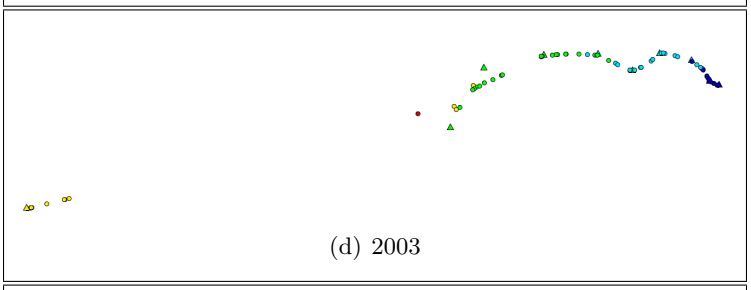


(a) 2000

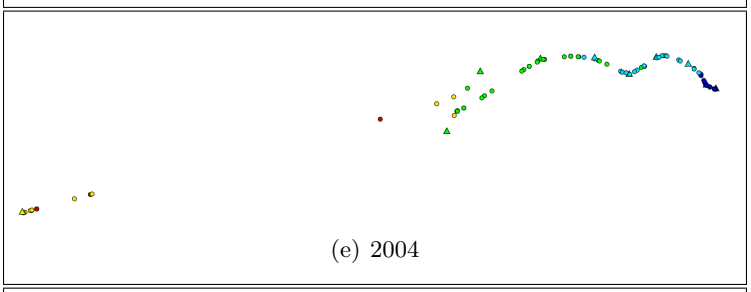
(b) 2001



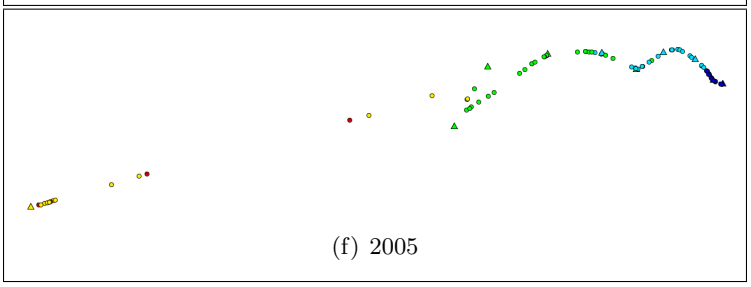
(c) 2002



(d) 2003



(e) 2004



(f) 2005

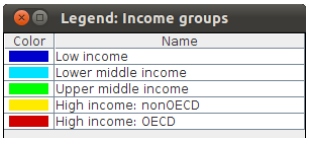
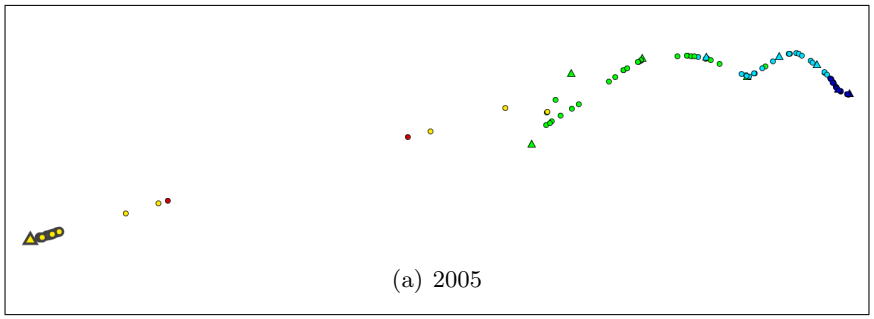
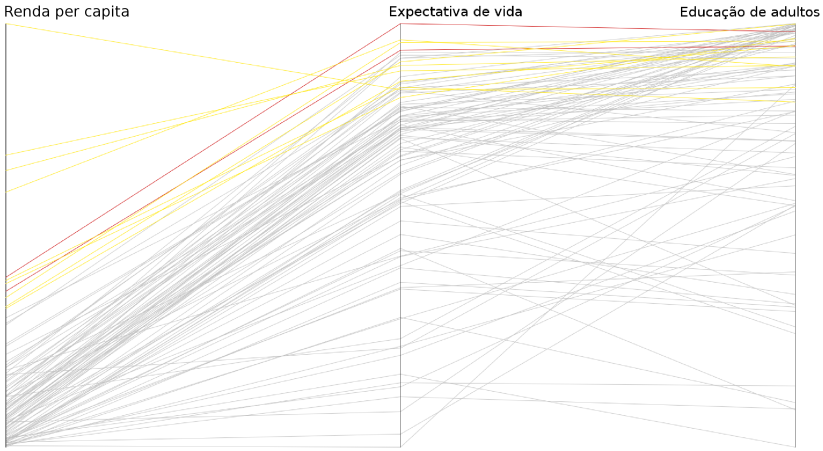


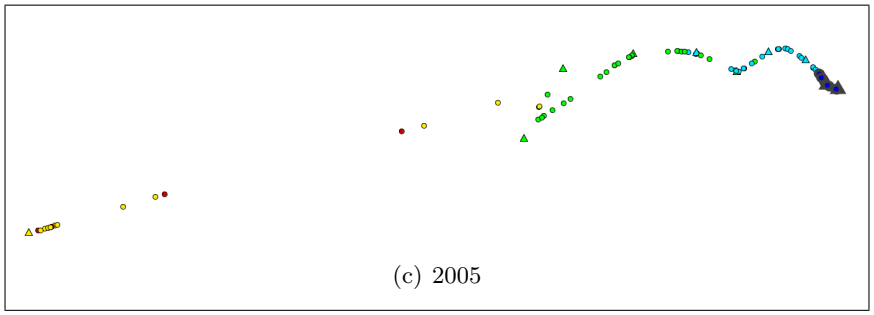
Figura 5.16: Sequência temporal de projeções para os dados que compõem o IDH, encontrados no *Gapminder*, entre os anos de 2000 e 2005.



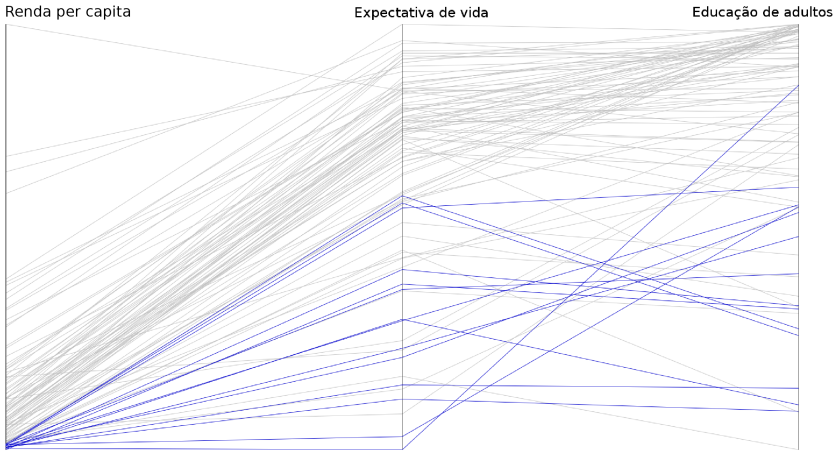
(a) 2005



(b)

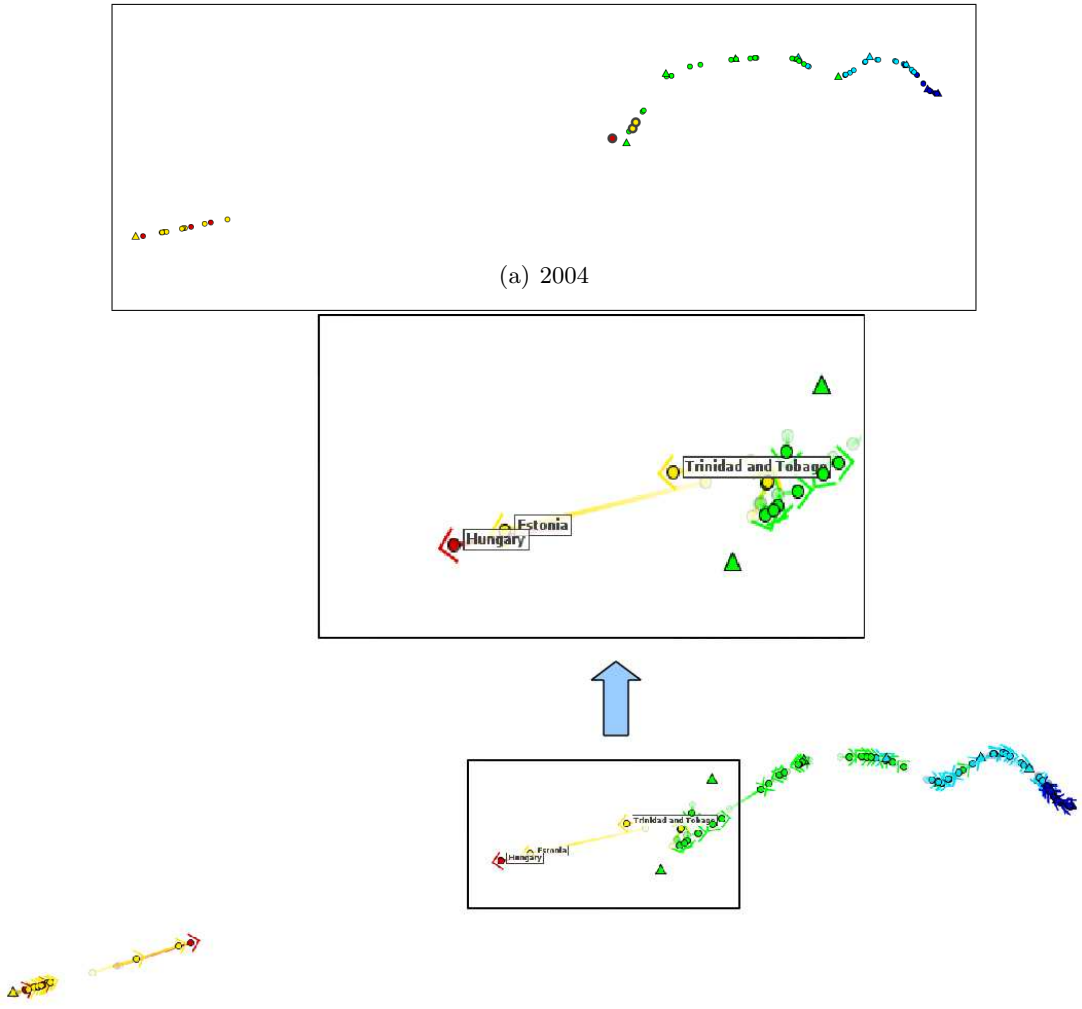


(c) 2005



(d)

Figura 5.17: Ilustração do papel da renda per capita no posicionamento dos pontos em 2005: em (a) os pontos mais à esquerda são selecionados e em (b) observa-se os mesmos destacados, assim como em (c) os pontos mais à direita são selecionados e em (d) observa-se os mesmos destacados.



(a) 2004

(b) 2004 – 2005

Figura 5.18: Entre os anos de 2004 e 2005, Trinidade e Tobago, Estônia e Hungria parecem seguir a tendência de aproximar das localidades de renda alta.

longo do tempo, dos países em dois grupos. Neste estudo, também foi possível observar o comportamento individual de elementos de interesse, em particular de casos atípicos. Por exemplo, a funcionalidade de exibição do rastro nos países Portugal e Coreia do Sul evidenciou a mudança de classes dos mesmos no intervalo de tempo.

Os estudos de caso com bases compostas pelos indicadores presentes no *Gapminder* reforçaram a utilidade da técnica para a observação do comportamento global e do comportamento de elementos de interesse. Além disso, estes estudos apontaram algumas limitações da técnica, como por exemplo a oclusão visual que é fruto da grande quantidade de pontos e arestas em um mesmo espaço de projeção.

O resultado dos estudos mostrou que a utilização da animação, e das funcionalidades de exibição do rastro, apoiam a observação e identificação do comportamento dos elementos ao longo do tempo. Porém, para o entendimento dos fatores que levam a um determinado comportamento é necessário o apoio de uma técnica auxiliar, como Coordenadas Paralelas, que ilustra o papel dos diferentes atributos individualmente.

Capítulo

6

Conclusões

Neste capítulo ressaltamos as principais contribuições deste trabalho para a análise exploratória de bases de dados multidimensionais referenciadas, bem como suas limitações. Também são discutidos desenvolvimentos futuros que podem tratar algumas dessas limitações.

O principal objetivo deste trabalho de mestrado foi propor e desenvolver uma abordagem que, ao gerar uma visualização baseada em projeção, trate de forma diferenciada os atributos temporais ou geográficos. A técnica *Least Square Projection* (LSP) (Paulovich et al., 2008) foi estendida para favorecer a observação do comportamento do conjunto de dados, seguindo uma coerência temporal ou incorporando as referências geográficas durante o processo de criação dos mapas multidimensionais. Esta nova abordagem foi chamada de LSP_{FCP}.

6.1 Contribuições

A técnica LSP_{FCP} permite maior flexibilidade e interferência do usuário na seleção de pontos de controle para o cálculo da LSP. É possível escolher tanto os pontos que serão utilizados como pontos de controle quanto se estes pontos vão utilizar as coordenadas geográficas como coordenadas projetadas, de modo que o usuário pode gerar diferentes projeções e escolher a mais adequada à atividade exploratória desejada.

A ideia de utilizar os mesmos pontos de controle nas projeções da sequência temporal garante uma coerência temporal entre as projeções, evitando distorções como rotação ou mudança de escala. Juntamente com as funcionalidades auxiliares implementadas, a sequência temporal de projeções permitiu a observação do comportamento dos elementos ao longo do tempo.

A utilização das coordenadas geográficas como coordenadas dos pontos de controle tem como objetivo introduzir alguma coerência espacial. Na prática, não identificamos situações em que ela foi realmente útil e por isso essa solução não foi muito explorada nos estudos de caso.

Os estudos de caso ilustram situações em que é possível descobrir a formação de grupos de elementos ao longo do tempo, sendo esta uma característica do comportamento global dos elementos, assim como a observação de elementos com comportamento diferenciado. Os estudos de caso apresentados ilustram potenciais usos do sistema, e as funcionalidades desenvolvidas podem ser úteis para explorar outras bases de dados multidimensionais que apresentam algum tipo de referência (temporal ou geográfica).

6.2 Limitações

A observação do comportamento global dos elementos projetados utilizando a animação requer uma concentração do usuário em uma determinada região do mapa, sendo necessárias várias repetições de modo a contemplar o mapa em sua totalidade. A utilização do rastro em todos os elementos recai no problema da oclusão visual, tornando a interpretação difícil quando todos os *timestamps* são agregados.

Outra limitação encontrada é a dificuldade de entender as mudanças de comportamento utilizando apenas a sequência temporal de projeções e as funcionalidades auxiliares. A interpretação do que causa o movimento dos pontos durante o intervalo de tempo é difícil quando o número de variáveis é grande, sendo necessário recorrer a outras técnicas que ajudem nessa tarefa. Após a identificação de comportamento de elementos de interesse, neste trabalho utilizou-se Coordenadas Paralelas para o entendimento do cenário.

A escolha dos pontos de controle em um *timestamp* não garante que estes sejam os mais representativos nos demais, o que pode causar perda de precisão. Tornar esta escolha dinâmica pode auxiliar o usuário na tarefa de gerar uma boa sequência temporal de projeções. A fixação dos pontos de controle em sua posição inicial, apesar de garantir uma coerência temporal, impede a observação do comportamento temporal destes pontos em particular.

Esta limitação também ocorre quando utiliza-se as coordenadas geográficas como coordenadas cartesianas dos pontos de controle. Uma outra abordagem para associar esta informação ao processo de análise se faz necessária como forma de evitar ou diminuir a perda de precisão das projeções.

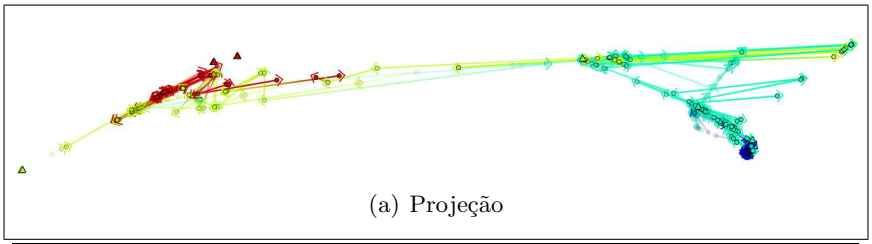
Caso a sequência temporal seja composta por uma quantidade grande de *timestamps*, a execução da abordagem pode ser demorada, devido ao cálculo para resolver o sistema linear para cada *timestamp*.

6.3 Trabalhos futuros

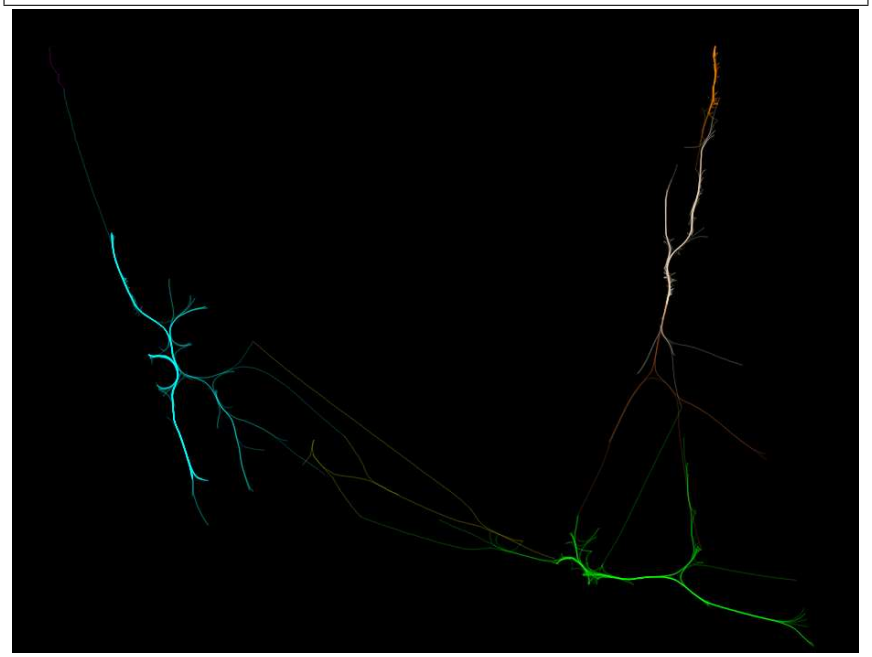
As limitações apresentadas na seção anterior e outras características desejáveis indicam os principais desenvolvimentos futuros, dentre os quais pode-se destacar:

- Visualização estática da sequência temporal de projeções, sem justapor as projeções que a compõem.
- Tornar dinâmica a escolha dos pontos de controle, permitindo que o usuário a realize sobre uma projeção.
- Permitir a utilização de pontos de controle que não estejam na base de dados, servindo como âncoras.
- Flexibilizar o módulo *PExParser*, de forma a contemplar uma quantidade maior de bases de dados iniciais para manipulação e formatação.
- Automatizar a criação dos arquivos utilizando apenas as coordenadas geográficas, como uma opção no módulo *PExParser*.
- Estudar e propor uma nova abordagem para incorporar a informação geográfica ao processo de análise.

É possível considerar alternativas para resolver ou amenizar a oclusão visual causada pela grande quantidade de arestas que unem os pontos na sequência temporal de projeções. Por exemplo, um trabalho que vem sendo desenvolvido no grupo tem como objetivo transformar as retas que representam as arestas em curvas, aplicando uma técnica conhecida como *bundling* (Ersoy et al., 2011). Esta técnica agrupa arestas que são similares com relação à informação de posicionamento, ou seja, origem e destino de cada aresta, como ilustrado na Figura 6.1. Em 6.1(a) tem-se a sequência temporal gerada para o conjuntos dos dados do IDH, apresentados na Seção 5.2, com todas as arestas e em 6.1(b) o resultado do *bundling*. Os valores das coordenadas foram normalizados de forma independente, primeiro apenas os valores de x e em seguida apenas os valores de y, o que altera o espaço para um cubo, causando uma deformação. Essa solução poderá ser adaptada para favorecer a observação do comportamento temporal.



(a) Projeção



(b) *Bundling*

Figura 6.1: Ilustração da aplicação de *bundling* nas arestas da sequência temporal: em (a) tem-se a sequência temporal para o conjuto de dados do IDH e em (b) tem-se o resultado da aplicação da técnica de *bundling* sobre as arestas.

Referências Bibliográficas

Alencar, A. B.; Oliveira, M. C. F. d.; Paulovich, F. V.; Minghim, R.; Andrade, M. G. d. Temporal-PEx: Similarity-based visualization of time series. In: Gonçalves, L.; Wu, S. T., eds. *Proceedings do 20º Simpósio Brasileiro de Computação Gráfica e Processamento de Imagens (SIBGRAPI)*, Porto Alegre: Sociedade Brasileira de Computação, 2007.
Disponível em `http://urlib.net/sid.inpe.br/sibgrapi@80/2007/09.18.22.07` (Acessado em 20 de setembro de 2011)

Andrienko, N.; Andrienko, G. *Exploratory analysis of spatial and temporal data: A systematic approach*. Springer, 2005.

Berkhin, P. *Survey of clustering data mining techniques*. Relatório Técnico, Accrue Software, San Jose, CA, 2002.
Disponível em `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.3739` (Acessado em 20 de setembro de 2011)

Berndt, D. J.; Clifford, J. Using dynamic time warping to find patterns in time series. In: *KDD Workshop*, 1994, p. 359–370.

Brodbeck, D.; Mazza, R.; Lalanne, D. Interactive visualization - a survey. In: Lalanne, D.; Kohlas, J., eds. *Human Machine Interaction*, v. 5440 de *Lecture Notes in Computer Science*, Springer, p. 27–46, 2009.

Card, S. K.; MacKinlay, J. D.; Shneiderman, B.; Card, M. *Readings in information visualization: Using vision to think*. San Francisco - CA, USA: Morgan Kaufmann Publishers Inc., 1999.

Chernoff, H. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, v. 68, n. 342, p. 361–368, 1973.
Disponível em `http://dx.doi.org/10.2307/2284077` (Acessado em 20 de setembro de 2011)

CLEVELAND, W. S. *Visualizing data*. Summit, New Jersey, U.S.A.: Hobart Press, 1993.

CLEVELAND, W. S.; MCGILL, R. Graphical perception and graphical methods for analyzing scientific data. *Science*, v. 229, n. 4716, p. 828–833, 1985.
Disponível em `http://dx.doi.org/10.1126/science.229.4716.828` (Acessado em 20 de setembro de 2011)

DRUCK, S.; CARVALHO, M. S.; CÂMARA, G.; MONTEIRO, A. M. V. *Análise espacial de dados geográficos*. Brasília: EMBRAPA, 2004.

EDSALL, R. M. The parallel coordinate plot in action: design and use for geographic visualization. *Computational Statistics & Data Analysis*, v. 43, n. 4, p. 605 – 619, 2003.

EDWARDES, A.; BURGHARDT, D.; WEIBEL, R. Portrayal and generalisation of point maps for mobile information services. p. 11–30, 2005.
Disponível em `http://dx.doi.org/10.1007/3-540-26982-7_2` (Acessado em 20 de setembro de 2011)

ELER, D.; NAKAZAKI, M.; PAULOVICH, F.; SANTOS, D.; ANDERY, G.; OLIVEIRA, M.; BATISTA NETO, J.; MINGHIM, R. Visual analysis of image collections. *The Visual Computer*, v. 25, p. 923–937, 10.1007/s00371-009-0368-7, 2009a.
Disponível em `http://dx.doi.org/10.1007/s00371-009-0368-7` (Acessado em 20 de setembro de 2011)

ELER, D. M.; PAULOVICH, F. V.; OLIVEIRA, M. C. F. D.; MINGHIM, R. Topic-based coordination for visual analysis of evolving document collections. In: *Proceedings of the 2009 13th International Conference Information Visualisation*, Washington, DC, USA: IEEE Computer Society, 2009b, p. 149–155.
Disponível em `http://portal.acm.org/citation.cfm?id=1588298.1589214` (Acessado em 20 de setembro de 2011)

ERSOY, O.; HURTER, C.; CANTAREIRA, G.; PAULOVICH, F.; TELEA, A. Image-based edge bundles: Simplified visualization of large graphs. *IEEE Transactions on Visualization and Computer Graphics (Visweek)*, (a ser publicado), 2011.

FALOUTSOS, C.; LIN, K.-I. FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. *SIGMOD Rec.*, v. 24, n. 2, p. 163–174, 1995.

FELIZARDO, K.; MARTINS, R. M.; MALDONADO, J. C.; LOPES, A. A.; MINGHIM, R. Content based visual mining of document collections using ontologies. In: *II Workshop on Web and Text Intelligence (WTI)*, São Carlos, SP, Brazil, publicado em CD-ROM, 2009, p. 1–8.

FLOATER, M. S. Parametrization and smooth approximation of surface triangulations. *Comput. Aided Geom. Des.*, v. 14, p. 231–250, 1997.

Disponível em `http://dx.doi.org/10.1016/S0167-8396(96)00031-3` (Acessado em 20 de setembro de 2011)

Giannotti, F.; Pedreschi, D. *Mobility, data mining and privacy: Geographic knowledge discovery.* Springer Publishing Company, Incorporated, 2008.
Disponível em `http://portal.acm.org/citation.cfm?id=1355342` (Acessado em 20 de setembro de 2011)

Grinstein, G.; Trutschl, M.; Cvek, U. High-dimensional visualizations. In: *Proceedings of 7th Data Mining Conference KDD Workshop*, San Francisco, CA, 2001, p. 7 – 19.

Harrie, L.; Stigmar, H.; Koivula, T.; Lehto, L. An algorithm for icon labelling on a real-time map. p. 493–507, 2005.
Disponível em `http://dx.doi.org/10.1007/3-540-26772-7_38` (Acessado em 20 de setembro de 2011)

Hornik, K. The R FAQ. ISBN 3-900051-08-9, 2010.
Disponível em `http://CRAN.R-project.org/doc/FAQ/R-FAQ.html` (Acessado em 20 de setembro de 2011)

Inselberg, A.; Dimsdale, B. Parallel Coordinates: a tool for visualizing multi-dimensional geometry. In: *Proc. First IEEE Conference on Visualization Visualization '90*, 1990, p. 361–378.

Keim, D. A. Designing pixel-oriented visualization techniques: theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, v. 6, n. 1, p. 59–78, 2000.

Keim, D. A.; Kriegel, H.-P. Visualization techniques for mining large databases: a comparison. *IEEE Transactions on Knowledge and Data Engineering*, v. 8, n. 6, p. 923–938, 1996.

Keim, D. A.; Panse, C.; Sips, M. *Exploring geovisualization*, Cap. Information Visualization: Scope, Techniques and Opportunities for Geovisualization , Elsevier, p. 23 – 52, 2005.

Keogh, E.; Lonardi, S.; Ratanamahatana, C. A.; Wei, L.; Lee, S.-H.; Handley, J. Compression-based data mining of sequential data. *Data Mining and Knowledge Discovery*, v. 14, n. 1, p. 99–129, 2007.

Lammarsch, T.; Aigner, W.; Bertone, A.; Gartner, J.; Mayr, E.; Miksch, S.; Smuc, M. Hierarchical temporal patterns and interactive aggregated views for pixel-based visualizations. In: *Proc. 13th Int Information Visualisation Conf*, 2009, p. 44–50.

Lee, M. D.; Reilly, R. E.; Butavicius, M. E. An empirical evaluation of chernoff faces, star glyphs, and spatial visualizations for binary data. In: *Proceedings of the Asia-Pacific symposium on Information visualisation - Volume 24*, Darlinghurst, Australia, Australia:

Australian Computer Society, Inc., 2003, p. 1–10.
Disponível em http://portal.acm.org/citation.cfm?id=857080.857081 (Acessado em 20 de setembro de 2011)

Maciejewski, R.; Rudolph, S.; Hafen, R.; Abusalah, A.; Yakout, M.; Ouzzani, M.; Cleveland, W.; Grannis, S.; Wade, M.; Ebert, D. Understanding syndromic hotspots - a visual analytics approach. *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, p. 35–42, 2008.

Maciejewski, R.; Tyner, B.; Jang, Y.; Zheng, C.; Nehme, R. V.; Glickman, L. T. LAHVA: Linked animal-human health visual analytics. 2009.
Disponível em http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.147.1120 (Acessado em 20 de setembro de 2011)

Mazza, R. *Introduction to information visualization*. Springer, 2009.
Disponível em http://books.google.com.br/books?id=JEhkpAH4BPgC (Acessado em 20 de setembro de 2011)

Minghim, R.; Paulovich, F. V.; Andrade Lopes, A. Content-based text mapping using multi-dimensional projections for exploration of document collections. SPIE, 2006, p. 60600S.
Disponível em http://link.aip.org/link/?PSI/6060/60600S/1 (Acessado em 20 de setembro de 2011)

Nascimento, H. A. D. D.; Ferreira, C. B. R. Visualização de informações – uma abordagem prática. In: *XXIV JAI / XXV Congresso da Sociedade Brasileira da Computação*, 2006, p. 1262 – 1312.

Nguyen, D.-Q.; Schumann, H. Taggram: Exploring geo-data on maps through a tag cloud-based visualization. In: *Information Visualisation (IV), 2010 14th International Conference*, 2010, p. 322 –328.

Oliveira, M. C. F.; Levkowitz, H. From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics*, v. 9, n. 3, p. 378–394, 2003.

Oliveira, M. C. F.; Minghim, R. Uma introdução a visualização computacional. In: *Jornadas de Atualização em Informática*, 1997, p. 85 – 131.

ONS *Vazões médias mensais nos aproveitamentos hidrelétricos – período 1931 a 2005*. Relatório Técnico, 2006.

Paulovich, F. V.; Maki, R.; Oliveira, M.; Colhone, M.; Santos, F.; Migliaccio, V.; Ciancaglini, P.; Perez, K.; Stabeli, R.; Perinoto, Â.; Oliveira, O.; Zucolotto, V. Using multidimensional projection techniques for reaching a high distinguishing ability in biosensing. *Analytical and Bioanalytical Chemistry*, v. 400, p. 1153–1159,

10.1007/s00216-011-4853-2, 2011a.
Disponível em `http://dx.doi.org/10.1007/s00216-011-4853-2` (Acessado em 20 de setembro de 2011)

Paulovich, F. V.; Moraes, M. L.; Maki, R. M.; Ferreira, M.; Oliveira Jr., O. N.; Oliveira, M. C. F. Information visualization techniques for sensing and biosensing. *Analyst*, v. 136, p. 1344–1350, 2011b.
Disponível em `http://dx.doi.org/10.1039/C0AN00822B` (Acessado em 20 de setembro de 2011)

Paulovich, F. V.; Nonato, L.; Minghim, R. Visual mapping of text collections through a fast high precision projection technique. In: *Proc. Tenth International Conference on Information Visualization IV 2006*, 2006, p. 282–290.

Paulovich, F. V.; Nonato, L.; Minghim, R.; Levkowitz, H. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. v. 14, n. 3, p. 564–575, 2008.

Paulovich, F. V.; Oliveira, M.; Minghim, R. The Projection Explorer: a flexible tool for projection-based multidimensional visualization. In: *Proc. XX Brazilian Symposium on Computer Graphics and Image Processing SIBGRAPI 2007*, 2007, p. 27–36.

Rao, R.; Card, S. K. The table lens: merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In: *Conference companion on Human factors in computing systems*, New York, NY, USA: ACM, 1994, p. 222.
Disponível em `http://doi.acm.org/10.1145/259963.260391` (Acessado em 20 de setembro de 2011)

Robertson, G.; Fernandez, R.; Fisher, D.; Lee, B.; Stasko, J. Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics*, v. 14, p. 1325–1332, 2008.
Disponível em `http://portal.acm.org/citation.cfm?id=1477066.1477431` (Acessado em 20 de setembro de 2011)

Schreck, T.; Tekušová, T.; Kohlhammer, J.; Fellner, D. Trajectory-based visual analysis of large financial time series data. *SIGKDD Explor. Newsl.*, v. 9, n. 2, p. 30–37, 2007.

Sen, A. K. A decade of human development. *Journal of Human Development*, , n. 1, p. 17–23, 2000.

Silva, H. A. O.; Andrade, M. G.; Minghim, R. Power Map Explorer: ferramenta para visualização e previsão de vazões. In: *XVII Simposio Brasileiro de Recursos Hídricos*, São Paulo, 2007, p. 1 – 20.

Siqueira, J. R.; Maki, R. M.; Paulovich, F. V.; Werner, C. F.; Poghossian, A.; Oliveira, M. C. F.; Zucolotto, V.; Oliveira, O. N.; Schoning, M. J. Use of information visualization methods eliminating cross talk in multiple sensing units investigated for a light-addressable potentiometric sensor. *Analytical Chemistry*, v. 82, n. 1, p. 61–65, pMID: 20041720, 2010.
Disponível em `http://pubs.acs.org/doi/abs/10.1021/ac9024076` (Acessado em 20 de setembro de 2011)

Sorkine, O.; Cohen-Or, D. Least-square meshes. In: *Proceedings of Shape Modeling International*, IEEE Computer Society Press, 2004, p. 191–199.

Tejada, E.; Minghim, R.; Nonato, L. G. On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization*, v. 2, n. 4, p. 218–231, 2003.

Tiburtino, T.; Oliveira, M. C. F. Visualização de séries temporais georeferenciadas. *13ª Escola de Séries Temporais e Econometria*, 2009.

Tufte, E. R. *The visual display of quantitative information.* 2. ed. Cheshire, CT: Graphics Press, 2001.

Tutte, W. T. How to draw a graph. *Proc Lond Math Soc*, v. 13, p. 743–767, 1963.
Disponível em `http://www.ams.org/mathscinet-getitem?mr=28:1610` (Acessado em 20 de setembro de 2011)

Wegman, E. J. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, v. 85, n. 411, p. 664–675, 1990.
Disponível em `http://dx.doi.org/10.2307/2290001` (Acessado em 20 de setembro de 2011)

Wijk, J. J. V.; Selow, E. R. V. Cluster and calendar based visualization of time series data. In: *INFOVIS '99: Proceedings of the 1999 IEEE Symposium on Information Visualization*, Washington, DC, USA: IEEE Computer Society, 1999, p. 4.

Xavier, S. S. A. Visualizações de informação para suporte a ead – ferramenta VisEd. Monografia de Projeto Final de Curso, Instituto de Informática, Universidade Federal de Goiás, 2004.