Universidade Federal da Bahia
Escola Politécnica

Departamento de Engenharia Elétrica

DIRETRIZES PARA ANÁLISE DE PROJEÇÕES MULTIDIMENSIONAIS E SUAS MÉTRICAS EM DIFERENTES CONFIGURAÇÕES DE BASES DE DADOS

Erick Roseira Pinheiro

TRABALHO DE GRADUAÇÃO

Salvador
30 de julho de 2018

ERICK ROSEIRA PINHEIRO

DIRETRIZES PARA ANÁLISE DE PROJEÇÕES MULTIDIMENSIONAIS E SUAS MÉTRICAS EM DIFERENTES CONFIGURAÇÕES DE BASES DE DADOS

Este Trabalho de Graduação foi apresentado ao Departamento de Engenharia Elétrica da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Bacharel em Engenharia da Computação.

Orientador: Prof. Dr. Danilo Barbosa Coimbra

Salvador
30 de julho de 2018

Sistema de Bibliotecas - UFBA

Pinheiro, Erick Roseira.
DIRETRIZES PARA ANÁLISE DE PROJEÇÕES MULTIDIMENSIONAIS E SUAS MÉTRICAS EM DIFERENTES CONFIGURAÇÕES DE BASES DE DADOS / Erick Roseira Pinheiro – Salvador, 2018.
72p.: il.

Orientador: Prof. Dr. Danilo Barbosa Coimbra.
Trabalho de Conclusão de Curso – Universidade Federal da Bahia, Escola Politécnica , 2018.

1. Projeções Multidimensionais. 2. Avaliação de Projeções Multdimensionais.. I. COIMBRA, Danilo B.. II. Universidade Federal da Bahia. Escola Politécnica . III Diretrizes para análise de projeções multidimensionais e suas métricas em diferentes configurações de bases de dados.

CDD – XXX.XX
CDU – XXX.XX.XXX

TERMO DE APROVAÇÃO

ERICK ROSEIRA PINHEIRO

DIRETRIZES PARA ANÁLISE DE PROJEÇÕES MULTIDIMENSIONAIS E SUAS MÉTRICAS EM DIFERENTES CONFIGURAÇÕES DE BASES DE DADOS

Este Trabalho de Graduação foi julgado adequado à obtenção do título de Bacharel em Engenharia da Computação e aprovado em sua forma final pelo Departamento de Engenharia Elétrica da Universidade Federal da Bahia.

Salvador, 30 de Julho de 2018

Prof. Dr. Danilo Barbosa Coimbra
Universidade Federal da Bahia

Prof. Dr. Maycon Leone Maciel Peixoto
Universidade Federal da Bahia

Profa. Dra. Tatiane Nogueira Rios
Universidade Federal da Bahia

Dedico este trabalho a Deus e a meus pais, Francisco e Jocely, que sempre me incentivaram nesta caminhada e na vida.

AGRADECIMENTOS

Agradeço primeiramente a Deus, meu amigo, por permitir chegar onde estou. A meus pais, Francisco e Jocely, pelo apoio, pelos ensinamentos, por sempre me incentivarem a lutar pelos meus sonhos e por todo carinho e amor dado. A meu irmão Felipe, aquele que por ser irmão mais velho sempre me espelhei, obrigado por todo o apoio, pelos conselhos e carinho. A todos os amigos, que me ajudaram nos momentos difíceis, nestes seis anos de caminhada. Ao meu orientador Prof. Dr. Danilo Barbosa Coimbra, pela humildade, maravilhosa orientação e paciência para comigo, obrigado por estes doze meses. Ao professor Dr. Maycon Peixoto pela disposição e vontade em ajudar-me no desenvolvimento deste trabalho. Ao professor Dr. Tácito Trindade, que juntamente com meu orientador Danilo e o professor Maycon Peixoto, contribuiu para o resultado final deste trabalho. A banca de professores presente na defesa deste trabalho pelo tempo e comentários acerca do mesmo. A todos os demais que de alguma forma contribuíram e fazem parte desta conquista. Obrigado a todos vocês.

The ultimate tragedy is not the oppression and cruelty by the bad people but the silence over that by the good people.

— KING, MARTIN LUTHER JR.

RESUMO

Dados de alta dimensionalidade são uma tendência no mundo moderno, com grande ocorrência nas mais diversas áreas da indústria e ciência. Devido a isto, inúmeras técnicas de projeções multidimensionais têm sido desenvolvidas como ferramenta de visualização e análise destes dados. Para assegurar a qualidade da técnica de projeção é necessário aplicar métricas de avaliação no mapeamento gerado além de utilizar diferentes configurações de bases de dados para confirmar a acurácia do método de projeção. No entanto, nem sempre é transparente para o usuário qual a influência da dimensionalidade, número de instâncias ou número de grupos da base de dados (a partir de agora nomeados fatores) considerando diferentes projeções e métricas de avaliação.

Este Trabalho de Conclusão de Curso objetiva realizar, por meio do método do fatorial 2^k, um estudo dos efeitos de cada um destes fatores quantificando o grau de influência dos mesmos nas variáveis de respostas escolhidas. A partir dos resultados adquiridos são traçadas diretrizes para o processo de análise de projeções multidimensionais e suas métricas de avaliação.

Palavras-chave: Projeções Multidimensionais, avaliação de Projeções Multidimensionais.

ABSTRACT

High-dimensional data are a trend in the modern world with prevalent occurrence in many areas of industry and science. As a result, several multidimensional projection techniques have been developed as a visualization and analysis tool. In order to ensure the projection quality, is necessary assess the lower-dimensional embedding, by using different data sets configurations as input. However, is not always clear to the user how the number of dimensions, instances or clusters (from now on called factors) can affect the projection mapping and its quality regarding different projection techniques and assessment metrics.

This work aims to perform an analysis about the effects of these factors using a two-level full factorial design (2^k) by quantifying the influence of each factor in the response variables. From the results guidelines are presented for the process of analysis of multi-dimensional projection and its assessment metrics.

Keywords: Multidimensional Projection, assessment of multidimensional projection.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

A	Autoencoders
CCA	Curvilinear Component Analysis
CPs	Componentes Principais
DR	Dimensionality Reduction
DM	Diffusion Maps
FA	Factor Analysis
Hybrid	Hybrid Model
ICA	Independent Component Analysis
ICs	Independent Components
IDMAP	Interactive Document Map
Isomap	Isomap
KPCA	Kernel Principal Component Analysis
LAMP	Local Affine Multidimensional Projection
LCMC	Local Continuity Meta-criterion
LE	Laplacian Eigenmaps
LLE	Locally Linear Embedding
LLC	Locally Linear Coordination
LLP	Linearity Preserving Projection
LLTSA	Linear Local Tangent Space Alignment
LTSA	Local Tangent Space Alignment
LoCH	Local Convex Hull
LSP	Least Square Projection

MDS Multidimensional Scaling

MRRE Mean Relative Rank Error

MVU Maximum Variance Unfolding

NPE Neighborhood Preserving Embedding

Optidigits Optical Recognition of Handwritten Digits

PCA Principal Component Analysis

PEx Projection Explorer

PLMP Part-Linear Multidimensional Projection

RMSE Root Mean Square Error

S Sammon Mapping

SMS Sammon's Mapping Speeding-up

SNE Stochastic Neighbor Embedding

SNP Smooth Neighborhood Preservation

SPE Stochastic Proximity Embedding

SPECTRAL Spectral Signatures

t-SNE t-Distributed Stochastic Neighbor Embedding

UCM UC Merced Land Use Dataset

WLD Weber Local Descriptors

Capítulo 1

INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Enquanto as diversas áreas da ciência e campos de atuação tem sido inundados pela excessiva quantidade de dados, a questão de como extrair, de forma eficaz, informações significativas ainda permanece.

Dentre esta variedade e tamanho de dados, há aqueles conhecidos como dados multidimensionais. Tais dados são peculiares devido à sua natureza intrínseca: cada instância possui inúmeras propriedades que podem ser mensuradas. Sendo assim, cada valor de medição resultante, para todos os dados, são normalmente chamados variáveis, dimensões ou atributos (SILVA; RAUBER; TELEA, 2016).

Armazenar dados multidimensionais é uma tarefa relativamente fácil. Porém, a maior dificuldade está em analisá-los. De fato, a preocupação não está somente na quantidade de informação a ser processada, mas principalmente na expansão dimensional das observações. Assim, uma medida de complexidade dos dados é o número de atributos associados a cada instância de dados (PAULOVICH et al., 2008).

Por exemplo, considere duas bases hipotéticas, B_1 e B_2, provindas de uma instuição de ensino superior. A base B_1 possui exatamente 1200 observações contendo um único atributo (dimensão): a idade de seus 1200 alunos matriculados naquele semestre. A base B_2, no entanto, possui 100 observações, que representam os professores de uma instituição, cada uma contendo 12 atributos, que são o número de dias trabalhados (assiduidade) em cada mês de um determinado ano. É possível perceber que o número de medições é exatamente o mesmo, 1200, porém a base de dados B_1 é muito mais simples de ser explorada devido a sua baixa dimensionalidade.

Neste caso, pode-se citar algumas abordagens. Por exemplo, a utilização de uma representação tabular onde as linhas representam cada um dos alunos e a única coluna a idade dos mesmos é uma opção. Uma outra alternativa é o gráfico de barras, onde no eixo das abscissas (X) seriam mapeados os 1200 alunos e no eixo das ordendas (Y) as suas respectivas idades. Por outro lado, entender a base de dados B_2 não é tão trivial. É necessário, por exemplo, analisar quais tipos de padrões existem nesta base, sejam

eles grupos, correlações, *outliers*, entre outros, além da difícil tarefa de representar tal informação.

Chega-se então à uma importante questão: como representar estes dados a fim de examiná-los e extrair importantes conclusões? A maneira mais indicada é através de representações visuais que possuem uma vantagem muito maior quando comparadas as técnicas textuais. De acordo com Mazza (2009), propriedades visuais como cor, tamanho, proximidade e movimento são processadas pela percepção visual humana de uma forma rápida e eficiente. Neste cenário se encontram os métodos de análise visual de dados multidimensionais.

Estas técnicas se caracterizam por mapearem visualmente o conjunto de dados de alta dimensionalidade. Algumas das mais famosas técnicas são: Table Lens (RAO; CARD, 1994), técnicas iconográficas (CHERNOFF, 1973), Coordenadas Paralelas (INSELBERG; DIMSDALE, 1990), dentre outras. Estes métodos, em sua maioria, representam a informação por meio de variados elementos gráficos, como pontos, segmentos de retas, ícones, etc.

Mais recentemente, novas metodologias de visualização vem surgindo, sendo estas, alternativas aos métodos convencionais. Denominadas projeções multidimensionais, tais técnicas lidam bem com o número de observações e dimensões, são intuitivas e podem ser usadas com esforço mínimo (SILVA; RAUBER; TELEA, 2016).

As projeções multidimensionais geralmente projetam uma nuvem de pontos em um espaço bidimensional (usualmente), agrupados em uma estrutura (*layout*) específica, a fim de refletir a similaridade das amostras. Desta forma é possível ter uma visão generalista ou global, a qual aliada com a interatividade proporcionada por estas técnicas, proporciona uma análise mais completa ao focar em grupos de interesse.

A popularidade das projeções multidimensionais vem aumentando ao longo do tempo, visto que apresentam eficiência. Porém, diferentes configurações de bases de dados, aliadas com o surgimento cada vez mais repentino de novas técnicas, tem dificultado o processo de decisão sobre qual escolha de projeção é mais assertiva. Neste momento então, é necessário um estudo mais preciso a respeito das vantagens e desvantagens de cada metodologia tendo como base o contexto dos dados a serem analisados.

1.2 OBJETIVOS

O principal objetivo deste trabalho é apresentar diretrizes no processo de execução e avaliação das projeções multidimensionais e suas métricas de avaliação, baseando-se nas características de diferentes bases de dados.

1.3 ORGANIZAÇÃO

O presente documento é, em termos de estrutura, organizado da seguinte maneira:

- No capítulo 2 é apresentada uma Revisão Sistemática da Literatura a respeito das métricas de avaliação de Projeções Multidimensionais.
- No capítulo 3 são apresentadas técnicas de visualização de dados multidimensionais, abordando os métodos mais tradicionais.

- No capítulo 4, são introduzidas as técnicas de projeção multidimensional, foco desta monografia, além de introduzir as métricas de avaliação utilizadas neste trabalho.
- No capítulo 5 são apresentados os resultados dos experimentos realizados em um conjunto de 8 bases sintéticas de diferentes configurações, utilizando as projeções multidimensionais Local Affine Multidimensional Projection (LAMP) e Local Convex Hull (LoCH), onde analisa-se o efeito do número de instâncias, dimensões e grupos nas projeções e métricas utilizadas.
- No Capítulo 6, são apresentadas as conclusões deste trabalho, descrevendo as principais contribuições para o processo de análise e avaliação de projeções multidimensionais e suas métricas, as limitações encontradas e possíveis trabalhos futuros.

Capítulo 2

REVISÃO SISTEMÁTICA DA LITERATURA

2.1 CONSIDERAÇÕES INICIAIS

O presente capítulo tem por intuito realizar um estudo sobre como o tema de pesquisa tem sido desenvolvido na literatura, quais as fronteiras do conhecimento e quais informações têm sido descobertas a respeito da área, isto é, realizar uma análise do estado da arte no assunto em questão. Para concretizar tal objetivo a Revisão Sistemática da Literatura será utilizada como ferramenta auxiliadora.

A Revisão Sistemática é definida como um meio de identificação, avaliação e interpretação de toda pesquisa relevante disponível para um problema de pesquisa particular, tópico de uma área ou fenômeno de interesse (KITCHENHAM, 2004). Enriquecendo ainda mais sua conceituação, Petticrew (2001) define a Revisão Sistemática como uma técnica eficiente para testar hipóteses, resumir os resultados de estudos existentes e avaliar a consistência entre estudos prévios. Percebe-se que estas duas definições podem ser expressas sucintamente em reunir todos os achados relevantes a um problema de pesquisa e analisá-los de forma crítica e sistemática.

A revisão sistemática apresenta características divergentes em relação a revisão tradicional, comumente empregada, possibilitando assim sua escolha em detrimento da outra. A revisão tradicional tende a ser principalmente descritiva, sem envolver uma busca sistemática da literatura, frequentemente focando-se em um subconjunto de estudos de determinada área, que é escolhido baseado na disponibilidade dos mesmos ou por livre arbítrio do autor (UMAN, 2011). Isto revela uma característica tendenciosa que muitas vezes tem o intuito de corroborar uma visão pré-concebida ou uma opinião pessoal (PAI et al., 2004). Já a revisão sistemática tipicamente envolve um planejamento detalhado e abrangente e uma estratégia de busca visando diminuir a tendenciosidade a partir da análise, avaliação e sintetização de todos os estudos relevantes sobre um assunto específico (UMAN, 2011).

O processo de revisão sistemática possui muitas variantes com relação ao conjunto e ordem dos passos a serem executados diferindo-se de autor para autor. Os estágios que

serão descritos a seguir e utilizados na presente revisão são baseados no modelo definido por (KITCHENHAM, 2004).

A revisão sistemática é constituída de três etapas fundamentais: planejamento da revisão, condução da revisão e análise da revisão, representadas na Tabela 2.1 a seguir:

Tabela 2.1: Fases da Revisão Sistemática.

Fase	**Estágios**
Planejamento da revisão	1. Identificação do problema de pesquisa 2. Definição de um protocolo de revisão
Condução da revisão	1. Seleção de estudos 2. Avaliação da qualidade dos estudos 3. Extração de dados
Análise da Revisão	1. Sintetização dos dados

Percebe-se que a fase de planejamento da revisão está relacionada a identificar a necessidade de uma revisão devido a um problema de estudo. Mediante a definição da questão da pesquisa uma estratégia de revisão é formulada. A próxima fase é conduzir a revisão por meio da minuciosa seleção de estudos, sempre avaliando a qualidade dos mesmos, ou seja, se estes são relevantes o suficiente à pergunta principal que norteia todo o estudo. Por fim, na análise da revisão ocorre a sintetização dos dados onde será possível aplicar a análise das informações obtidas.

Dada esta breve introdução, é realizada a seguir uma revisão sistemática da literatura concernente ao tema de avaliação de técnicas de projeções multidimensionais. Todo o processo será descrito de forma detalhada nas próximas seções.

2.2 PLANEJAMENTO DA REVISÃO

2.2.1 Identificação do Problema de Pesquisa

A necessidade de análise de dados de várias dimensões tem sido requisitada bastante nos últimos anos, o que tem fomentado grandemente a utilização e criação de técnicas de projeção multidimensional que mapeiam dados n-dimensionais, $n > 3$, para um espaço p-dimensional, com p=1,2,3, sendo p bidimensional mais comumente. No entanto, o grande desafio da área está em garantir a qualidade e corretude destas projeções, visto que intrinsecamente existem erros associados ao processo. Mediante este contexto é definida

a pergunta principal da pesquisa como:

Quais as abordagens/técnicas comumente utilizadas para avaliar a qualidade de projeções multidimensionais?

A pergunta acima é a formulação em si do problema de pesquisa que será respondida ao final desta revisão com os trabalhos mais relevantes relacionados ao tema. A fim de detalhar/especializar ainda mais a pergunta principal, enriquecendo o conjunto de estudos coletados, perguntas secundárias são definidas a seguir:

1. **PS.1** - Qual o domínio dos dados?
2. **PS.2** - Quem é o pesquisador principal da área?
3. **PS.3** - Existe alguma métrica mais relevante que as demais?
4. **PS.4** - Qual a frequência de estudos publicados por ano?

2.3 DEFINIÇÃO DE UM PROTOCOLO DE REVISÃO

Após a definição da pergunta principal e das perguntas secundárias, um protocolo de revisão foi adotado seguindo as subtarefas definidas nas subseções 2.3.1, 2.3.2 e 2.3.3.

2.3.1 Seleção da Fonte de Busca

Decidiu-se que a base de dados utilizada para busca dos artigos/estudos relacionados ao tema, seria unicamente o repositório online Scopus (https://www.scopus.com/home.url).

O Scopus é uma base de dados de resumos e citações de artigos publicados em livros ou revistas/jornais acadêmicos, sendo assim amplamente utilizado para fins de consulta e trabalhos, como este, pela comunidade acadêmica. Um dos fatores da escolha do Scopus como fonte de busca é alta disponibilidade de estudos relevantes e mais relacionados a determinado tema a partir de uma string de busca específica.

2.3.2 Seleção do Idioma

Definiu-se também uma padronização no quesito idioma dos artigos pesquisados, adotando-se como aptos apenas artigos escritos em língua inglesa, uma vez que a maioria dos trabalhos encontram-se na referida língua. Desta forma os demais idiomas foram desconsiderados.

2.3.3 Definição das Palavras-Chave e Strings de Busca

Analisando a pergunta principal foi possível definir algumas palavras-chave relacionadas como:

- dimensionality reduction
- multidimensional projection

- evaluation
- assessment
- analysis

A *string* de busca final após combinação das palavras-chave tomou a seguinte forma:

("dimensionality reduction"OR "multidimensional projection") AND (evaluation OR assessment OR analysis)

2.4 CONDUÇÃO DA REVISÃO

2.4.1 Seleção dos Estudos

A condução da revisão é a fase da revisão sistemática na qual realiza-se a seleção dos artigos mais relevantes ao problema de pesquisa, utilizando critérios de inclusão e exclusão.

Para a seleção dos artigos/estudos fez-se a pesquisa utilizando a *string* de busca *("dimensionality reduction"OR "multidimensional projection") AND (evaluation OR assessment OR analysis)* na base de dados Scopus, como é demonstrado na Figura 2.1. Como resultado, foram encontrados exatamente 8792 artigos.

2.4.2 Avaliação da Qualidade dos Estudos

Nesta primeira vez, a base de dados retornou todos os artigos que contivessem a string ou no título (*article title*) ou no resumo (*abstract*) ou nas palavras-chave (*keywords*), como pode-se verificar novamente na Figura 2.1. Isto acabou gerando uma grande quantidade de artigos sendo em sua maioria pouco relevantes para o problema de pesquisa.

Como ilustração, um dos artigos encontrados tem por título *"Prognosis cancer prediction model using deep belief network approach"* ou em tradução livre *"Modelo de previsão de câncer utilizando a abordagem de rede de crença profunda"*. Percebe-se que o título não tem relação nenhuma com o tema de pesquisa. No entanto, por conter em seu resumo, a citação da técnica de redução de dimensionalidade Principal Component Analysis (PCA) como ferramenta para análise do prognóstico de câncer, o artigo foi retornado como resultado, uma vez que em seu *abstract* há o termo *analysis*, sendo este parte integrante da *string* de busca. Adicionalmente, ao ler o resumo do artigo ficou claro, mais detalhadamente, que o trabalho era irrelevante, visto que somente aplicava uma técnica de redução de dimensionalidade, mas não tinha o intuito de abordar o problema de estudo. Esta situação ocorreu em diversos documentos.

A fim de trazer um conjunto de estudos mais conciso e mais relevante, a busca foi novamente realizada considerando a *string* apenas nos títulos dos trabalhos. Como resultado foram obtidos 320 artigos relacionados.

Desta vez, foi realizada a leitura dos títulos e resumos de todos os artigos a fim de selecionar os trabalhos mais relacionados à pesquisa.

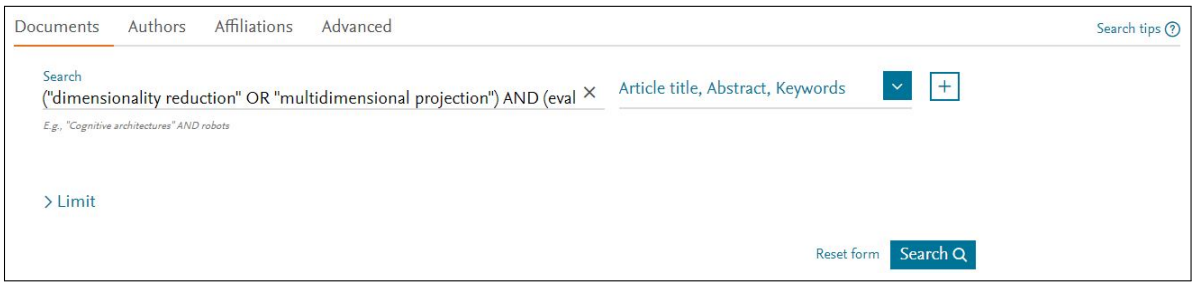


Figura 2.1: Exemplo de busca na base de dados Scopus.

2.4.3 Extração dos Dados

1. **Critérios de Exclusão e Inclusão**

Os artigos selecionados devem relatar o contexto da utilização das projeções multidimensionais, bem como analisá-las utilizando uma ou mais métricas/técnicas de avaliação de qualidade, com o intuito de investigar quão bom/confiável é o método. Artigos que apenas relatam a utilização de alguma projeção multidimensional para analisar uma determinada questão ou chegar a alguma conclusão de pesquisa, não atendem aos objetivos desta revisão sistemática e são descartados. Em adição, artigos duplicados ou que não abordavam as métricas de avaliação de forma detalhada foram também excluídos.

1. **Dados de Extração dos Artigos Selecionados**

Com o intuito de auxiliar o processo de extração de dados e organizá-los de uma forma relevante, definiu-se o seguinte conjunto de informações a serem extraídas de cada artigo selecionado:

- Título
- Autor
- Contexto
- Objetivo
- Técnicas de Redução de Dimensionalidade Utilizadas
- Domínio de Dados
- Métricas de Avaliação
- Objetivo das Métricas

A sintetização e análise dos dados obtidos, a partir da leitura dos artigos, serão apresentados na seção subsequente.

2.5 RESULTADOS

Aplicando a *string* de busca apenas nos títulos dos artigos, foram retornados 320 documentos, como mencionado anteriormente.

Analisando os resultados deste primeiro conjunto de trabalhos, em termos de qual país tem maior contribuição científica no assunto de projeção multidimensional, percebe-se que a China, os Estados Unidos e a Índia figuram como os países com maior número de publicações. A Figura 2.2 apresenta a classificação dos 10 primeiros países com relação ao número de publicações.

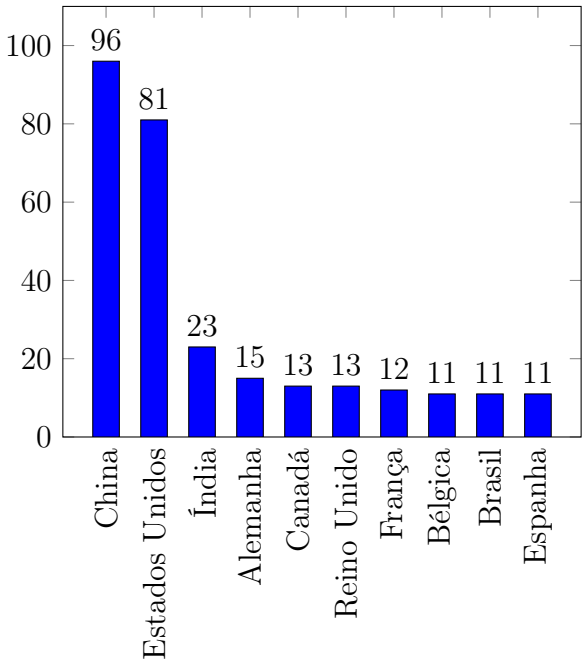


Figura 2.2: Distribuição do número de artigos publicados por país.

Uma outra conclusão a se observar é que o aumento da análise de dados multidimensionais tem tornado-se frequente, especialmente em mineração de dados. Pode-se verificar que nos últimos anos houve um crescimento importante no que tange ao número de trabalhos desenvolvidos na área. Isto é explicitado na Figura 2.3, onde é mostrado o número de artigos publicados em um período de 20 anos, de 1998 até a presente data (01/2018). Vale ressaltar que a aparente queda no ano 2018, é explicada por esta análise considerar unicamente o primeiro mês de 2018, onde houve produção de apenas 2 artigos.

Com o intuito de tornar a revisão sistemática mais precisa e selecionar apenas os trabalhos mais relevantes foram aplicados os critérios de inclusão e exclusão, definidos anteriormente, neste conjunto de 320 artigos em duas fases descritas a seguir.

Primeiramente analisou-se cada um dos artigos a partir da leitura de seus títulos e resumos. Desta forma foram selecionados 25 artigos para leitura completa. Isto revela que a maioria dos artigos retornados pela *string* de busca estavam relacionados à projeção multidimensional de uma forma geral ou apenas utilizavam-se dela para realizar um

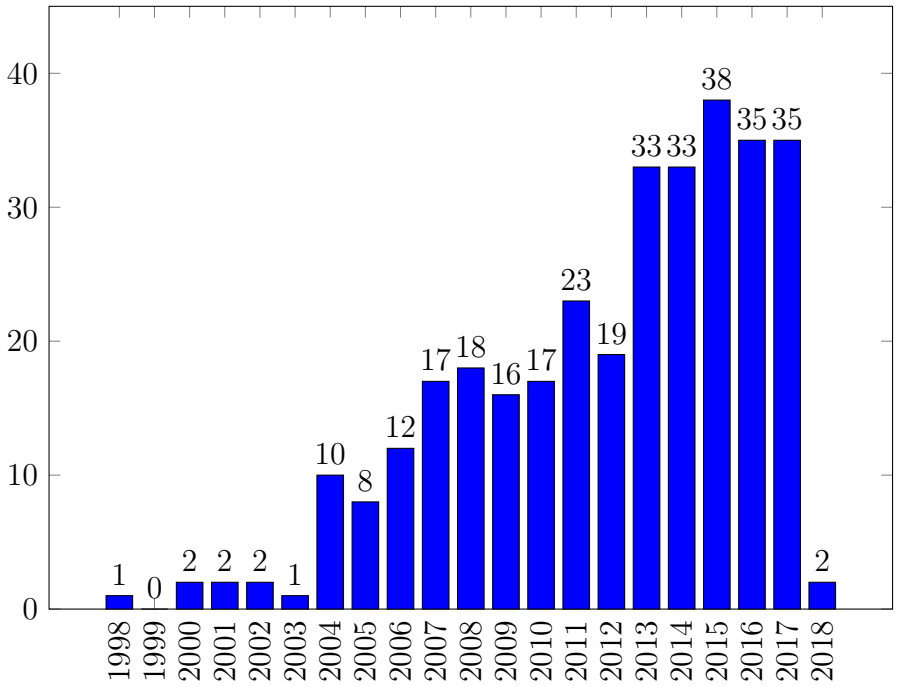


Figura 2.3: Número de artigos publicados por ano.

estudo, mas apenas uma pequena parcela abordava diretamente a questão da avaliação de técnicas de redução de dimensionalidade.

Na segunda fase, que compreendeu a leitura detalhada de cada artigo, chegou-se, após a remoção de artigos duplicados e que não abordavam minuciosamente a técnica de avaliação das projeções multidimensionais, chegou-se ao número final de 11 artigos altamente correlacionados à pesquisa.

Apresenta-se então, nesta fase, informações quantitativas sobre os 11 artigos. A Figura 2.4 mostra a quantidade de publicações por país dos autores, revelando que a Alemanha e o Brasil lideram as pesquisas relacionadas ao uso de métricas de avaliação dos mapeamentos multidimensionais. Como alguns artigos foram produzidos por mais de um autor de nacionalidades distintas, cada ocorrência foi computada como uma unidade no gráfico. Outra informação interessante é mostrada na Tabela 2.2. Nela há a frequência de publicação dos artigos ao longo dos anos. Observa-se que o ano de 2015 foi o mais produtivo para área, tendo um total de 4 publicações, enfatizando a crescente produção científica no tema abordado, respondendo assim a pergunta secundária **PS.4**.

Os autores mais influentes, tomando como base o número de publicações na área de estudo, são demonstrados na Tabela 2.2. Como um autor pode ter publicado mais de um artigo, cada publicação equivale a uma unidade na tabela. Percebe-se que a autora com maior número de artigos publicados é a Rosane Minghim, com um total de 3 publicações, respondendo assim a pergunta secundária **PS.2**. Ela é professora da Universidade Federal de São Paulo (USP) e atua na área de pesquisa de visualização de informações e análise visual. Na Tabela 2.3 é mostrada a relação dos 28 autores dos artigos e seus respectivos

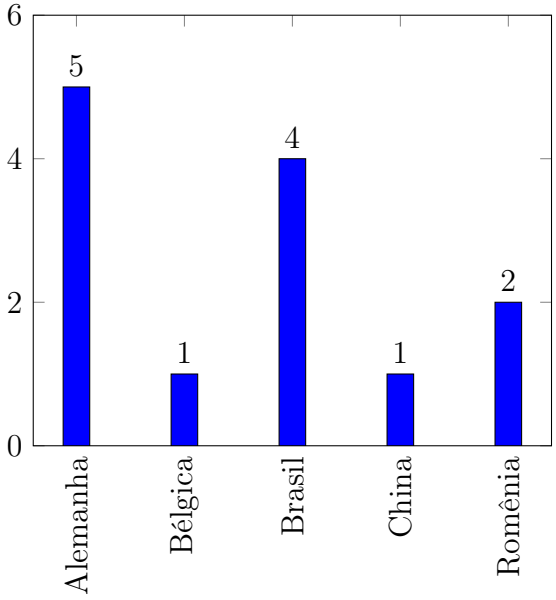


Figura 2.4: Número de artigos publicados por país dos autores.

Tabela 2.2: Número de publicações por ano.

Ano	Frequência
2017	1
2016	2
2015	4
2014	1
2013	1
2011	1
2009	1

números de publicações.

Tabela 2.3: Quantidade de Publicações por Autor.

Autor	Quantidade de Publicações
Rosane Minghim	3
Andreea Gripars	2
Bastian Rieck	2
Daniela Faur	2
Heick Liette	2
Mihai Datcu	2
Alneu de Andrade Lopes	1
Alexandru C. Telea	1
Danilo Barbosa Coimbra	1
Danilo Medeiros Eler	1
Deyu Meng	1
Fernando V. Paulovich	1
Gerhard Rigoll	1
Haim Levkowitz	1
Jaqueline Batista Martins	1
John Aldo Lee	1
Luis Gustavo Nonato	1
Maria Cristina F. Oliveira	1
Michel Verleysen	1
Mihai P. Datcu	1
Mohammadreza Babaee	1
Paulo Pagliosa	1
Priscila Alves Macanha	1
Rafael Messias Martins	1
Robson Motta	1
Rogério Eduardo Garcia	1
Yee Leung	1
Zongben Xu	1

É também de interesse desta revisão sistemática identificar quais são as métricas e suas frequências de utilização. A Figura 2.5 traz esta informação relacionando a quantidade de utilização de cada técnica nos artigos selecionados. Como uma determinada técnica pode ter sido utilizada mais de uma vez em diferentes artigos, cada ocorrência foi computada como uma unidade. Ao analisar tal figura é possível perceber que as técnicas preservação de vizinhança (Neighborhood Preservation) e stress são as métricas mais relevantes, respondendo assim a pergunta secundária **PS.3**.

Com relação às bases de dados utilizadas em cada um dos artigos, foram utilizadas, no total, 31 bases, sendo 8 bases sintéticas e 23 bases reais, algumas provenientes de

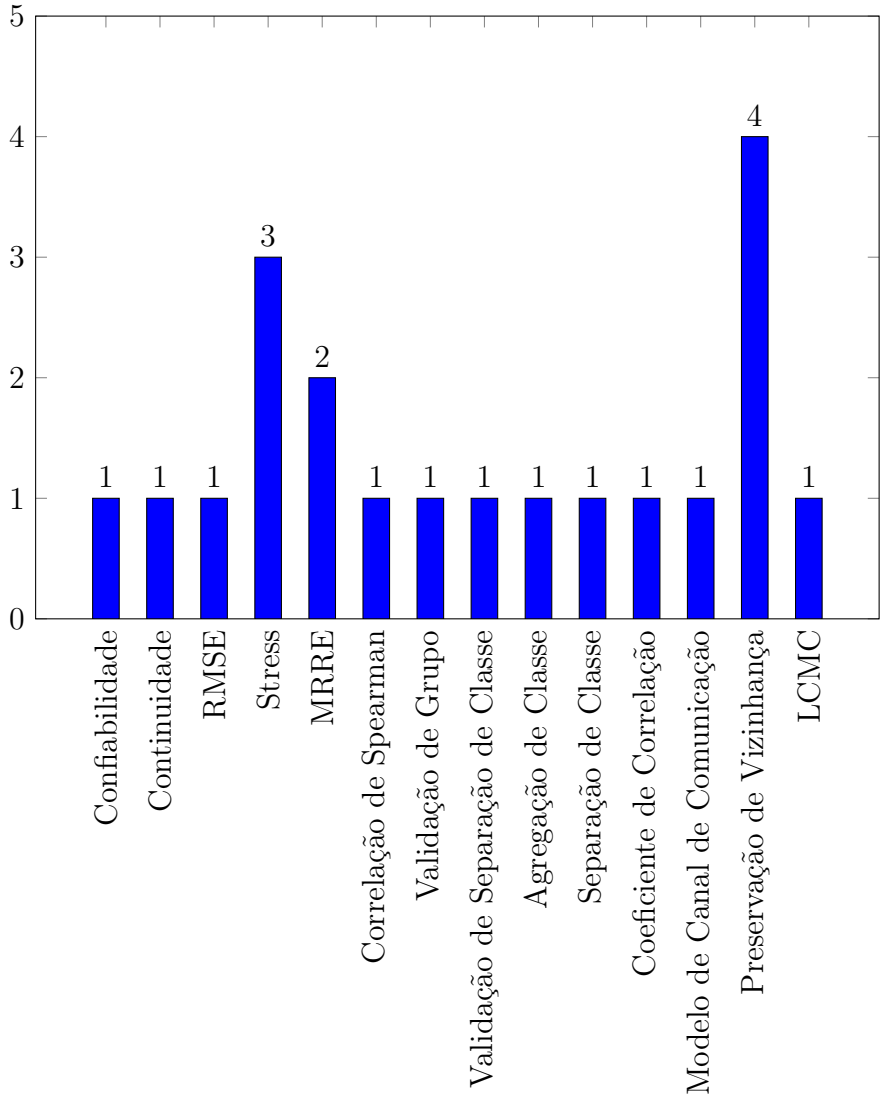


Figura 2.5: Frequência de utilização de cada técnica entre todos os artigos.

sensoriamento remoto, outras formadas de feeds RSS, etc. Uma descrição mais detalhada é apresentada na Tabela 2.4. Com as informações aqui providas e as adicionais na coluna **Domínio(s) de Dados** da Tabela 2.4, é possível responder a pergunta secundária **PS.1.** A Tabela 2.5 traz a referência do artigo citado na Tabela 2.4 com base no seu identificador (ID).

Tabela 2.4: Extração de Dados dos Artigos Selecionados.

ID	Contexto	Objetivo	Técnica(s) de Redução de Dimensionalidade Avaliada(s)	Domínio(s) de Dados	Métrica(s) de Avaliação	Objetivo da(s) Métrica(s)
1	O sensoriamento remoto é uma área que provê terabytes de imagens diariamente, sendo normalmente os dados representados num espaço multidimensional, necessitando de técnicas de redução de dimensionalidade para a análise dos mesmos. Existem muitas técnicas para alcançar tal objetivo e a comparação da performance de cada uma delas é interessante para a escolha da melhor técnica.	Analisar seis métodos de redução de dimensionalidade, um linear e seis não-lineares, utilizando duas técnicas de avaliação denominadas confiabilidade e continuidade a partir de uma base sintética e duas bases provenientes de sensoriamento remoto real.	PCA, Kernel Principal Component Analysis (KPCA), Diffusion Maps (DM), Sammon Mapping (S), Autoencoders (A) e Locally Linear Coordination (LLC).	Três bases de dados: uma sintética, contendo distribuições gaussianas de 20 dimensões com baixa variância e duas provenientes de sensoriamento remoto, a saber UC Merced Land Use Dataset (UCM), consistindo de 21 classes com 90 imagens de sensoriamento remoto de 256x256 pixels. Processando cada imagem a partir das técnicas Spectral Signatures (SPECTRAL) e Weber Local Descriptors (WLD) foram obtidos um espaço de características de 192 e 432 dimensões respectivamente. **Em resumo: a base de dados 1 possui dimensionalidade 20, a base de dados 2 possui 90 instâncias de 192 dimensões e a base 3 possui 90 instância de 432 dimensões.**	Confiabilidade e Continuidade.	A confiabilidade avalia a quantidade dos k-vizinhos presentes no espaço de alta dimensão também preservados no espaço de baixa dimensão. A continuidade, por sua vez, avalia quantidade dos k-vizinhos presentes no espaço de alta dimensão, mas não preservados no espaço de baixa dimensão.
2	Conjunto de dados de alta de dimensão ocorrem comumente em diversos domínios de aplicação. Eles são frequentemente analisados através de métodos de redução de dimensionalidade. Para analisar a fidelidade destas técnicas de redução dimensional, usuários necessitam avaliar suas qualidades.	Analisar comparativamente diferentes métodos de avaliação da qualidade de redução de dimensionalidade.	PCA, t-Distributed Stochastic Neighbor Embedding (t-SNE).	Conjunto de dados Swiss Roll (sintético) e base de dados proveniente de reconhecimento ótico de manuscritos contendo 5620 instâncias com um vetor de características de 64 dimensões. **Em resumo: Sobre a base de dados 1 não é mencionado maiores informações e a base de dados 2 possui 5620 instâncias de 64 dimensões.**	Root Mean Square Error (RMSE), Kruskal's stress, Residual variance, Spearman's rank correlation, Neighborhood loss, Mean Relative Rank Error (MRRE).	O RMSE avalia o desvio da distância quadrática média. O Kruskal's stress avalia o desvio da distância quadrática média penalizando pequenos desvios. A técnica Residual Variance (Variância Residual) avalia a correlação entre a distância no espaço original (alta dimensão) e no espaço mapeado (baixa dimensão). A técnica Spearman's rank correlation (Correlação de Classificação de Spearman) avalia a correlação entre os *ranks* no espaço original (alta dimensão) e no espaço mapeado (baixa dimensão). A técnica Neighbourhood loss (perda de vizinhança) analisa mudança nos k-vizinhos mais próximos e na preservação dos grupos. Por fim, a técnica MRRE analisa a extrusão e intrusão de k vizinhos mais próximos.

3	O aumento do uso de imagens no processo de sensoriamento remoto vem aumentando continuamente. O desafio para análise destes dados é reduzir sua complexidade preservando informações relevantes. Para alcançar tal objetivo e mapear esta base de dados de imagens em um espaço tridimensional é necessário o uso de técnicas de redução dimensional, observando qual destas é a mais apropriada para a base de dados em questão, por meio de critérios de avaliação.	Analisar quais técnicas são mais apropriadas para a análise de uma base de dados de imagens provenientes de sensoriamento remoto.	PCA, Factor Analysis (FA), Linear Local Tangent Space Alignment (LLTSA), t-SNE, LLC e Stochastic Proximity Embedding (SPE).	A primeira base de dados é a UCM consistindo de 1890 *patches* (blocos de pixels de tamanhos definidos) igualmente distribuídos em 21 classes sendo que cada *patch* tem dimensionalidade 256 x 256 pixels. Cada *patch* foi descrito pelo seu histograma de cores (Hist) e WLD gerando um espaço de 192 e 432 dimensões. A segunda base de dados é sintética formada por 600 vetores aleatórios com distribuição gaussiana de 20 dimensões cada. **Em resumo: A base de dados um possui 1890 instâncias, contendo 192 ou 432 dimensões (a depender da técnica de descrição de cada instância) e a base de dados 2 possui 600 instâncias de 20 dimensões.**	Classificação de distância (distant rank) e preservação de vizinhança.	A técnica Neighborhood preservation (preservação de vizinhança) analisa se os n-vizinhos mais próximos presentes no espaço de alta dimensão ainda continuam vizinhos, após a redução dimensional, no espaço de baixa dimensão. A classificação de distância da mesma forma analisa se a distância de cada elemento da base de dados é mantida após a redução dimensional.
4	Várias técnicas de projeção multidimensional têm sido propostas ao longo dos anos. A qualidade destas projeções pode ser avaliada pela qualidade das técnicas de redução dimensional. Duas destas técnicas são a análise de stress e análise de grupo (silhouette coefficient). A primeira avaliação de qualidade objetiva verificar se as similaridades presentes no espaço de alta dimensionalidade são preservadas no espaço projetado. A segunda métrica de avaliação verifica se as instâncias de uma mesma classe são mantidas no mesmo grupo no espaço projetado.	Apresentar duas novas abordagens de técnicas de avaliação de projeções multidimensionais a saber, Simplified Stress (Stress Simplificado) and Simplified Silhouette Coefficient (Coeficiente de Silhueta Simplificado), que possuem um tempo de processamento menor que as técnicas originais.	PCA, Multidimensional Scaling (MDS), Interactive Document Map (IDMAP), FASTMAP, Least Square Projection (LSP), Local Affine Multidimensional Projection (LAMP) e Part-Linear Multidimensional Projection (PLMP).	Várias bases de dados foram utilizadas como: CBR-ILP-IR que é uma base de dados composta por artigos científicos de três áreas de inteligência artificial. WDBC, base de dados de câncer de mama obtido a partir de imagens digitalizadas de nódulos nos seios. Shuttle, base de dados composta de informações de registro de naves espaciais. Mammals-8 e Mammals-20, bases de dados criadas artificialmente para representar características dos mamíferos. **Em resumo: Não foram citadas maiores informações quantitativas sobre as bases de dados.**	Stress e Silhouette Coefficient.	A técnica Silhouette Coefficient objetiva aferir a qualidade dos grupos no espaço projetado, medindo a coesão e separação entre instâncias de um mesmo grupo.

5	Nos últimos anos várias técnicas de redução de dimensionalidade foram propostas para análise visual de dados multidimensionais. Dado um conjunto de observações de n-dimensões, tais algoritmos criam uma projeção em 2 ou 3 dimensões. Usuários julgam dificultosa a análise da eficiência e da qualidade destas projeções em relação à manutenção das características do espaço original, além de comparar duas técnicas entre si.	Propor um conjunto de visualizações interativas com o objetivo de ajudar os usuários na análise e verificação da qualidade de projeções multidimensionais com respeito a preservação de vizinhança, além de analisar erros de projeção inerentes ao processo.	LSP, LAMP, PLMP, Pekalska e ISOMAP .	Base de dados Freefoto, contendo 3462 imagens (instâncias) agrupadas em 9 classes não balanceadas, extraindo para cada imagem 130 BIC (border-interior pixel classification). Outra base de dados utilizada foi a Corel, composta de 1000 fotografias (instâncias) que abrangem 10 assuntos específicos, extraindo de cada foto um vetor de 150 descritores SFIT. A base de dados News, contendo 1771 feeds RSS provenientes do BBC, CNN, Reuters e Associated Press foi utilizada. Por fim foi utilizada a base de dados Sourceforge contendo 24 métricas de software computadas em 6773 projetos C++ open-source provenientes do website sourceforge.net. **Em resumo: a base de dados Freefoto possui 3462 instâncias e 9 classes. A base de dados Corel possui 1000 instâncias e 10 classes. A base de dados News possui 1771 instâncias. A base de dados Sourceforge possui 24 instâncias e 6773 classes.**	Preservação de Vizinhança (Neighborhood Preservation).	A técnica Neighborhood preservation (preservação de vizinhança) analisa se os n-vizinhos mais próximos presentes no espaço de alta dimensão ainda continuam vizinhos, após a redução dimensional, no espaço de baixa dimensão.
6	Conjuntos de dados de alta dimensão são uma ocorrência prevalente em muitos domínios de aplicação. Estes dados são comumente visualizados usando métodos de redução de dimensionalidade, do inglês Dimensionality Reduction (DR). Os métodos DR proporcionam um mapeamento bidimensional dos dados mantendo características relevantes da alta dimensão, como distâncias locais entre pontos. Uma vez que muitas técnicas de DR tem surgido, avaliar suas qualidades têm se tornado cada vez mais importante.	Criar um novo método para quantificar e comparar a qualidade de técnicas de redução de dimensionalidade baseado em persistência homológica. O método informa qual a melhor técnica de redução de dimensionalidade para uma base de dados e uma métrica específica e provê dados sobre a qualidade local do mapeamento.	t-SNE, MDS, Isomap, PCA e SPE.	Base de dados Isomap Faces contendo 698 imagens (instâncias) de 64x64 pixels cada. Outra base de dados utilizada foi o German Climate Computing Centre (DKRZ) contendo uma matriz de 192 x 96 de diferentes localidades na Terra e 6 variáveis contínuas. **Em resumo: a base de dados Isomap Faces possui 698 instâncias. A base de dados DKRZ possui 18432 instâncias.**	Preservação de vizinhança (neighborhood preservation).	A preservação de vizinhança avalia a quantidade dos k-vizinhos presentes no espaço de alta dimensão também preservados no espaço de baixa dimensão.

7	Projeções multidimensionais são ferramentas utilizadas para análise exploratória de uma grande variedade de dados de alta dimensão. Alguns métodos de análise da qualidade da projeção falham em capturar dados que são essenciais para a interpretação do usuário como a capacidade de transporte de informações das classes ou a preservação de grupos e da vizinhança do espaço original.	Propor um framework unificado a fim de obter medidas objetivas do comportamento local dos mapeamentos das projeções. São utilizadas no framework análises das medidas das propriedades visuais e da preservação destas propriedades originais.	LSP, PCA, t-SNE e S.	Base de dados Optical Recognition of Handwritten Digits (Optidigits), contendo ocorrências dos dígitos de 0-9, 10 classes, 1797 itens, 64 atributos e dissimilaridade euclidiana. Já a segunda base de dados News2011, formada por coleções de feeds RSS provenientes de vários provedores, contendo 23 classes, 1771 itens, 834 atributos e dissimilaridade de cossenos. **Em resumo: a base de dados Optidigits possui 1797 instâncias, 10 classes e dimensionalidade 64. A base de dados News2011 possui 1771 instâncias, 23 classes e dimensionalidade 834.**	Preservação de vizinhança (neighborhood preservation), validação de grupo (Group Validation), Validação de separação de classe (Class Separation Validation), Agregação de Classe (Class Aggregation) e Separação de Classe (Class Separation).	A preservação de vizinhança avalia a quantidade dos k-vizinhos presentes no espaço de alta dimensão também preservados no espaço de baixa dimensão. A validação de grupo verifica se grupos observados na projeção são de fato formados por pontos próximos no espaço original. A Validação de Separação de Classe mede a pureza da classe na vizinhança do ponto de referência no espaço projetado com relação ao espaço original. A Agregação de Classe mede a proximidade visual dos pontos em uma classe específica. Por fim a Separação de Classe mede a pureza da classe na vizinhança de um ponto de referência, ou seja, se as classes estão bem segregadas.
8	Como o número e a complexidade de técnicas de visualização tem crescido, cada vez mais é mais dificultosa a decisão de qual técnica escolher para uma situação ou aplicação específica. Da mesma forma inúmeras métricas de avaliação têm sido desenvolvidas.	Apresentar o Projection Inspector como uma abordagem que permite entender diferenças entre projeções e como método interativo de avaliação das mesmas. Propor uma variação da técnica de preservação de vizinhança (Neighborhood Preservation) chamada Smooth Neighborhood Preservation (SNP).	PLMP, LSP, LAMP, Sammon's Mapping Speeding-up (SMS) e Hybrid Model (Hybrid).	Foram utilizadas 4 bases de dados: Caltech, wdbc, segmentation e fibers. Caltech é uma base de dados formada por imagens. Wdbc é uma base de dados de câncer de mama obtida a partir de 569 imagens de nódulos nos seios, possuindo cada instância 30 dimensões, sendo classificadas em dois grupos: malignos e benignos. Segmentation é uma base de dados de imagens, com 7 classes diferentes. Por fim fibers é uma base de dados de 19000 instâncias classificadas em 8 classes diferentes. **Em resumo: não há informações quantitativas sobre a base de dados Caltech. A base de dados wdbc possui 569 instâncias, 2 classes e dimensionalidade 30. A base de dados Segmentation possui 7 classes. A base de dados Fibers possui 19000 instâncias e 8 classes.**	Stress, Smooth Neighborhood Preservation e Correlation Coefficient.	O Stress calcula uma medida de preservação das distâncias do espaço original com relação ao espaço projetado, ou seja, quão bem as distâncias originais são preservadas. O coeficiente de correlação visa medir como as distâncias no espaço original estão correlacionados com aquelas no espaço projetado. Por fim o Smooth Neighborhood Preservation é uma pequena variação do já conhecido método Neighborhood Preservation.

9	Lidar com base de dados de imagens de alta dimensão requer novas abordagens na mineração de dados, onde a vistualização é a mais importante. Técnicas de redução de dimensionalidade em termos de preservação da estrutura dos dados. Modelar a redução de dimensionalidade como um canal de comunicação, onde pontos no espaço de alta dimensão são transferidos para um espaço de baixa dimensão, a fim de medir a qualidade deste canal e consequentemente a qualidade da técnica de redução de dimensionalidade.	Propor uma nova métrica para avaliar a qualidade de técnicas de redução de dimensionalidade em termos de preservação da estrutura dos dados. Modelar a redução de dimensionalidade como um canal de comunicação, onde pontos no espaço de alta dimensão são transferidos para um espaço de baixa dimensão, a fim de medir a qualidade deste canal e consequentemente a qualidade da técnica de redução de dimensionalidade.	Laplacian Eigenmaps (LE), Stochastic Neighbor Embedding (SNE) e Locally Linear Embedding (LLE).	Duas bases de dados: UCM e Corel image. A primeira base de dados é formada por 2100 imagens classificadas em 21 grupos. A segunda base contém 1500 imagens classificados em 15 grupos. **Em resumo: a base de dados UCM possui 2100 instâncias e 21 classes. A base de dados Corel image possui 1500 instâncias e 15 classes.**	Communication Channel Model.	Preservação da estrutura global.	Enquanto os dados são transferidos do espaço de alta dimensão para o espaço de baixa dimensão matrizes são construídas a partir dos dados. Essas matrizes são mescladas para definir uma distribuição de probabilidade, a partir da qual é possível avaliar a qualidade da técnica de redução de dimensionalidade.
10	Dados coletados de várias aplicações em diversos campos como ciências biológicas e processamento de informações multimídias são frequentemente de alta dimensionalidade dificultando o processo de mineração de dados. Técnicas de redução de dimensionalidade são utilizadas para capturar a representação de baixa dimensão do espaço de alta dimensão, sendo que, idealmente, estas representações podem preservar a configuração local.	Propor um novo critério de avaliação para medir a qualidade de técnicas de redução de dimensionalidade não-lineares.	Três técnicas lineares: PCA, MDS, Independent Component Analysis (ICA) e nove métodos não lineares: LLE, Laplacian eigenmap, Local Tangent Space Alignment (LTSA), Hessian LLE, Maximum Variance Unfolding (MVU), LLC, Neighborhood Preserving Embedding (NPE), Linearity Preserving Projection (LLP).	Duas bases de dados sintéticas, A e B, sendo a primeira constituída de 1500 instâncias e segunda de 1200 instâncias. Outra base de dados constituída de imagens, contêm 698 vetores de 4096 dimensões cada. **Em resumo: a base de dados A possui 1500 instâncias. A base de dados B possui 1200 instâncias. A terceira base de dados possui 698 instâncias e dimensão 4096.**		Mensurar o quão bem a estrutura global do mapeamento é mantida após técnica de redução de dimensionalidade não linear.

11	Técnicas de redução de dimensionalidade proveem representação de baixa dimensão de base de dados de alta dimensão. Muitos métodos não lineares foram propostos nos últimos anos, enquanto a questão de avaliação e comparação dos mesmos permanece aberta.	Revisar alguns dos critérios de avaliação existentes baseados na classificação de distâncias e na vizinhança. Propor um novo critério que quantifica dois aspectos dos mapeamentos: sua qualidade geral e a tendência em favorecer a extrusão (erro de classificação negativa) ou intrusão (erro de classificação positiva).	Curvilinear Component Analysis (CCA), S.	Uma base de dados artificial composta de 1000 instâncias geradas aleatoriamente a partir de uma esfera oca de raio 1. Uma base de dados real formada por 1965 imagens da mesma face, com diferentes orientações e expressões, sendo cada imagem convertida em um vetor de 560 dimensões. **Em resumo: a primeira base de dados possui 1000 instâncias. A segunda base de dados possui 1965 instâncias e dimensão 560.**	Confiabilidade, Continuidade, Local Continuity Meta-criterion (LCMC) e MRRE.	A confiabilidade avalia a quantidade dos k-vizinhos presentes no espaço de alta dimensão também preservados no espaço de baixa dimensão. A continuidade, por sua vez, avalia quantidade dos k-vizinhos presentes no espaço de alta dimensão, mas não preservados no espaço de baixa dimensão.

Tabela 2.5: Identificador para cada artigo.

ID	**Artigo**
1	(GRIPARIS; FAUR; DATCU, 2016a)
2	(RIECK; LEITTE, 2015a)
3	(GRIPARIS; FAUR; DATCU, 2016b)
4	(ELER et al., 2015)
5	(MARTINS et al., 2014)
6	(RIECK; LEITTE, 2015b)
7	(MOTTA et al., 2015)
8	(PAGLIOSA et al., 2015)
9	(BABAEE; DATCU; RIGOLL, 2013)
10	(MENG; LEUNG; XU, 2011)
11	(LEE; VERLEYSEN, 2009)

2.6 CONCLUSÃO

É notório que a demanda de análise de conjuntos multidimensionais tem aumentado consideravelmente, o que consequentemente contribui para o desenvolvimento de novas técnicas de redução dimensionalidade e por conseguinte novas métricas de avaliação das mesmas. Isto tem sido evidenciado pelos trabalhos descritos na Tabela 2.4.

Analisando-se os artigos descritos na tabela supracitada, percebe-se que foram utilizadas 19 métricas de avaliação de projeções multidimensionais distintas, a saber, confiabilidade, continuidade, Root Mean Square Error (RMSE), Kruskal's Stress, Residual Variance, Spearman's Rank Correlation, Neighborhood Preservation, Mean Relative Rank Error (MRRE), distant Rank, Silhouette Coefficient, Validação de Grupo, Validação de Separação de Classe, Agregação de Classe, Separação de Classe, Smooth Neighborhood Preservation, Correlation Coefficient, Communication Channel Model, Preservação da estrutura Global e Local Continuity Metacriterion. Dentre todas as citadas, há uma popularidade enorme entre a Preservação de Vizinhança (Neighborhood Preservation) e Stress, sendo, portanto, as técnicas mais utilizadas entre todos os artigos estudados. Mais adiante no Capítulo 4, estas duas técnicas serão apresentadas detalhadamente.

A partir da leitura dos estudos apresentados é possível perceber que tais métricas foram aplicadas nas mais diversas técnicas de projeção multidimensional, sendo estas lineares ou não, como por exemplo, PCA, LSP, Isomap (Isomap), t-SNE, MDS dentre outras. No entanto, a técnica PCA foi a mais utilizada nos experimentos, confirmando assim sua popularidade como técnica de redução de dimensionalidade.

2.7 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentados a definição, as etapas de uma revisão sistemática da literatura, bem como todo o processo de sua execução na busca por métricas de avaliação

de projeções multidimensionais. Esta revisão possibilitou a construção de conhecimentos sólidos a respeito do problema de pesquisa, além de evidenciar o atual estado arte, sobre a área, até o presente momento.

A partir de toda análise realizada percebe-se que a área de projeções multidimensionais tem ganhado expressivo foco, principalmente pela crescente demanda da área de análise visual que trabalha frequentemente com dados de alta dimensão. A todo momento novas técnicas de visualização surgem e o processo de escolher a técnica mais adequada se torna mais difícil.

Ficou claro que a quantidade de estudos/artigos que abordam técnicas de avaliação da qualidade de projeções multidimensionais é em menor escala quando comparada ao montante daqueles que apenas se utilizam delas para realizar uma análise específica. Porém, como já explicado anteriormente, a necessidade de analisar cada vez mais dados multidimensionais, acrescida da grande gama de opções de técnicas de visualização, tem forçado a comunidade científica a utilizar métricas de avaliação, além de conceber novas técnicas para tal finalidade.

A partir dos dados coletados nesta revisão, ficou claro que há diversas técnicas de avaliação de projeções multidimensionais, no entanto as mais utilizadas são a preservação de vizinhança (Neighborhood Preservation) e o conjunto de funções stress.

Também pode-se afirmar que o estudo na área de avaliação de projeções multidimensionais é válido, visto que a área tem sido visada fortemente nos últimos anos, como apresentado nesta revisão, além de possuir grande perspectiva de crescimento nos próximos anos.

Neste capítulo foram apresentadas algumas métricas de avaliação das projeções multidimensionais, técnicas estas provenientes da área de visualização de dados multidimensionais. Nesta área é encontrada uma grande variedade de métodos que permitem a análise exploratória dos dados, sendo alguns destes métodos mais tradicionais. No próximo capítulo são apresentadas estas técnicas mais convencionais, evidenciando suas características e limitações.

Capítulo 3

VISUALIZAÇÃO DE DADOS MULTIDIMENSIONAIS

3.1 CONSIDERAÇÕES INICIAIS

A visualização de dados multidimensionais objetiva permitir a análise exploratória de dados de uma forma clara e efetiva através de representações gráficas. A partir de tal representação é possível descobrir relações entre os dados que não são facilmente observadas olhando dados brutos (HAN; PEI; KAMBER, 2011).

Neste âmbito, a maior contribuição das representações visuais é a de identificar padrões, tendências e relacões complexas entre itens a partir da visualização de suas estruturas (CHEN; HÄRDLE; UNWIN, 2007).

Uma das maiores dificuldades da área de visualização é projetar estes dados em espaços bi ou tridimensionais, ou seja, o desafio está em visualizar base de dados com mais de três dimensões. Em adição, a peculiaridade dos dados multidimensionais dificulta o processo de *insight*, ou seja, um entendimento mais profundo e acurado, por parte do usuário final. Diferentes métodos têm sido desenvolvidos, a fim de vencer tal barreira e dar melhor compreensão do fenômeno estudado.

Neste capítulo são apresentados alguns dos principais métodos tradicionais de visualização de dados multidimensionais. Utilizando-se das divisões encontradas em Han et al. (2011) e em Mazza (2009), a Seção 3.2 aborda as técnicas orientadas a pixel, a Seção 3.3 introduz métodos geométricos, a Seção 3.4 aborda as técnicas iconográficas e a Seção 3.5 discorre sobre as técnicas hierárquicas. Por fim, na Seção 3.6, é realizada uma análise comparativa entre as técnicas.

3.2 TÉCNICAS ORIENTADAS A PIXEL

Uma maneira bastante elementar de mensurar visualmente uma dimensão é utilizar um pixel, onde sua cor associada representa o valor do atributo. Neste cenário, o pixel seria a subdivisão máxima da representação gráfica, sendo impossível ir além.

Em tese, o número máximo de elementos a serem mapeados na tela de um computador, por exemplo, é restrito à sua resolução. Sendo assim, uma tela com resolução de 1024 x

768 pixels poderia representar um máximo de 786,432 componentes distintos. Uma outra tela com resolução *Full HD* (1920 x 1080 pixels) seria capaz de representar 2,073,600 elementos. No entanto, nas aplicações práticas, este limite nunca é atingido, uma vez que elementos funcionais e estéticos, como botões, barras de rolagem, bordas, dentre outros, ocupam áreas reservadas da tela, além de que mapear uma unidade de informação a um único pixel é restrito a um número limitado de situações (MAZZA, 2009). Um ponto positivo destas técnicas é que, estas maximizam a quantidade de informação exibida de uma só vez sem o problema de sobreposição. Elas acabam preservando a percepção de pequenos grupos de interesse enquanto mantém uma visão global (PATRO, 2004).

Mas como é a aplicação da orientação a pixel em um conjunto de dados multidimensionais? Supondo uma base de dados m-dimensional, são geradas m janelas, cada uma para uma dimensão. Cada dimensão é então mapeada em sua janela específica. Isto é representado na Figura 3.1, onde é possível perceber o mapeamento de uma instância da base de dados no agrupamento de janelas.

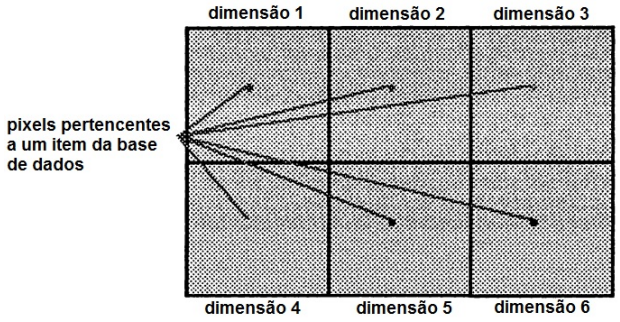


Figura 3.1: Agrupamentos de janelas para base de dados de 6 dimensões.
Fonte: (KEIM, 1996)

Alguns fatores influenciam a aplicação da referida técnica, como definido em Keim (2000): *mapeamento de cor, forma da janela e arranjo de pixels.*

No *mapeamento de cor* a questão central é como as cores dos pixels serão representadas. Neste caso, a propriedade a ser explorada é o brilho, onde a partir dela é possível diferenciar cores para a representação de instâncias de dados distintas. Sendo assim, a utilização de uma escala de cor com variação de brilho é suficiente ao invés de utilizar todas as possibilidades de cores.

O outro fator a se analisar neste tipo de representação é o *arranjo de pixels.* A problemática está em como organizar espacialmente os pixels dentro da área de cada janela. Esta questão é de extrema importância pois, a partir de um bom arranjo, é possível descobrir a existência de grupos, bem como possíveis relações entre as dimensões da base de dados. Alguns dos mais comuns arranjos de pixels são mostrados na Figura 3.2.

Por fim, a forma da janela é um aspecto também a ser considerado. Tradicionalmente utiliza-se o formato retangular. Este possibilita um bom aproveitamento em termos de área, no entanto, ocasiona uma dispersão pronunciada o que dificulta a percepção de

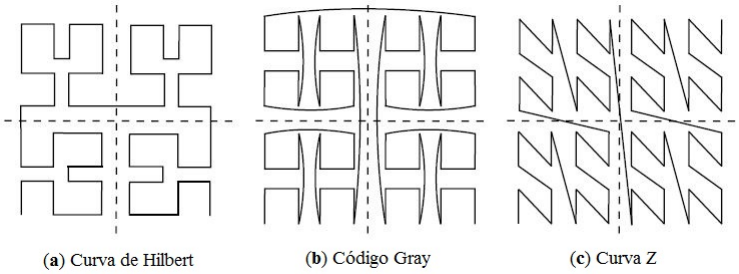


Figura 3.2: Alguns arranjos comumente utilizados segundo Han, Micheline e Kamber (2012).

correlações entre os dados, por exemplo. Como alternativa, tem-se o formato circular, como apresentado na Figura 3.3, onde o item (a) representa o agrupamento circular das janelas e o item (b) o arranjo dos pixels no formato circular.

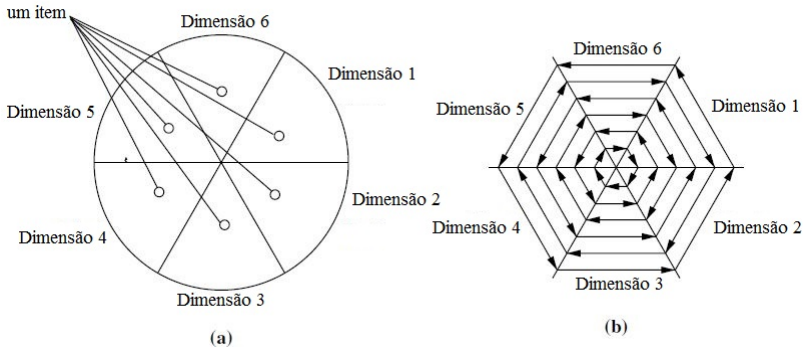


Figura 3.3: Agrupamento das janelas e arranjo de pixels no formato circular.
Fonte: (HAN; PEI; KAMBER, 2011)

Como ilustração da técnica de orientação a pixels, segue-se a Figura 3.4, exemplo este retirado de Han et al. (2011). Considere uma loja fictícia que guarda dados dos clientes em uma tabela onde as linhas representam os clientes e as únicas quatro colunas a *renda*, o *limite de crédito*, o *volume de transações* e a *idade*. É possível então investigar uma possível correlação entre a *renda* e as outras dimensões utilizando a técnica discutida nesta seção. Uma abordagem é ordenar os clientes por renda em ordem crescente e mapear cada dimensão em 4 janelas distintas representadas na Figura 3.4. De acordo com o valor da *renda* o sombreamento de cada pixel muda. Valores menores são mapeados em pixels de cor próxima ao branco enquanto que valores maiores são representados por pixels com colaração mais próxima do preto. A partir de uma visualização baseada em pixels é possível perceber que a medida que a renda aumenta o limite de crédito segue o mesmo padrão, revelando que ambas as dimensões são correlacionadas. Além disso, pode-se chegar a uma outra conclusão: clientes que possuem renda mediana normalmente são

os que mais compram. Isto é claro ao analisar as janelas da *renda* e *volume de transações*. Porém, não há uma correlação aparente entre *renda* e *idade*.

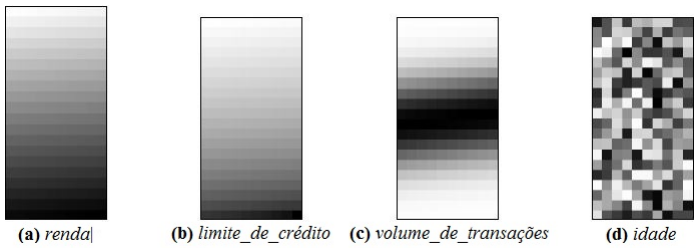


Figura 3.4: Visualização de quatro atributos baseados na renda de todos os clientes.
Fonte: (HAN; PEI; KAMBER, 2011)

3.3 TÉCNICAS GEOMÉTRICAS

Uma desvantagem importante das técnicas orientadas a pixels é que as mesmas não auxiliam no entendimento da distribuição dos dados em um espaço multidimensional (HAN; PEI; KAMBER, 2011). Com elas não é possível, por exemplo, verificar a existência de uma área densa. Neste quesito às técnicas geométricas superam as baseadas em pixels.

As técnicas geométricas consistem em mapear os valores dos atributos em um espaço geométrico (MAZZA, 2009). Tal característica facilita o processo exploratório dos dados a partir de informações de geometria.

A seguir são apresentadas algumas das técnicas geométricas mais comumente utilizadas.

3.3.1 Gráficos de Dispersão

O gráfico de dispersão é uma técnica geométrica bastante simples que mostra dados bidimensionais utilizando coordenadas cartesianas (HAN; PEI; KAMBER, 2011). A ideia é mostrar a distribuição entre atributos a fim de analisar possíveis correlações.

Quando se tem apenas duas dimensões a representação é bem simplista. Por exemplo, seja a base de dados de qualidade do ar da linguagem R. Tal base contém medições diárias da qualide do ar, em Nova Iorque, de Maio à Setembro de 1973. A Figura 3.5 ilustra esta base sendo possível perceber a relação entre a média de ozônio, em partes por bilhão (ppb), e o dia do mês.

Porém, na maioria das situações é necessário analisar mais do que dois atributos. Para isto, é possível utilizar novas propriedades gráficas como cor, textura, tamanho e formato. Com a adição destas propriedades e mais um terceiro eixo é possível representar, no máximo, sete atributos.

Um exemplo deste tipo extensão de atributos visuais é mostrado na Figura 3.6. Nela são mapeados quatro atributos. O dia do mês é mapeado no eixo das abscissas e ozônio no eixo da ordenadas, enquanto que os meses e a velocidade do vento são representados por cores e tamanhos distintos, respectivamente.

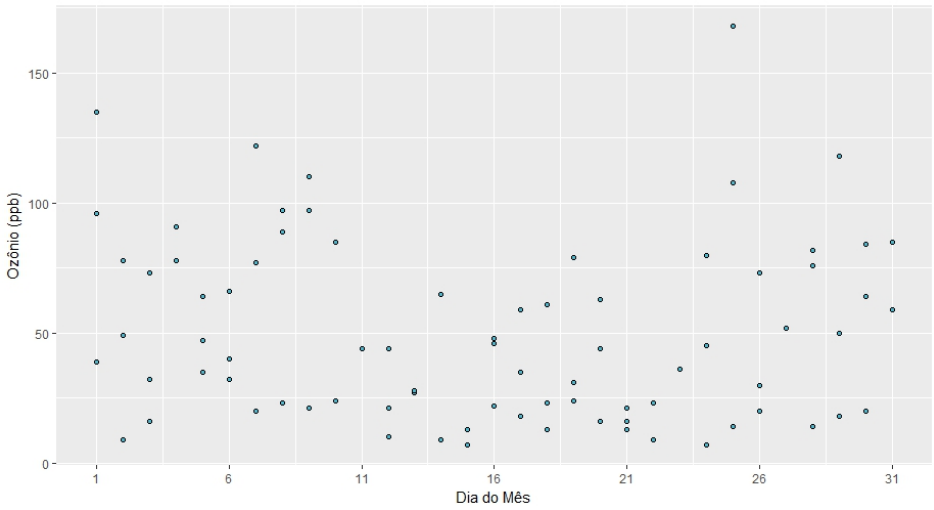


Figura 3.5: Qualidade do ar na cidade Nova Iorque diariamente.

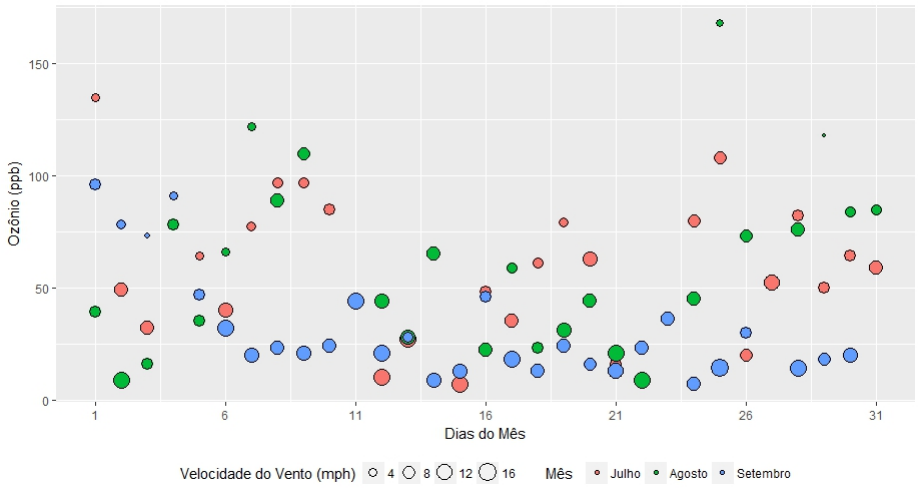


Figura 3.6: Extensão do gráfico de dispersão com representação de quatro atributos.

No entanto, apesar desta abordagem possibilitar um mapeamento maior de dimensões, os gráficos de dispersão não se adequam satisfatoriamente a todos os tipos de dados. É necessário então, partir para outras abordagens.

3.3.2 Matriz de Gráficos de Dispersão

A matriz de gráficos de dispersão é uma extensão dos gráficos de dispersão, onde vários destes são organizados simultaneamente em uma matriz, a fim de prover informações de correlação entre os atributos (CHAN, 2016).

Este método é mais eficiente que os gráficos de dispersão porque consegue mapear um número genérico de dimensões. Por exemplo, a partir de uma base de dados n-

dimensional é formada uma matriz n x n constituída por vários gráficos de dispersão. A partir desta geometria é então possível correlacionar cada atributo com os demais, uma vez que todos eles estarão presentes na matriz.

No entanto, por vezes, devido a questões de economia computacional, a matriz quadrada n x n não é gerada, uma vez que informação redundante é exibida. Por exemplo, ao correlacionar dois atributos a_1 e a_2 dois gráficos serão gerados: o primeiro conterá o a_1 e a_2 nos eixos das abscissas e ordenadas, respectivamente, e um outro gráfico mapeará os mesmos atributos, mas com os eixos trocados, a saber a_1 no eixo das ordenadas e a_2 no eixo das abscissas. As mesmas tendências, padrões e grupos podem ser observados em ambos os gráficos, sendo um espelho do outro. Desta forma, ao invés de gerar uma matriz n x n, dá-se preferência a geração de uma matriz com $\frac{n^2-n}{2}$ elementos, o que em termos práticos é considerar apenas elementos abaixo da diagonal principal.

Como ilustração da técnica, segue-se na Figura 3.7 uma representação, em matriz de gráficos de dispersão, da base de dados Iris. Esta é uma base de dados que contém 50 amostras de cada uma das três espécies da flor Iris, a saber: setosa, versicolor, e virginica. As cinco dimensões de cada amostra são: espécie, comprimento e largura das pétalas e comprimento e largura das sépalas, sendo as mesmas mapeadas na matriz de dispersão de tamanho 5 x 5, como já explicado previamente.

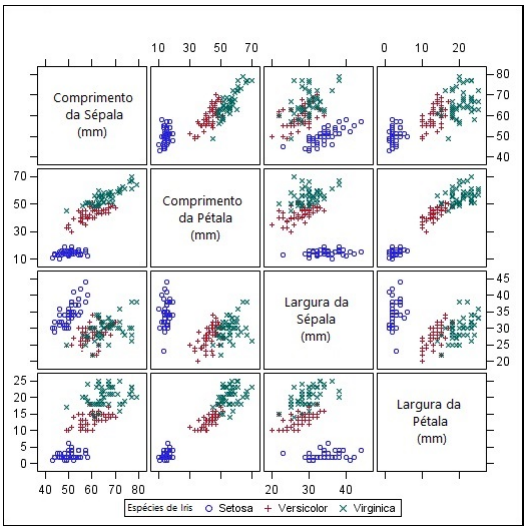


Figura 3.7: Matriz de dispersão da base de dados Iris.
(Fonte: [1].)

Uma vantagem perceptível desta técnica é que correlações entre atributos são facilmente percebidas. No entanto, um fator limitante para sua visualização é a grande utilização de área, uma vez que é necessário mostrar uma matriz de tamanho n x n. Sendo assim, a matriz de dispersão torna-se menos efetiva a medida que a quantidade de dimensões aumenta (HAN; PEI; KAMBER, 2011).

[1] Traduzida de http://support.sas.com/documentation/cdl/en/grstatproc/62603/HTML/default/images/gsgscmat.gif. Acessado em 11 de Maio de 2018

3.3.3 Coordenadas Paralelas

A técnica de coordenadas paralelas possui este nome por causa da forma como os atributos são representados: cada atributo corresponde a um eixo, sendo estes paralelos e igualmente espaçados (MAZZA, 2009). Cada observação corresponde a uma linha poligonal que intersecta todos os eixos, em locais específicos, de acordo com os valores n-dimensionais. As correlações neste tipo de representação são observadas a partir das linhas. Casos em que as linhas de eixos vizinhos cruzam entre si em formato similar ao da letra X revelam correlações inversas. Por exemplo, a Figura 3.8 foi construída na ferramenta High-D, uma ferramenta de visualização de dados multidimensionais, utilizando uma base de dados sintética de 392 carros e 7 dimensões. É fácil perceber que há uma relação inversa entre o consumo do carro (milhas por galão) e o número de cilindros, pois a medida que o valor do atributo *milhas por galão* aumenta o valor do atributo *cilindro* diminui. Uma outra informação explícita na Figura 3.8 é que quanto maior o peso do carro menor sua aceleração.

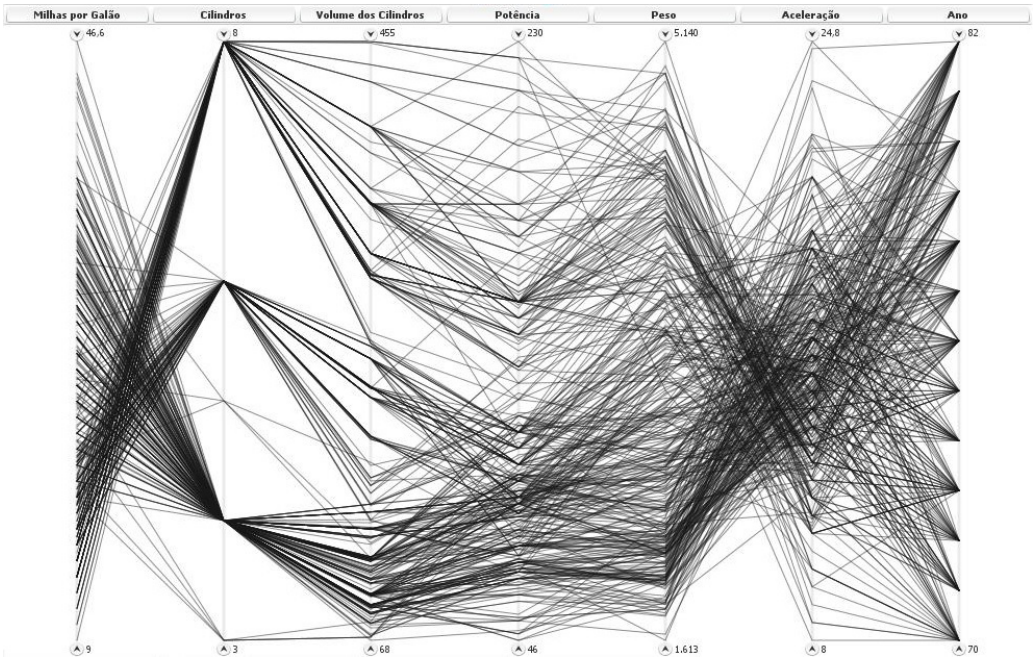


Figura 3.8: Coordenadas Paralelas geradas no software High-D.

Apesar de permitir uma análise exploratória dos dados de uma forma intuitiva e poderosa, as coordenadas paralelas têm duas limitações essenciais: a primeira é que torna-se necessária a proximidade dos eixos representativos de dois atributos em análise, ou seja, adjacência é um fator importante para uma boa interpretação. Eixos muito distantes dificultam a percepção de alguma correlação. A segunda limitação é que a representação visual de bases de dados com extenso número de observações é ineficiente. Em casos extremos, o que antes eram linhas distinguíveis, acabam tornando-se um polígono de cor uniforme, devido a enorme densidade de linhas (instâncias). Isto pode ser visto na Figura

3.9 que ilustra um caso hipotético da mesma base, agora contendo 1000 carros. Ações de movimento dos eixos paralelos e coloração das linhas poligonais podem ser cogitadas a fim de amenizar o efeito da oclusão visual.

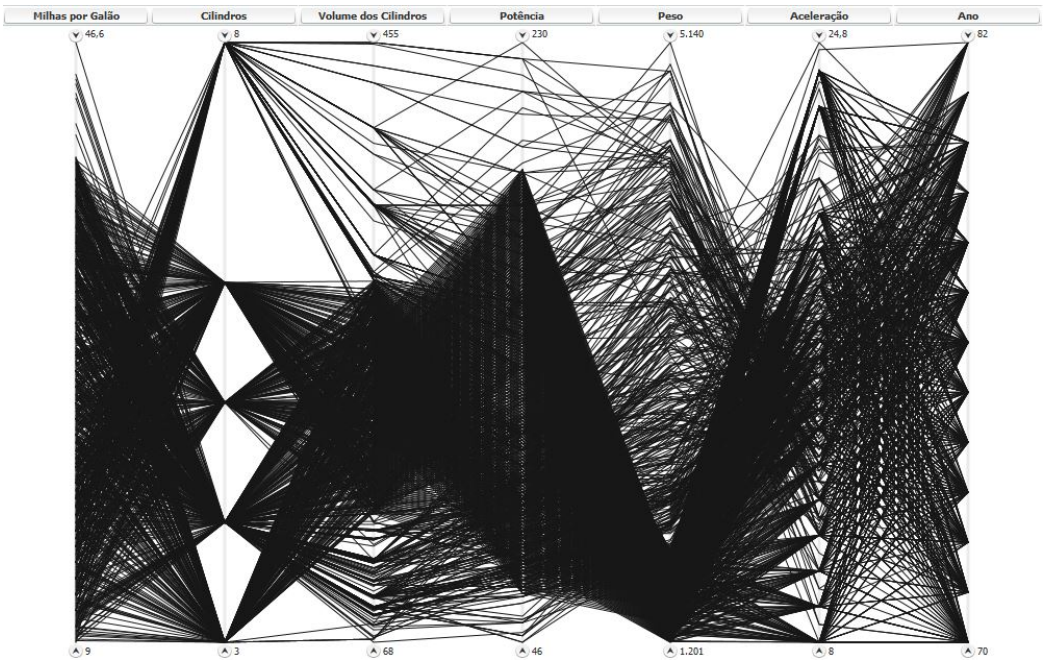


Figura 3.9: Oclusão na técnica de Coordenadas Paralelas gerada no software High-D.

3.3.4 TableLens

A técnica de Visualização *TableLens*, proposta por John Lamping e Ramana Rao em 1994, é inspirada em uma ferramenta bastante conhecida utilizada para dispor dados e realizar cálculos: as planilhas. A ideia é representar dados por meio de barras horizontais ao invés de valores numéricos, como é feito usualmente nas aplicações de planilhas.

Na técnica de *TableLens* cada linha representa uma instância de dados enquanto que as colunas representam os atributos. Barras horizontais são utilizadas para representar os valores das dimensões. Sendo assim, as colunas são vistas como histogramas ou uma representação similar de gráfico de barras. Uma ilustração deste tipo de técnica é mostrada na Figura 3.10.

A interatividade é um fator importante na análise exploratória dos dados e esta característica é presente na TableLens. O usuário consegue facilmente perceber tendências e correlações a partir do reposicionamento das colunas, bem como da ordenação dos valores a elas associadas, por exemplo.

3.4 TÉCNICAS ICONOGRÁFICAS

As técnicas iconográficas utilizam pequenos ícones para representar dados multidimensionais (HAN; PEI; KAMBER, 2011). A ideia se baseia no mapeamento dos atributos

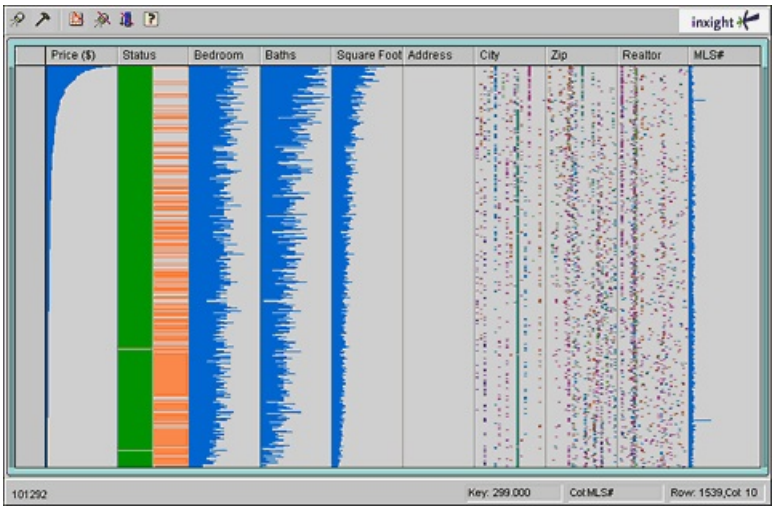


Figura 3.10: Exemplo de Visualização no Table Lens.

a características geométricas em ícones. Estes podem então variar em tamanho, cor, formato, orientação, etc.

3.4.1 StarPlots

Os StarPlots, como o nome sugere, é uma representação geométrica formada por um polígono em formato estrelar, onde os vértices são formados por eixos de uma mesma origem.

Os atributos de cada instância de dados são mapeados em pontos ao longo de cada eixo que, após serem conectados, formam uma "estrela". Para analisar uma base de dados, basta apenas comparar o formato de cada figura geométrica (representante de uma instância) uma com as outras.

Como exemplo, pode-se citar a utilização de uma base de dados contendo os índices de crime dos 50 estados americanos. Na Figura 3.11 é representada as estatísticas do estado de Georgia, um dos estados mais violentos, onde também são representados os 7 atributos da base de dados, a saber, furto, roubo, estupro, assalto, roubo de carros, lesão corporal e assassinato. A partir dela é possível perceber por exemplo, que quanto menor o índice de estupro maior é a ocorrência de assaltos, roubos de carros, furto, etc.

3.4.2 Chernoff faces

Chernoff faces é provavelmente a mais famosa técnica iconográfica. Foi criada e apresentada pelo estatístico Herman Chernoff em 1973.

Esta técnica procura usufruir da habilidade que os seres humanos têm em perceber os mínimos detalhes em características faciais. A ideia é mapear atributos de uma base multidimensional em características e expressões da face humana, como olhos, boca, nariz, dentre outros, a partir de seus tamanhos, formatos, posições e orientações.

Uma representação da técnica é dada na Figura 3.12. Nela há a representação das

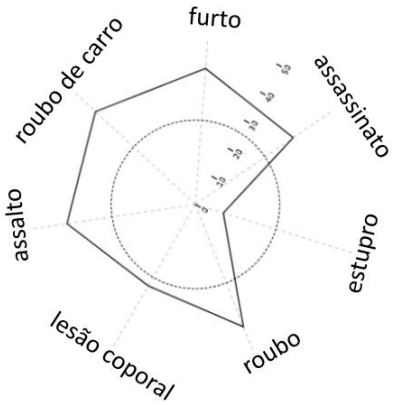


Figura 3.11: StarPlot dos crimes do Estado de Georgia.
(Fonte: Traduzida de https://developer.ibm.com/predictiveanalytics/2012/03/14/reference-lines-for-star-plots-aid-interpretation/. Acessado em 12 de Maior de 2018.)

características climáticas de dezesseis cidades ao redor do mundo por meio das características faciais humanas. Em particular, a área da face mapeia a precipitação média de chuva, a curvatura do rosto representa a média de temperatura, a dimensão do nariz mapeia a média máxima de temperatura e por fim a curvatura, largura e posição da boca mapeiam os valores de temperatura máxima registrada, temperatura mínima e média de temperatura mínima, respectivamente. Assim pode-se ver claramente que a cidade de Hong Kong é mais chuvosa de todas. Porém, deve-se tomar cuidado ao utilizar tal técnica, pois a má interpretação das expressões faciais pode levar o usuário a interpretações erradas. Por exemplo, uma vez que os recordes de temperatura máxima e mínima estão representados pela curvatura e largura da boca, respectivamente, pode-se ter a má interpretação de que o Rio de Janeiro, juntamente com Tunis são as cidades com as melhores condições climáticas, devido ao fato de que os Chernoff faces das mesmas parecem estar sorrindo.

Apesar da facilidade e simplicidade a técnica possui algumas limitações. De acordo com Han et al. (2011), cada face consegue representar até no máximo 18 dimensões. Uma outra limitação é que os valores de cada dimensão não são mostrados na representação das faces, além de que relacionar múltiplos atributos não é tão eficiente.

Visto que a economia de espaço visual tem sido problema em várias técnicas de visualização de dados multidimensionais, as Chernoff faces assimétricas foram propostas como uma melhoria a técnica tradicional. Como a face possui simetria vertical, com relação ao eixo Y, não há necessidade de representar os dois lados da face, mas apenas um. A partir desta pequena modificação, as Chernoff faces assimétricas conseguem representar até 36 dimensões, o dobro da técnica tradicional.

3.5 TÉCNICAS HIERÁRQUICAS

As técnicas até aqui apresentadas e discutidas, possuem uma característica comum: elas representam várias dimensões simultaneamente. Quando a base de dados analisada é pequena isto não é um problema. Porém, a medida que a base cresce, tanto em número

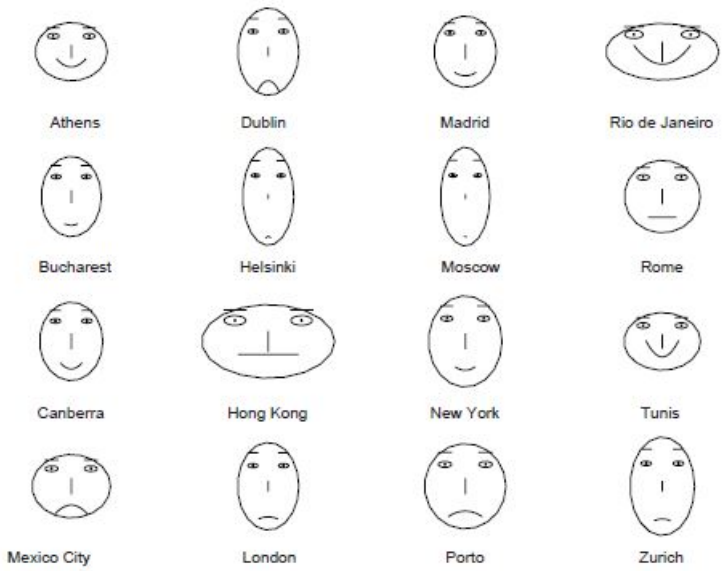


Figura 3.12: Chernoff Faces.
Fonte: (MAZZA, 2009)

de instâncias como em dimensões, a visualização torna-se um pouco ineficiente.

As técnicas Hierárquicas particionam todas as dimensões em subconjuntos (HAN; PEI; KAMBER, 2011). Isto quer dizer que a base de dados passa a ser vista de uma maneira hierarquizada. A seguir são abordadas duas das técnicas mais famosas: TreeMaps e World within Worlds.

3.5.1 World Within Worlds

A técnica denominada World Within Worlds, que em tradução livre seria, "mundo dentro de mundos", tem este nome porque consegue subdividir uma base de dados multidimensional em diversos subespaços do plano cartesiano.

Por exemplo, a partir de uma base formada por 6 dimensões, a saber F, X_1, X_2, X_3, X_4 e X_5, é possível analisar a correlação entre atributos. Supondo que é necessário verificar como a dimensão F varia em relação as demais, pode-se mapear os valores de algumas dimensões em valores fixos, por exemplo X_3, X_4 e X_5 para V_3,V_4,V_5. Estes valores se tornarão pontos em um plano tridimensional. Destas forma, tem-se um gráfico 3-D formado por F, X_2 e X_3 onde o ponto (V_3,V_4,V_5) é a origem de outro subespaço interno tridimensional formado pelas dimensões X_3, X_4 e X_5. Assim o usuário pode, de forma interativa, mudar a posição da origem do subespaço interno e verificar quais os impactos desta mudança nos demais atributos. Além disso é possível também intercambiar entre os diversos "mundos" (dimensões) disponíveis. Caso a dimensão aumente, basta apenas aumentar o número de subespaços. Como forma de ilustração na Figura 3.13 é mostrada a representação da técnica no processo anteriormente descrito.

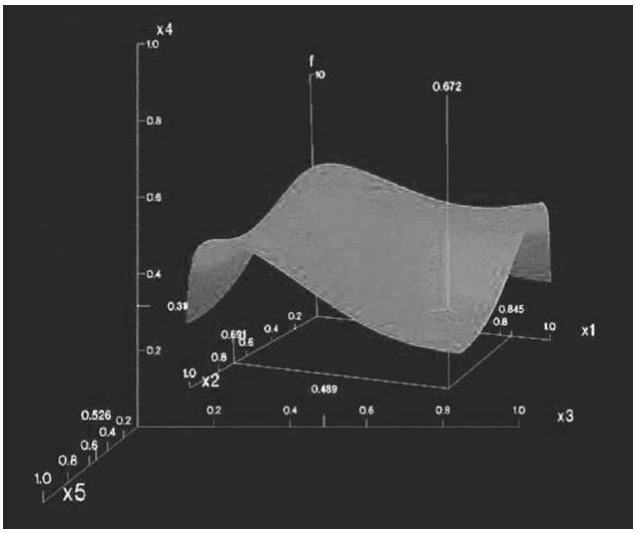


Figura 3.13: Técnica World Within Worlds. (Fonte: Han et al. (2011))

3.5.2 Tree-Maps

Uma outra técnica hierárquica bastante conhecida é a TreeMaps. Ela se baseia na subdivisão hierárquica da tela em regiões, a depender dos valores dos atributos (CHAN, 2016).

Tal representação utiliza retângulos aninhados que podem ser subdividos tanto verticalmente como horizontalmente. Cada retângulo possui uma cor onde esta pode representar outros atributos.

Uma das vantagens das técnicas de Tree-Maps é que elas são adequadas na visualização de grandes bases de dados com diversos atributos ordinários (valores numéricos). Além do mais a técnica tenta ter o maior aproveitamento de tela ocupando ao máximo a área disponível.

A Figura 3.14 ilustra a técnica de tree-map para o Google New Stories. A ferramenta reúne as últimas notícias circuladas na plataforma Google. As histórias são particionadas em sete diferentes grupos representados por retângulos de cores distintas, como pode-se observar no canto inferior direito da imagem. Mais internamente, dentro de cada retângulo, as histórias são novamente agrupadas em retângulos menores, ou seja, subcategorias.

3.6 ANÁLISE COMPARATIVA

Uma vez que as principais técnicas tradicionais de visualização de dados multidimensionais foram abordadas, é interessante realizar uma sumarização comparativa entre as mesmas, de forma a explicitar em qual tipo de situação cada técnica, até aqui vistas, se encaixa melhor.

A seguir é feita uma análise teórica das diversas técnicas com respeito a três aspectos: tipo de dados, informação a ser extraída e escalabilidade.

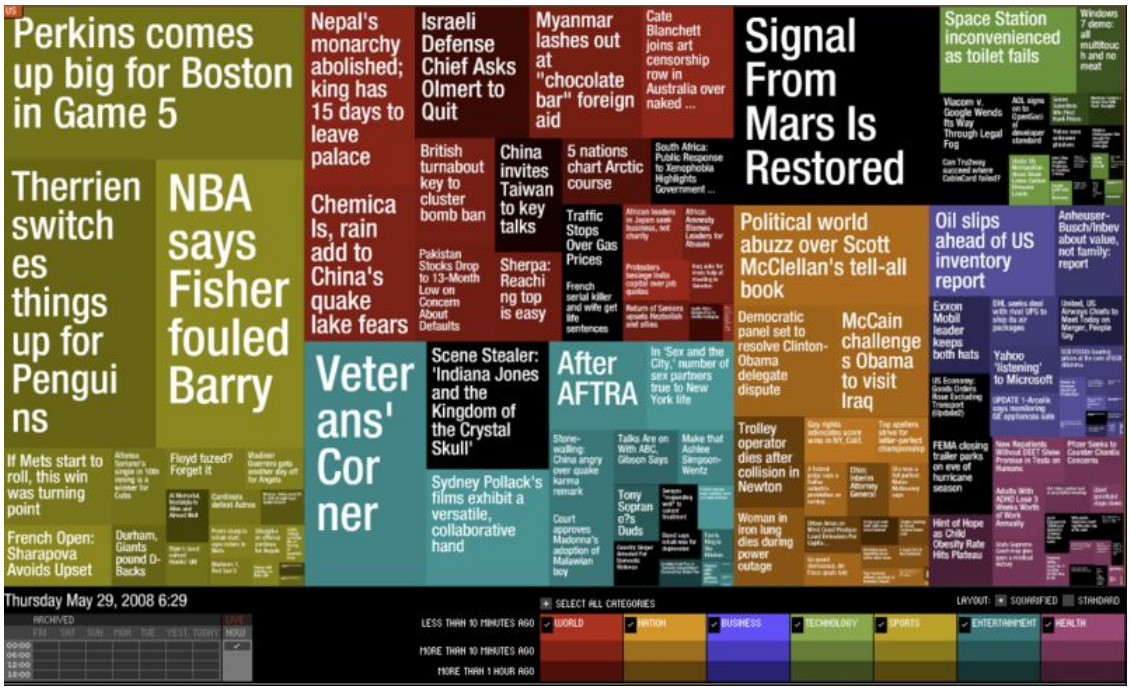


Figura 3.14: Técnica de TreeMap para o Google New Stories.
(Fonte: http://www.science.smith.edu/dftwiki/images/thumb/e/ea/Newsmap.png/800px-Newsmap.png acessado em 14 de Maio de 2018)

3.6.1 Tipo de dados

Um fator importante na hora de escolher a projeção multidimensional é o tipo de dados que se está trabalhando. Estes são dados quantitativos ou qualitativos? Ainda mais, são dados contínuos ou discretos? A partir destes questionamentos é possível escolher a técnica mais adequada.

Seguindo a ordem apresentada, as técnicas orientadas a pixel normalmente são utilizadas em dados quantitativos contínuos ou discretos. No entanto, tais técnicas não são adequadas para dados categóricos (KEIM, 2001). Dados categóricos, basicamente, são dados não numéricos que variam um número finito de vezes. Sendo assim, técnicas baseadas em pixel são recomendadas para dados que podem ser quantizados.

Por outro lado, as técnicas geométricas, geralmente, são flexíveis o bastante para trabalhar com os dois tipos de dados. No entanto, mesmo entre elas há técnicas que se adaptam melhor a tipos de dados específicos. Por exemplo, o método das coordenadas paralelas se ajusta bem à dados quantitativos e qualitativos, sejam eles discretos ou não. No entanto, gráficos de dispersão são mais adequados para dados contínuos.

As técnicas iconográficas são mais apropriadas para dados quantitativos, uma vez que as características geométricas dos ícones variam de acordo com os valores das dimensões (DIAS et al., 2012). Por exemplo, na Chernoff faces, mudanças nos atributos da base modificam as expressões faciais de seus ícones. Para dados nominais as técnicas iconográficas não são uma boa escolha.

Por fim, as técnicas hierárquicas são uma ótima escolha quando os dados a serem

analisados possuem algum tipo de hierarquia e/ou relacionamento. Textos, por exemplo, são um bom exemplo de dados com tais características.

A seguir, na Tabela 3.1, é apresentada a correlação entre cada técnica de visualização de dados multidimensionais com os tipos de dados mais adequados.

Tabela 3.1: Técnicas e seus tipos de dados adequados.

Categoria	**Técnica**	**Dado**			
		Quali.[1]		**Quant.**[2]	
		Nom.[3]	**Ord.**[4]	**Cont.**[5]	**Disc.**[6]
Orientadas a Píxel				x	x
Geométricas	Gráficos de Dispersão			x	
	Matrizes de Dispersão			x	
	Coordenadas paralelas	x	x	x	x
Iconográficas	Star Plots			x	x
	Chernoff Faces			x	x
Hierárquicas	World Within Worlds			x	x
	Tree Maps	x	x	x	x

[1] Qualitativo, [2] Quantitativo, [3] Nominal, [4] Ordinal, [5] Contínuo e [6] Discreto

3.6.2 Informação a ser extraída

É usual que algumas técnicas sejam mais eficientes que as outras em tarefas exploratórias distintas. Isto vai depender de qual informação se deseja extrair da base de dados analisada.

As técnicas orientadas a pixel são ótimas na análise de correlação entre as diversas dimensões, permitindo assim a identificação de padrões. Além do mais, é possível reposicionar pixels a procura de grupos (*clusters*) de dados mais correlacionados.

As técnicas geométricas, por sua vez, são mais indicadas quando o intuito é ter uma visão global dos dados e na busca por padrões. Relacionamentos entre os dados são fáceis de serem percebidos, principalmente na utilização de gráficos ou matrizes de dispersão. A visualização de grupos é possível, porém pode haver a ocorrência de oclusão visual.

As técnicas iconográficas, como já vistas anteriormente, utilizam ícones para mapear os atributos. Cada dado é mapeado individualmente em um único ícone permitindo o reconhecimento de padrões. É possível também agrupar ícones com características geométricas semelhantes permitindo assim a formação de grupos.

Por fim, através das técnicas hierárquicas é possível ter uma visão geral da estrutura dos dados, perceber possíveis relacionamentos, além de permitir a formação de grupos, no caso do TreeMaps, por exemplo. A seguir, a Tabela 3.2 sumariza os tipos de informações mais comuns a serem extraídos em cada técnica.

Tabela 3.2: Técnicas e tipos de informação a serem extraídos.

Categoria	**Técnica**	**Informação**			
		visão geral	**corr.** [1]	**clust.**[2]	**padrões**
Orientadas a Píxel			x	x	x
Geométricas	Gráficos de Dispersão		x		x
	Matrizes de Dispersão		x		x
	Coordenadas paralelas	x	x		x
Iconográficas	Star Plots			x	x
	Chernoff Faces			x	x
Hierárquicas	World Within Worlds		x		x
	Tree Maps	x	x	x	x

[1] Correlações e [2] *Clusters*

3.6.3 Escalabilidade

A escalabilidade é uma característica muito importante quando se trata de visualização multidimensional. A análise deste parâmetro permite perceber como as técnicas se comportam com o aumento do número de instâncias e da dimensionalidade (atributos), além de verificar se uma boa visão geral dos dados é ainda mantida.

Uma boa visão geral da base de dados é essencial, uma vez que em todas as técnicas, com excessão das baseadas em pixel, é possível ocorrer, por exemplo, a oclusão (quando várias instâncias se sobrepõem uma as outras) no processo de visualização das instâncias.

De fato, não há um consenso estabelecido sobre o que é baixa e alta dimensionalidade (RABELO et al., 2008). Esta definição é um pouco arbitrária variando de autor para autor. Por exemplo, de acordo com Böhm, Kriegel (2000), alta dimensão é a partir de 25 atributos. Outra classificação adota a dimensionalidade baixa como até 4 atributos, média a partir de 5 até 9 dimensões e alta acima de 10 de atributos (OLIVEIRA; LEVKOWITZ, 2003). Esta última será a utilizada na discussão a seguir. Sendo assim fica evidente que não há uma convenção a respeito deste assunto.

As técnicas orientadas a pixel conseguem lidar com grandes bases de dados de média a alta complexidade em telas de alta resolução (OLIVEIRA; LEVKOWITZ, 2003). Uma vez que cada dimensão é mapeada em único pixel, a ocorrência de oclusão visual é descartada. Assim tais técnicas possuem uma boa escalabilidade tanto em termos de instâncias como de atributos. No entanto, é importante lembrar que esta escalabilidade será limitada pela resolução da tela. Como visto na Seção 3.2, uma tela com resolução *Full HD* consegue representar um pouco mais de 2 milhões de elementos. Sendo assim, uma base de dados com dezenas ou centenas de milhões de instâncias, por exemplo, não conseguiria ser representada eficientemente pelas técnicas orientadas a pixel, até mesmo porque dificilmente se encontrará uma resolução de tela que consiga representar tamanho número de elementos.

As técnicas geométricas, em sua maioria, conseguem mapear um número médio de instâncias, podendo representar bases de dados de média a alta dimensionalidade. Quando acopladas com técnicas interativas conseguem representar grandes bases de dados (OLIVEIRA; LEVKOWITZ, 2003). No entanto, são menos eficientes, que as técnicas baseadas em pixel, por exemplo.

As técnicas iconográficas conseguem representar bases de pequeno e médio porte, além de poder ser aplicada em conjunto de dados de média a alta dimensionalidade. Sobreposição de instâncias apenas ocorre caso os atributos sejam mapeados para as posições dos ícones na tela.

Por fim, as técnicas hierárquicas conseguem representar dados hierárquicos, sendo mais apropriados para bases de baixa a média dimensionalidade, tendo, no entanto, algumas limitações de área. A seguir é apresentada a Tabela 3.3 que sumariza a escalabilidade de cada técnica.

Tabela 3.3: Técnicas e suas escalabilidades.

Categoria	**Técnica**	**Escalabilidade**	
		Base de dados[1]	**Dimensões**[2]
Orientadas a Píxel		Grande - Muito Grande	Média - Alta
Geométricas	Gráficos de Dispersão	Pequena	Baixa
	Matrizes de Dispersão	Média	Média - Alta
	Coordenadas paralelas	Média	Média - Alta
Iconográficas	Star Plots	Pequena - Média	Média - Alta
	Chernoff Faces	Pequena - Média	Média - Alta
Hierárquicas	World Within Worlds	Média	Baixa - Média
	Tree Maps	Média	Baixa - Média

[1] Refere-se a escabilidade de instâncias, ou seja, ao número de instâncias na base de dados e [2] refere-se a escalabilidade de dimensões/atributos.

3.7 CONSIDERAÇÕES FINAIS

Neste capítulo foram vistas algumas técnicas tradicionais de visualização de dados multidimensionais, abordando métodos mais simples e comuns, como os gráficos de dispersão, e métodos mais robustos como TreeMaps. Cada uma delas possuem vantagens e desvantagens, abordadas e discutidas ao decorrer deste capítulo, sendo sumarizadas em tabelas de acordo com o tipo de dados, informação a ser extraída e escalabilidade.

Porém, uma das técnicas de visualização mais recentes, promissoras e bastante utilizada não foi abordada: as projeções multidimensionais, foco deste trabalho. O *background* adquirido neste capítulo, serve como fator motivacional para abordar as técnicas de projeção multidimensionais a partir de um ponto de vista mais crítico, percebendo suas vantagens e o porquê de estarem em alta.

Sendo assim, o Capítulo 4 encarrega-se de discorrer sobre as projeções multidimensionais.

Capítulo

4

TÉCNICAS DE PROJEÇÃO MULTIDIMENSIONAIS

4.1 CONSIDERAÇÕES INICIAIS

A crescente complexidade dos dados é um padrão notório que precisa ser lidado corretamente, a fim de gerar informações de valor.

Algumas técnicas, vistas no capítulo anterior, foram propostas para tal problemática, cada uma sobressaindo-se individualmente em alguns aspectos. Algumas delas se destacam em proporcionar uma visão global da base de dados permitindo a identificação de padrões, outras conseguem satisfatoriamente denunciar correlações e grupos semelhantes. Dentre estas técnicas é possível identificar aquelas que garantem uma boa escalabilidade, permitindo a visualização de um número relativamente extenso de instâncias e/ou dimensões. Tais particularidades, de certa forma, tornam o usuário refém de mais de um destes métodos, visto que nenhum provê simultaneamente estes aspectos discutidos.

Neste cenário, as técnicas de projeção multidimensional, também chamadas de técnicas de redução de dimensionalidade, têm sido utilizadas como uma abordagem que contornam os desafios supramencionados. Estas ferramentas se caracterizam por permitirem a análise visual de bases de dados multidimensionais de forma eficiente e efetiva (SILVA; RAUBER; TELEA, 2016).

De forma geral, as projeções multidimensionais sobrepõem-se às limitações das técnicas tradicionais, focando-se principalmente na observação das instâncias, ao invés de dedicar-se, em maior ênfase, a um bom entendimento de cada dimensão e seu valor, marca esta registrada das técnicas do Capítulo 3.

As projeções multidimensionais, conceitualmente, mapeiam as instâncias da base de dados em um conjunto (nuvem) de pontos em espaços bi ou tridimensionais, sendo o primeiro caso mais comum. A partir da observação desta nuvem é então possível perceber relações de distâncias e similaridade, encontrar pontos que se distanciam excessivamente dos demais, os chamados *outliers*, além de identificar grupos de interesse. Pontos próximos identificam instâncias similares enquanto os distantes são dissimilares, desde que a projeção seja de qualidade, ou seja, preserve ao máximo a estrutura dos dados no espaço original.

Um exemplo, bastante ilustrativo, é dado na Figura 4.1. Nela é possível perceber uma tabela com múltiplas instâncias e dimensões, onde deseja-se aplicar alguma técnica de projeção multidimensional. Sendo assim, cada linha é mapeada em um único ponto no espaço bidimensional, de forma a preservar as distâncias do espaço original e a cada ponto é atribuída uma coloração referente ao valor de um atributo (coluna) no espaço n-dimensional.

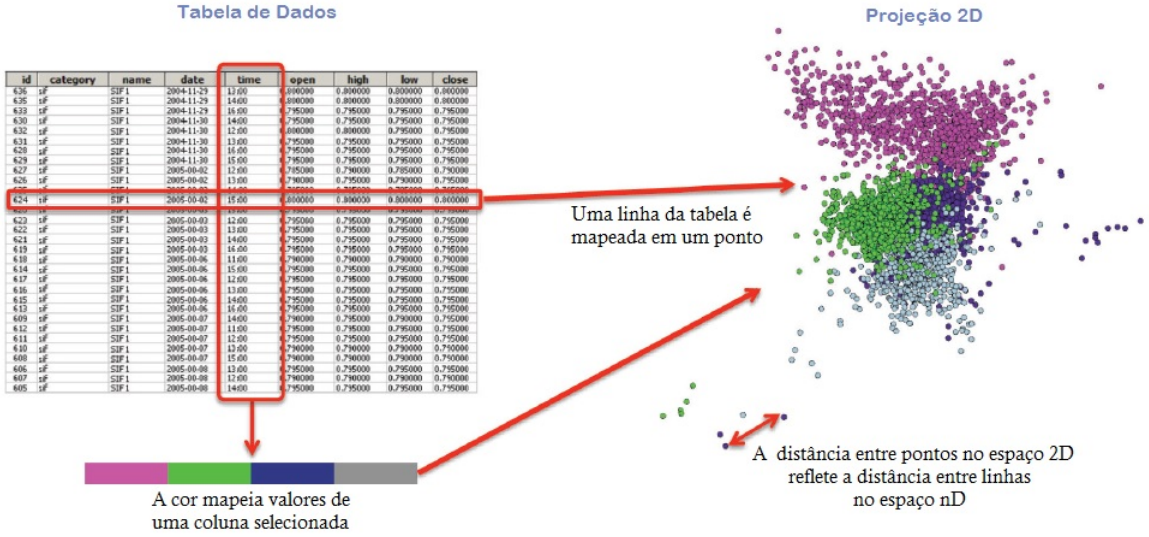


Figura 4.1: De uma tabela multidimensional para uma projeção.
(Fonte: Traduzida de (SILVA; RAUBER; TELEA, 2016))

Partindo para o formalismo matemático, de acordo com Tejada, Minghim e Nonato (2003) uma projeção multidimensional tem a seguinte definição: seja X um conjunto de pontos em $\mathbb{R}^n$ sendo $\delta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ um critério de proximidade entre instâncias em $\mathbb{R}^n$. Seja Y um conjunto de pontos em $\mathbb{R}^p$, onde p $\ll$ n (p = {1,2,3}) e $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ um critério de proximidade em $\mathbb{R}^p$. Uma projeção multidimensional é então definida como uma função de mapeamento $f : X \to Y$ que visa tornar $|\delta(x_i, x_j) - d(f(x_i), f(x_j))|$ o mais próximo possível de zero, $\forall x_i, x_j \in X$.

4.2 CLASSIFICAÇÃO DAS TÉCNICAS DE PROJEÇÃO MULTIDIMENSIONAL

Há uma grande variedade de técnicas de redução de dimensionalidade e elas podem ser classificadas a partir de diferentes critérios. A seguir são apresentadas algumas destas classificações.

4.2.1 Dimensão Vs Distância

Uma classificação a ser adotada é a baseada em *dimensão* x *distância*. Tal classificação é fundamentada no tipo de informação usada como *input* para construir a projeção.

Técnicas baseadas em distância apenas necessitam das distâncias ou similariades entre observações multidimensionais (SILVA; RAUBER; TELEA, 2016). Desta forma, a entrada para a técnica de projeção é uma matriz n x n contendo as similaridades entre todos os pares de dados, sendo normalmente empregada a distância Euclidiana e de cossenos.

Estes métodos são denominados Multidimensional Scaling (MDS) (TORGERSON, 1952) e mapeiam as distâncias no espaço n-dimensional em distâncias no espaço bidimensional. Tais métodos não necessitam do conhecimento da dimensionalidade original da base de dados, apenas a matriz de similaridades é suficiente, ou seja, uma vez que seja provida a matriz de distâncias desconsidera-se a ciência de quais atributos explicam estas similaridades. Porém, uma desvantagem importante é que calcular e armazenar uma matriz n x n, onde n pode tender a dezenas de milhares de instâncias, torna-se computacionalmente custoso.

Por outro lado, as técnicas baseadas em dimensões utilizam expressamente todas as dimensões de todas as instâncias de dados. Quando $n \ll m$, ou seja o número de observações é muito maior que o de dimensões, é possível ter uma boa "economia" computacional, porém, tornando-se necessário acessar diretamente os valores da base. A técnica mais conhecida desta classificação é o Principal Component Analysis (PCA) (JOLLIFFE, 2011).

4.2.2 Global vs Local

A classificação *global* vs *local* é definida de acordo com o funcionamento algorítmico das técnicas para a construção da projeção.

Métodos globais realizam um único mapeamento para todas as instâncias, considerando apenas as propriedades globais da base de dados. Por consequência, as propriedades locais são ignoradas. Exemplos clássicos de métodos globais são o PCA e MDS.

Uma vez que métodos globais utilizam uma única transformação, o maior desafio é encontrar uma função de mapeamento que, ao projetar uma base de dados complexa, preserve ao máximo suas distâncias, tornando-se esta sua principal desvantagem. Além disto esta abordagem pode requerer muito esforço computacional.

Os métodos locais, por sua vez, selecionam pequenos subconjuntos de instâncias e aplicam sobre estes métodos de projeção mais acurados. As observações restantes, próximas a cada um destes subconjuntos, são posteriormente agregadas ao redor dos mesmos. Uma das vantagens é que estes métodos não demandam muito processamento computacional, uma vez que o número de elementos em cada subgrupo é pequeno. Exemplos de métodos locais podem ser o Local Affine Multidimensional Projection (LAMP) (JOIA et al., 2011) e o Part-Linear Multidimensional Projection (PLMP) (PAULOVICH; SILVA; NONATO, 2010).

4.2.3 Preservação de distância vs Preservação de vizinhança

Este tipo de classificação é definido de acordo com a característica que se deseja preservar após a aplicação da projeção.

Quando deseja-se avaliar, com precisão, as similaridades das instâncias, técnicas de preservação de distância devem ser preferidas. No entanto, raramente é possível projetar

uma base de dados multidimensional em um espaço de baixa dimensão sem modificação de nenhuma distância (YANG, 2011). Exemplos de técnicas de preservação de distâncias são o PCA, MDS, dentre outras.

Quando a dimensionalidade da base é muito grande, as distâncias Euclidianas entre os pares de instâncias são muito similares, tornando a alta acurácia na preservação de distância sem muito valor (SILVA; RAUBER; TELEA, 2016). Neste caso, por exemplo, é mais interessante preservar a vizinhança, a fim de identificar os grupos e *outliers* presentes na base de dados multidimensional. Sendo assim, a representação dos grupos são mais perceptíveis, uma vez que há uma maior "liberdade" ao algoritmo para projetar as observações no espaço bidimensional (SILVA; RAUBER; TELEA, 2016). A técnica de preservação de vizinhança mais conhecida e melhor classificada é a t-Distributed Stochastic Neighbor Embedding (t-SNE) (MAATEN; HINTON, 2008), sendo utilizada em aplicações de aprendizado de máquina, dentre outros.

4.2.4 Linearidade vs Não-linearidade

A principal classificação entre técnicas de projeção multidimensional (redução de dimensionalidade) é a distinção entre técnicas lineares e não-lineares (MAATEN; POSTMA; HERIK, 2009).

Os métodos lineares assumem que os dados do espaço original (alta dimensão) possuem relações lineares, enquanto que os não-lineares não supõe a linearidade. Sendo assim, se a projeção é uma combinação linear das instâncias do espaço original, esta é uma técnica linear. Caso contrário, é uma técnica não-linear.

Técnicas lineares, como PCA e MDS, existem há muito tempo, mas a maioria delas só conseguem lidar com dados naturalmente lineares (TSAI; CHAN, 2007). Desta forma, métodos lineares não conseguem eficientemente projetar dados que não possuem relação linear intrínseca.

De acordo com Fang, Sakellaridi e Saad (2009) , métodos de projeção lineares podem tornar-se inadequados, uma vez que as informações significativas de baixa dimensão extraídas de dados de alta dimensão são frequentemente não-lineares. Devido a esta limitação, métodos não-lineares têm sido propostos ao longo das últimas décadas.

Ao contrário dos métodos tradicionais, as técnicas não-lineares conseguem lidar com dados de alta complexidade não linearmente correlacionados. De fato, tais métodos possuem certa vantagem, uma vez que a maioria das aplicações de tempo real geram dados não-lineares. Em adição, a partir de estudos anteriores, métodos não-lineares demonstram melhor eficiência do que os tradicionais métodos lineares em tarefas artificiais complexas (MAATEN; POSTMA; HERIK, 2009). Apesar da aparente supremacia dos métodos não-lineares, estes possuem desvantagens. De acordo com Tsai e Chan (2007), duas delas são que eles são muito tendenciosos a ruído e muitos requerem uma complexidade computacional quadrática o que os tornam praticamente inviáveis para bases de dados muito grandes.

4.2.5 Supervisão

Uma outra classificação a ser adotada é a capacidade de supervisão. Algumas projeções multidimensionais levam em consideração a informação de *label* (também conhecidos como rótulos) para realizar o mapeamento e assim agrupar instâncias pertencentes a uma mesma classe próximas as outras.

A maioria dos métodos de projeção multidimensional podem ser adaptados a fim de serem capazes de aplicarem a supervisão. Desta forma, há algumas variantes de métodos como Stochastic Neighbor Embedding (SNE) (HINTON; ROWEIS, 2003), Isomap (TENENBAUM; SILVA; LANGFORD, 2000), Locally Linear Embedding (LLE) (ROWEIS; SAUL, 2000), Least Square Projection (LSP) (PAULOVICH et al., 2008), entre outros, que atendem tal requisito.

Em teoria, métodos que conseguem lidar com dados de dissimilaridades são passíveis de serem adaptados a trabalharem com supervisão por meio da alteração da informação de dissimilaridade de acordo com a classe, por exemplo (NONATO; AUPETIT, 2018). Uma outra abordagem seria aplicar um mecanismo de aprendizagem de métricas de forma a estimar as dissimilaridades a partir de dados rotulados (*labels*).

Outras técnicas como LAMP, que utilizam informação dos coeficientes de controle no processo de mapeamento, também são adaptáveis de modo a trabalhar em uma forma supervisionada utilizando as informações de classificação nos coeficientes de controle (NONATO; AUPETIT, 2018).

4.2.6 Estabilidade

O termo estabilidade pode ter significados diversos, porém neste contexto o conceito é bem definido.

Estabilidade significa o quão sensível é o método de projeção multidimensional a variações nos dados de entrada, ou seja, a mudanças no número de instâncias, etc. Se o método é considerado estável, então pequenas variações nos dados de entrada ocasionam pequenas mudanças no espaço projetado. Em adição, métodos estáveis não podem mudar a posição das instâncias já projetadas no espaço visual na presença de novos dados de entrada.

As técnicas baseadas em decomposição de autovetores e autovalores, como PCA e LLE não são estáveis, onde mesmo pequenas variações nos dados podem ocasionar grandes impactos no espaço projetado (NONATO; AUPETIT, 2018).

A LAMP, por sua vez, consegue ser estável desde que os seus pontos de controle permaneçam fixos. Uma vez que estes pontos definem o posicionamento dos pontos no espaço, caso não mudem, o espaço projetado não muda.

4.3 TÉCNICAS DE PROJEÇÕES MULTIDIMENSIONAIS

A seguir são apresentados dois métodos de projeção multidimensional que serão utilizados nos experimentos do próximo capítulo. Sendo assim, o enfoque é direcionado para estas duas técnicas em específico.

4.3.1 Local Affine Multidimensional Projection (LAMP)

LAMP é uma técnica de projeção multidimensional que utiliza os conceitos da teoria da transformação ortogonal.

LAMP pode ser entendida como uma técnica local. Sendo assim, ela utiliza um subconjunto da base de dados, que será denominado pontos de controle, além de suas localizações no espaço visual. A informação proveniente destes pontos de controle é então utilizada para construir uma família de funções afim, uma para cada instância a ser projetada (JOIA et al., 2011). LAMP é altamente flexível, permitindo que o usuário possa manipular os pontos de controle no espaço projetado (visual), de forma a organizá-los da melhor forma possível, permitindo assim uma ótima análise exploratória.

O método LAMP pode ser definido da seguinte maneira: seja X a base de dados, com $x \in \mathbb{R}^m$, onde x são seus elementos. A partir desta base são selecionados um conjunto de k pontos de controle, denominados X_S, onde $X_S = x_1, x_2, ..., x_k$ e $X_S \subset X$. Então define-se k como muito menor que o número de instâncias de X. Em seguida seja $Y_S = y_1, y_2, ..., y_k$, a projeção dos pontos de controle X_S. Sendo assim, dada uma instância $x \in X$, LAMP mapeia x no espaço visual a partir da melhor transformação afim $f_x(p) = pM + t$ que minimiza:

$$\sum^{i} \alpha \left\| f_x(x_i) - y_i \right\|^2, \ sujeito\ a\ M^T M = I \tag{4.1}$$

onde, a matriz M e o vetor t serão determinados via otimização, I é a matriz identidade e α_i são pesos escalares definidos por:

$$\alpha_i = \left\| x_i - x \right\|^2 \tag{4.2}$$

4.3.2 Local Convex Hull (LoCH)

A técnica Local Convex Hull (LoCH) é uma técnica de projeção multidimensional especialmente voltada para espaços de alta dimensão esparsos, como por exemplo coleções de documentos (FADEL et al., 2015).

Normalmente, nestes espaços, as instâncias estão organizadas em *manifolds* locais possuindo, em sua maioria, uma alta dissimilaridade entre si (MARTIN-MERINO; MUNOZ, 2004). A LoCH é capaz de segregar grupos (*clusters*) de instâncias similares nestes espaços de uma maneira superior às técnicas globais, além de ser mais rápida que técnicas locais e de implementação mais simples (FADEL et al., 2015).

Esta técnica é constituída de três etapas: primeiramente é realizada a procura dos k vizinhos mais próximos para cada instância, posteriormente se aplica a amostragem e projeção de instâncias representativas e por último as instâncias restantes são interpoladas com base nas amostras. Além disso, na última etapa executa-se um processo iterativo a fim de posicionar o mais próximo possível as instâncias de seus vizinhos mais próximos.

O cálculo destes vizinhos mais próximos é realizado por meio de uma estratégia de agrupamento e uma vez que se têm as amostras aplica-se uma técnica global, que preserva ao máximo as distâncias originais, projetando-as no espaço.

Finalmente, realiza-se o processo chamado de aproximação do fecho convexo. Sendo $N_i = \{x_{i_1}, x_{i_2}, ..., x_{i_{k_i}}\}$ a lista do K_i vizinhos mais próximos da instância x_i, qualquer posição dentro do fecho convexo no espaço projetado, composto pelas instâncias de N_i, pode ser calculado com a seguinte formulação matemática:

$$\widehat{y}_i = \sum_{x_j \in N_i} \alpha_j y_j \tag{4.3}$$

com $\alpha_j > 0$ e $\sum \alpha_j = 1$.

4.4 MÉTRICAS DE AVALIAÇÃO

A qualidade de uma projeção multidimensional é um fator decisivo na difícil tarefa de escolher qual projeção aplicar em um determinado domínio. Uma vez que este processo não é tão simples e direto algumas métricas de avaliação foram desenvolvidas, com o intuito de prover, quantitativamente, uma medida do grau de qualidade/perfeição destas técnicas.

Na literatura, pode-se encontrar duas classes de técnicas de avaliação de projeção multidimensionais. A primeira mede a qualidade técnica quanto à preservação das similaridades entre as instâncias no espaço original e no espaço projetado. E a segunda analisa a qualidade dos *clusters* gerados pelo método, verificando se suas instâncias são da mesma classe (ELER et al., 2015).

De modo geral, a qualidade das projeções multidimensionais, em termos de preservação das similaridades entre as instâncias, é medida pelas técnicas de Stress e Preservação de Vizinhança, enquanto que a qualidade dos *clusters* é medida pelo método da Silhueta (do inglês, Coefficient Silhouette). A seguir estas três técnicas são apresentadas brevemente.

4.4.1 Preservação de Vizinhança (Neighborhood Preservation)

Seja uma base de dados $D^m = \{x_i \in \mathbb{R}\}_{1 \leq i \leq M}$ formada por M pontos m-dimensionais e uma técnica de redução de dimensionalidade representada por uma função $f : \mathbb{R}^m \to \mathbb{R}^p$ que mapeia cada $x_i \in D^m$ em um ponto $y_i \in D^p$.

A técnica de preservação de vizinhança objetiva analisar como os k-vizinhos mais próximos são afetados pela projeção e idealmente deve explorar 2 casos: preservação da vizinhança de D^m e a confiabilidade de D^p.

Na preservação da vizinhança é verificado se pontos vizinhos de D^m são projetados como vizinhos em D^p. Já na confiabilidade é verificado se pontos vizinhos em D^p também são vizinhos em D^m. De forma geral, a preservação de vizinhança avalia a quantidade de pontos vizinhos, presentes no espaço multidimensional original, que são mantidos como tais no espaço projetado, após o mapeamento. Matematicamente o cálculo deste índice é dado pela equação:

$$NP_k = \frac{1}{n} \sum_{n}^{1} \frac{\left| N_{k_i}^m \bigcap N_{k_i}^p \right|}{k}, \tag{4.4}$$

sendo $N_{k_i}^m$ os k vizinhos mais próximos de x_i no espaço m-dimensional (original) e $N_{k_i}^p$ os k vizinhos mais próximos de x_i no espaço p-dimensional (projetado). A variação de NP_k ocorre no intervalo $[0, 1]$, sendo que quanto maior este valor maior o grau de preservação da vizinhança.

Em termos mais específicos, ao se analisar a preservação de vizinhança de um dado ponto i e considerando os k vizinhos do mesmo como a lista $v_k(i)$, é necessário identificar três conjuntos para este ponto:

- **vizinhos ausentes** - $\{x_j \in v_k^m(i) \wedge y_j \notin v_k^p(i)\}$. Representam pontos presentes no espaço multidimensional original que foram considerados como não importantes. Desta forma foram descartados e não são encontrados ao se analisar a vizinhança do ponto i.

- **vizinhos falsos** - $\{x_j \notin v_k^m(i) \wedge y_j \in v_k^p(i)\}$. Representam pontos originalmente distantes do ponto i no espaço de alta dimensão, mas devido ao mapeamento foram posicionados próximos ao ponto i.

- **vizinhos verdadeiros** - $\{x_j \in v_k^m(i) \wedge y_j \in v_k^p(i)\}$. Representam pontos originalmente próximos no espaço de alta dimensão que são mantidos próximos no espaço de baixa dimensão após o mapeamento.

4.4.2 Stress

O Stress é uma outra medida que visa avaliar, de forma objetiva, o quão bem a projeção multidimensional preserva as características do espaço multidimensional original. Entende-se por forma objetiva, o descarte de análises unicamente visuais, do espaço projetado, tomando como base critérios subjetivos.

O stress busca quantificar o quão bem as distâncias calculadas no espaço original são preservadas após o mapeamento no espaço projetado. Matematicamente, a função stress é representada por:

$$S_1 = \sqrt{\frac{\sum_{i<j}(\delta(x_i, x_j) - d(y_i, y_j))^2}{\sum_{i<j}\delta(x_i, x_j)^2}} \tag{4.5}$$

onde $\delta(x_i, x_j)$ é a distância entre i e j no espaço original e $d(y_i, y_j)$ é a distância entre i e j no espaço transformado. O valor de stress ocorre no intervalo $[0, 1]$, sendo que quanto menor este valor mais "perfeita" é a representação, em termos de preservação das distâncias do espaço original.

4.4.3 Coeficiente de Silhueta

Apenas tomando como base uma análise visual da projeção multidimensional não é possível determinar qual método produz melhores *clusters*. Para isto, é necessário utilizar o método da silhueta.

O método da Silhueta, tem por objetivo medir a coesão e separação entre instâncias dos *clusters* criados pela projeção (ELER et al., 2015). Dado uma instância d_i, a sua

coesão a_i é definida como a média das distâncias entre a mesma e as demais instâncias presentes no seu *cluster*. Por outro lado, a separação b_i é definida como a distância mínima entre d_i e todas as outras instâncias pertencentes aos demais *clusters*. Sendo assim, o Coeficiente de Silhueta é representado, matematicamente, pela seguinte equação:

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{(b_i - a_i)}{max(a_i, b_i)} \tag{4.6}$$

O valor de S pode variar no intervalo $-1 \leq S \leq 1$. Quanto maior o valor, melhor a coesão e separação entre os grupos da projeção.

4.5 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentadas algumas definições a respeito das projeções multidimensionais, abordando seus conceitos norteadores, além de apresentar de forma clara como tais técnicas superam as limitações enfrentadas pelas técnicas mais tradicionais vistas no Capítulo 3.

Além disso, na Seção 4.2 deste capítulo, foram apresentadas diversas possibilidades de classificação dos métodos de projeção dimensional como por exemplo, dimensão vs distância, global vs local, entre outras.

As técnicas LAMP e LoCH foram abordadas estrategicamente devido ao fato de serem utilizadas na execução dos experimentos deste trabalho. Por fim, foram abordas as três métricas de avaliação também utilizadas nos experimentos. Em adição, foi também apresentada as métricas de avaliação de projeções multidimensionais, onde foram citadas as três principais: stress, preservação de vizinhança e coeficiente de silhueta, cada uma delas para um tipo específico de análise.

Capítulo 5

RESULTADOS EXPERIMENTAIS

5.1 CONSIDERAÇÕES INICIAIS

Normalmente no processo de desenvolvimento e validação de uma nova técnica de projeção multidimensional, costuma-se avaliar o método no intuito de verificar sua plausibilidade, montando assim um ambiente de testes. Tal informação também é válida quando deseja-se comparar projeções par-a-par evidenciando qual possui melhor acurácia, por exemplo.

Para isto, as projeções são sensibilizadas com diferentes configurações (instâncias, dimensões, grupos) de bases de dados, sendo posteriormente aplicadas métricas de avaliação. Atualmente, na literatura, como visto nos trabalhos de Joia et al. (2011), Martin, Minghim e Telea (2015) e Martins et al (2014), a qualidade da redução de dimensionalidade é medida pelo stress e preservação de vizinhança, enquanto que a qualidade dos grupos é medida pela silhueta (ELER et al., 2015). Sendo assim, os experimentos aqui realizados seguiram as recomendações e padrões já observados na literatura, o que justifica a escolha destas três técnicas, além da criação de diferentes bases de dados.

Este capítulo apresenta todo o processo de condução dos experimentos realizados, metodologia e resultados finais utilizando duas técnicas de projeção multidimensionais, a saber, LAMP (JOIA et al., 2011) e LoCH (FADEL et al., 2015). Além disso, foi utilizado o método do fatorial completo 2^k, método este proveniente da avaliação de desempenho, que serviu como aparato para execução e análise dos experimentos.

5.2 EXPERIMENTOS

Esta seção e por conseguinte suas subseções, descrevem todo o processo de execução dos experimentos, abordando a ferramenta VisPipeline (abordada na próxima seção), por meio da qual foi possível a análise dos dados multidimensionais, as bases de dados utilizadas como *inputs* de cada projeção, as métricas de avaliação das projeções, o modelo experimental utilizado e o planejamento e execução dos experimentos.

5.2.1 VisPipeline

A ferramenta VisPipeline, abordada por Neves et al. (2015) é uma ferramenta de visualização de dados, escrito em Java, desenvolvida no Instituto de Ciências e de Matemática da USP (ICMC-USP). Por meio dela foi possível analisar as bases de dados, realizar as projeções e aplicar as métricas de avaliação nas técnicas de projeções multidimensionais.

O Vispipeline consiste de um conjunto de vários componentes organizados de tal forma a permitir a comunicação entre os mesmos. Baseado na flexibilidade e na facilidade de construções de visualizações, os componentes podem ser vistos como *pipelines*, sendo cada um deles uma função bem definida. Além disso, tais componentes podem estar relacionados as etapas do *pipeline* de visualização, seja alterando o tipo de entrada, a escolha do tipo de distância, método de avaliação, entre outros.

Uma das principais características do VisPipeline é a possibilidade de criação e execução de diversos cenários de análise, seja por meio dos componentes já presentes na ferramenta ou de customizações e/ou desenvolvimento de novos por usuários especialistas.

5.2.2 Bases de dados

Para a realização dos experimentos foram criadas um total de 8 bases sintéticas de diferentes configurações com relação ao número de instâncias, dimensões e *clusters* (grupos), por meio do software Elki Data Mining (SCHUBERT et al., 2015). Cada base foi gerada por meio da semente de valor 428956419. O número de instâncias por grupo, o número de dimensões e o número de grupos de cada base são, respectivamente, 1000 ou 10000, 10 ou 100 e 3 ou 5. A seguir a Tabela 5.1 sumariza as configurações de cada uma das 8 bases.

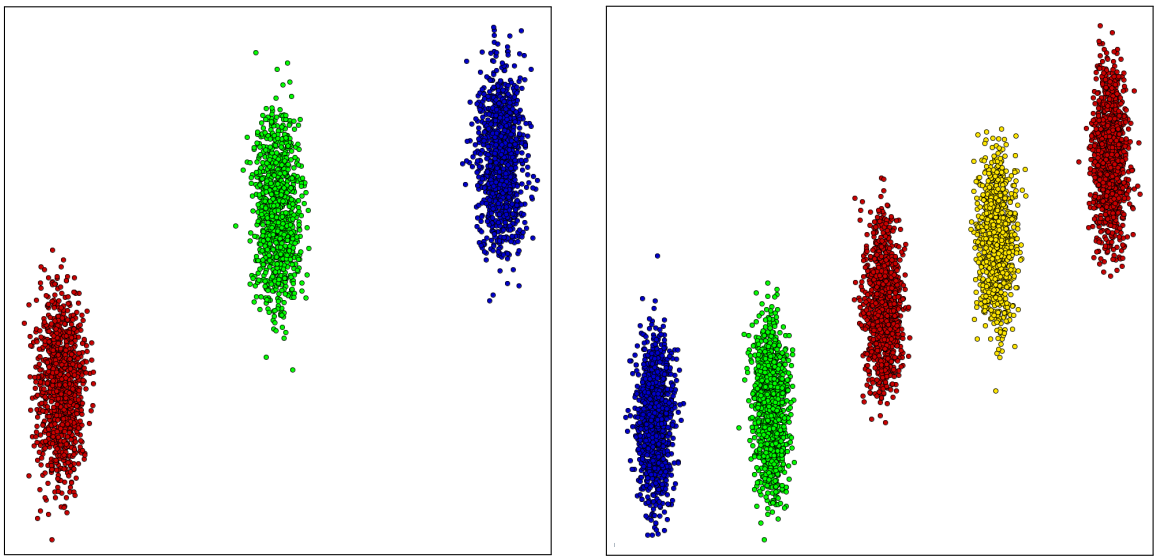
Tabela 5.1: Configurações das bases de Dados.

Base	**Configurações**		
	Inst. por Clusters [1]	**Dimensões** [2]	**Clusters** [1]
1	1k	10	3
2	1k	10	5
3	1k	100	3
4	1k	100	5
5	10k	10	3
6	10k	10	5
7	10k	100	3
8	10k	100	5

[1] Número de Instâncias por Cluster, onde $1k = 1000$, [2] Número de Dimensões, [3] Número de Clusters

Para fins de ilustração gerou-se a Figura 5.1 a partir de outra ferramenta de análise

de dados conhecida como Projection Explorer (PEx) também discutida e apresentada em Neves et al. (2015). Na Figura 5.1 é possível perceber a representação visual das bases de dados 1 e 2 (na Tabela 5.1), onde observa-se claramente a formação de três grupos distintos em 5.1 (a) e cinco grupos em 5.1 (b). As demais bases aqui não representadas possuem, em termos de grupos, características semelhantes.



(a) 1k de instâncias, 10 dimensões e 3 grupos (b) 1k de instâncias, 10 dimensões e 5 grupos

Figura 5.1: Bases de dados 1 e 2

5.2.3 Projeções Multidimensionais e Métricas de Avaliação

Para a realização dos experimentos foram escolhidas duas técnicas de projeção multidimensional: LAMP e LoCH. Tais técnicas foram selecionadas estrategicamente devido a possuírem algumas características semelhantes como por exemplo, são técnicas locais, permitem a supervisão, etc . Para avaliação da qualidade das mesmas, três métricas de avaliação foram designadas: stress, silhueta e preservação de vizinhança. Tais métricas já foram abordadas no Capítulo 4, mas recaptulando:

- **Stress**: Tem por intuito medir o quanto de informação foi perdida durante a projeção. O stress pode variar entre 0 e 1, sendo que quanto menor o valor, menor é a perda de informação.
- **Silhueta**: Tem por intuito medir a consistência dos grupos. Seu valor varia entre -1 e 1. Quanto maior o valor melhor a coesão e separação entre os grupos.
- **Preservação de Vizinhança**: Avalia a projeção em termos de preservação da vizinhança. Seu valor varia entre 0 e 1, sendo maior o valor, melhor a preservação de vizinhança.

5.2.4 Método Experimental do Fatorial Completo 2^k

O método utilizado na condução dos experimentos foi o método do *Fatorial Completo*. Tal técnica se caracteriza por utilizar todas as possibilidades de combinações em todos os níveis de todos os fatores (JAIN, 1990).

Conceituando a terminologia até aqui apresentada, define-se um fator como uma variável que afeta/influencia a variável de resposta e os níveis como os valores que um fator pode assumir. Por exemplo, em um caso hipotético, onde a variável de resposta é o tempo de execução de um algoritmo qualquer, pode-se adotar como fatores que influenciam diretamente este tempo o número de ciclos do processador e a quantidade de memória RAM. Supondo que para o experimento há disponibilidade de apenas duas configurações de computadores: o primeiro possui um processador com *clock* de 1.8 GHz e 4 GB de memória RAM e o segundo possui um processador com *clock* de 3.2 GHz e 8 GB de memória RAM. Neste cenário, pode-se definir os fatores como *clock* do processador, possuindo dois níveis: 1.8 GHz ou 3.2 GHz, e quantidade de memória RAM, com seus dois níveis: 4 GB ou 8 GB.

Matematicamente o método do fatorial completo pode ser descrito pela seguinte equação:

$$n = \prod_{i=1}^{k} n_i \tag{5.1}$$

onde o número de experimentos n, é determinado a partir de k fatores, onde o i-ésimo fator tem n_i níveis.

A vantagem principal do fatorial completo é que todas as possibilidades de configuração do sistema é verificada, permitindo dessa forma uma análise de cada fator e sua relação com fatores secundários. A maior dificuldade desta técnica é justamente o alto custo da avaliação, tanto em termos de tempo como recursos financeiros, principalmente quando tem-se a possibilidade de repetição do experimento várias vezes. Porém, nos experimentos aqui realizados, foi adotada uma variante do fatorial completo que é o modelo 2^k, onde cada fator tem apenas 2 níveis.

É de suma importância ressaltar que este é um método advindo da avaliação de desempenho, mas o mesmo apenas foi utilizado como aparato/arcabouço para a execução/análise experimental. Desta forma, o intuito não é realizar uma avaliação de desempenho, mas apenas utilizar tal método como planejamento experimental.

5.2.5 Planejamento dos Experimentos

Os experimentos foram conduzidos por meio do método do fatorial completo 2^k, discutido na seção anterior, onde para cada fator k há 2 níveis. Neste trabalho foram considerados 4 fatores, mostrados na Tabela 5.2 com seus respectivos níveis.

Os níveis do fator (A) foram definidos de acordo com as técnicas de projeções definidas para experimentação e para os fatores (B), (C) e (D), os níveis foram definidos de acordo com a configuração de cada base de dados, já demonstrada previamente na Tabela 5.1.

A partir do planejamento fatorial 2^k, levando em consideração os devidos fatores e

Tabela 5.2: Fatores e níveis.

Fator	**Níveis**	
	1	**2**
Algoritmo (A) *	LoCH	LAMP
Clusters (B) **	3	5
Dimensões (C) **	10	100
Instâncias (D) **	1k	10k

* Projeção Utilizada, ** Quantidade associada ao respectivo fator.

níveis, se construiu a Tabela 5.3 que representa os 16 cenários possíveis desta experimentação.

Tabela 5.3: Experimentos.

Nº	**A: Algoritmo**	**B: Clusters**	**C: Dimensões**	**D: Instâncias** *
1	LAMP	3	10	1k
2	LAMP	3	10	10k
3	LAMP	3	100	1k
4	LAMP	3	100	10k
5	LAMP	5	10	1k
6	LAMP	5	10	10k
7	LAMP	5	100	1k
8	LAMP	5	100	10k
9	LoCH	3	10	1k
10	LoCH	3	10	10k
11	LoCH	3	100	1k
12	LoCH	3	100	10k
13	LoCH	5	10	1k
14	LoCH	5	10	10k
15	LoCH	5	100	1k
16	LoCH	5	100	10k

* k representa unidade do milhar. $1k = 1000$.

Todos os experimentos foram avaliados tomando como base um intervalo de confiança de 95% ($\alpha = 0,05\%$) obtido a partir da tabela de distribuição t-student. Para cada experimento houve 10 replicações. Além disso, foram definidas oito variáveis de respostas, a saber:

- **Proj-Tempo**: O tempo da projeção.
- **Stress-Valor**: O valor do cálculo do stress da projeção.
- **Stress-Tempo**: O tempo gasto para o cálculo do stress.
- **Silhueta-Valor**: O valor do cálculo da silhueta da projeção.
- **Silhueta-Tempo**: O tempo gasto para o cálculo da silhueta.
- **Preser.Viz-Valor**: O valor do cálculo da preservação de vizinhança.
- **Preser.Viz-Tempo**: O tempo gasto para o cálculo da preservação de vizinhança.
- **Tempo-Total**: O tempo total da replicação somando-se os tempos anteriores (*tempo da projeção + tempo do stress + tempo da silhueta + tempo da preservação de vizinhança*).

5.3 ANÁLISE DOS RESULTADOS

Na análise dos resultados foram gerados gráficos de pareto e gráficos normais dos efeitos (através do software Minitab), onde é possível verificar qual o grau de influência de cada um dos fatores nas variáveis de resposta. Nas seções a seguir são analisadas cada uma das oito variáveis.

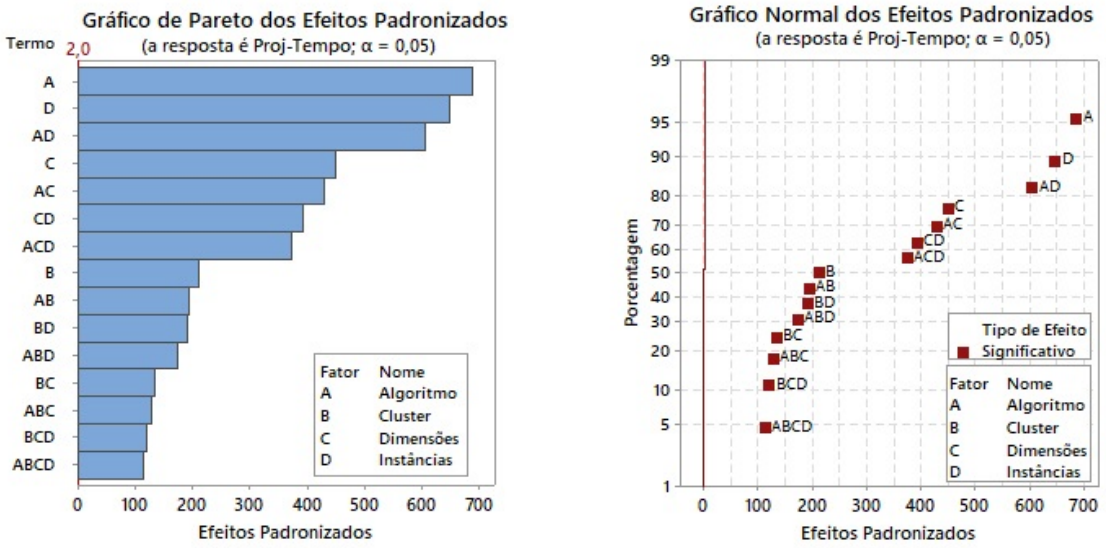
5.3.1 Tempo de Projeção

A Figura 5.2 (a) mostra o gráfico de pareto dos efeitos para o projeto fatorial 2^k, onde é possível perceber o grau de influência de cada um dos fatores na variável de resposta Proj-Tempo. É evidente que os efeitos mais significativos são o Algoritmo (A), Instâncias (D), a interação entre os fatores A e D, e Dimensões (C).

Isto de fato já era esperado uma vez que a complexidade do algoritmo de projeção tem influência direta no tempo da projeção. A complexidade da técnica LAMP é $O(pn)$ e no caso do método LoCH é $O(n\sqrt{n})$, onde p e n são o número de dimensões e instâncias, respectivamente. Sendo assim, corrobora-se a análise inferida da Figura 5.2 (a), onde basicamente o fator com maior influência no tempo da projeção é o algoritmo, sendo sua complexidade determinada pelo número de instâncias e/ou dimensões.

Enquanto que o gráfico de pareto informa o grau de influência dos fatores em termos absolutos, o gráfico normalizado mostra a magnitude, a importância e direção dos efeitos. Os pontos mais afastados do valor zero denotam significância estatística. Pontos exatamente em cima da linha de referência vermelha são não significativos. Além disso, os pontos mais à esquerda revelam um efeito padronizado negativo, enquanto que os mais à direita revelam um efeito positivo. Por exemplo, fatores localizados mais à esquerda de zero, indicam que a medida que crescem, o valor da variável de resposta diminui, enquanto que pontos à direita, à medida que crescem, aumentam a resposta.

Através da análise da Figura 5.2 (b) é possível perceber que nenhum fator tem efeito padronizado negativo, visto que todos estão à direita do valor zero, ou seja, todos contribuem para o aumento da variável de resposta (Proj-Tempo) a medida que crescem.



(a) (b)

Figura 5.2: Influência dos fatores em Proj-Tempo

5.3.2 Stress

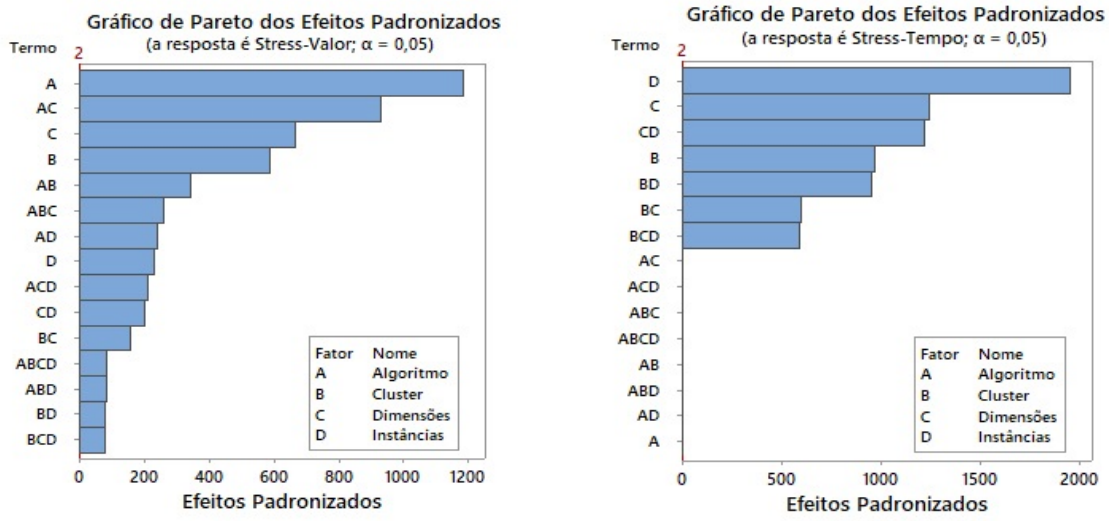
Duas outras variáveis de resposta analisadas nos experimentos foram o Stress-Valor e o Stress-Tempo da projeção. As Figuras 5.3 (a) e 5.3 (b), mostram o gráfico de pareto onde é possível avaliar o grau de influência de cada um dos fatores.

Observando a Figura 5.3 (a), percebe-se que no caso do valor do stress, o fator que mais influencia é o Algoritmo (A). Uma vez que o stress é uma métrica de qualidade de mensuração de perda de informação (neste caso distância) no processo de projeção, tal valor é variável de acordo com o algoritmo utilizado, pois ele pode ou não preservar, de forma ideal, as distâncias presentes no espaço original. Por exemplo, para a mesma configuração de base (1k de instâncias por *cluster*, 10 dimensões e 3 *clusters*) a técnica LoCH obteve um valor médio de stress de 0.038612524 enquanto que o LAMP obteve um valor de 0.004843758, ou seja, uma menor perda. Sendo assim, o que mais influencia este valor é a técnica de projeção multidimensional utilizada.

Por outro lado, a partir da Figura 5.3 (b), percebe-se que o maior grau de influência no tempo de cálculo do stress é o número de instâncias. Esta afirmação faz sentido uma vez que ao analisar novamente a equação do cálculo do stress:

$$S_1 = \sqrt{\frac{\sum_{i<j}(\delta(x_i, x_j) - d(y_i, y_j))^2}{\sum_{i<j}\delta(x_i, x_j)^2}} \tag{5.2}$$

onde $\delta(x_i, x_j)$ é a distância entre i e j no espaço original e $d(y_i, y_j)$ é a distância entre i e j no espaço transformado, é necessário computar para todos os pares de instâncias



(a) (b)

Figura 5.3: Gráficos de Pareto: Influência dos fatores em Stress-Valor e Stress-Tempo.

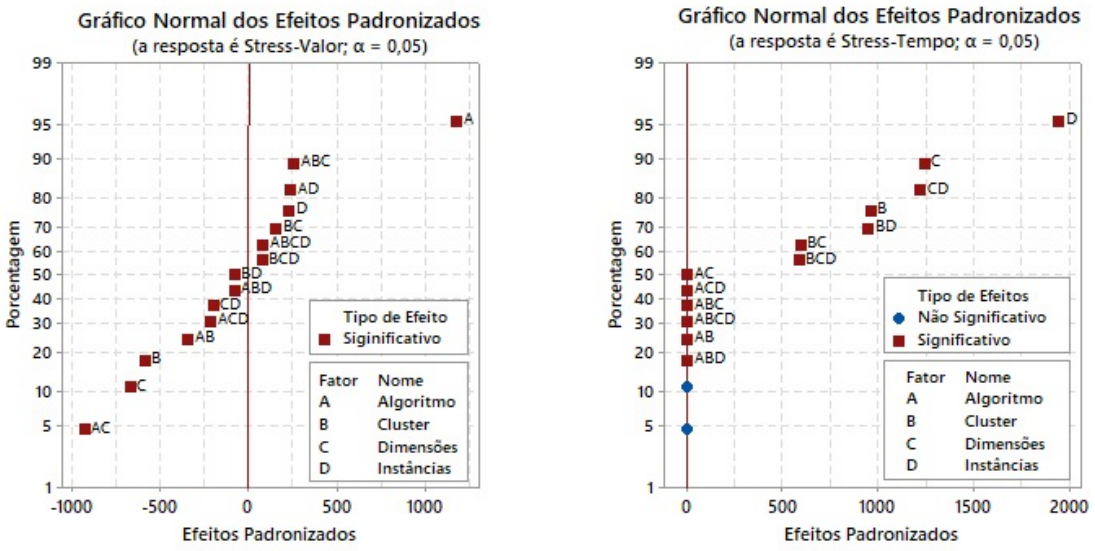
as distâncias entre as mesmas tanto no espaço original como no espaço projetado. A medida que o número de instâncias cresce o número de cálculos de distância também cresce proporcionalmente, o que justifica um acréscimo de tempo.

Já na Figura 5.4 (a) é possível inferir que o fator Dimensões (C), tem um efeito negativo. Isto quer dizer que a medida que a dimensionalidade cresce o valor do stress diminui, de acordo com o que foi observado por Sturrock e Rocha (2000). Em contrapartida, ao analisar a Figura 5.4 (b) observa-se que o fator Instâncias (D) tem um efeito positivo, à medida que cresce aumenta a variável de resposta Stress-Tempo, como mencionado anteriormente, pois está à direita do valor zero.

5.3.3 Silhueta

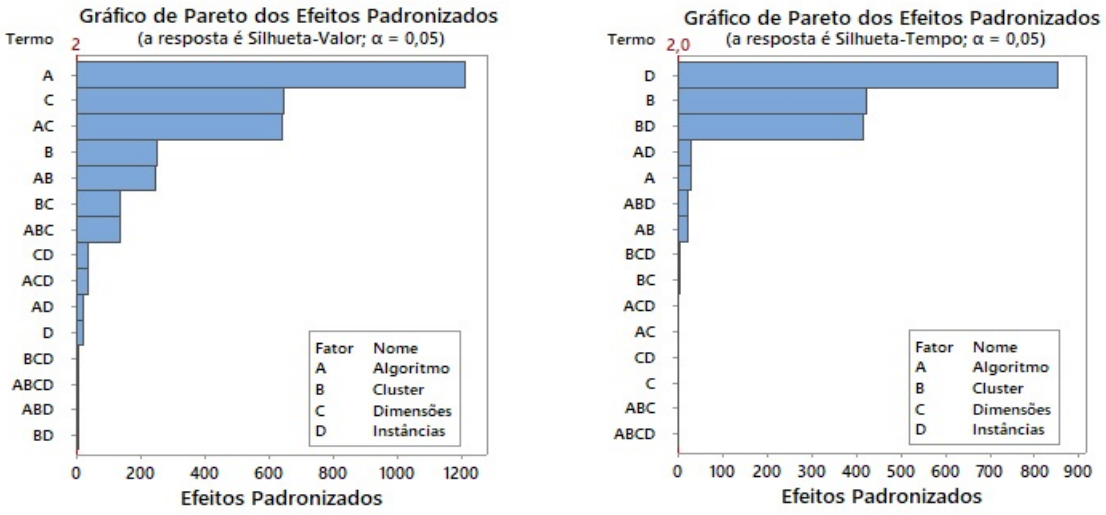
Ainda analisando variáveis de respostas relacionadas as métricas de avaliação, nesta seção apresenta-se os gráficos normais e gráficos de pareto para as variáveis Silhueta-Valor e Silhueta-Tempo.

A partir da Figura 5.5 (a) infere-se novamente que o fator que influencia em maior grau o valor da silhueta é o algoritmo, uma vez que seu cálculo utiliza informações de distância, que podem ser preservadas de forma fiel ou não a depender da técnica de projeção utilizada. Além desta, uma outra conclusão pode ser tirada a partir da Figura 5.6 (a): o fator Dimensões (C) tem um efeito padronizado positivo na variável de resposta, ou seja, a medida que o número de dimensões aumenta o valor da silhueta também aumenta. Isto já era esperado, pois a medida que a dimensionalidade aumenta, a representação dos grupos é melhor, uma vez que a projeção terá maior "liberdade" para posicionar



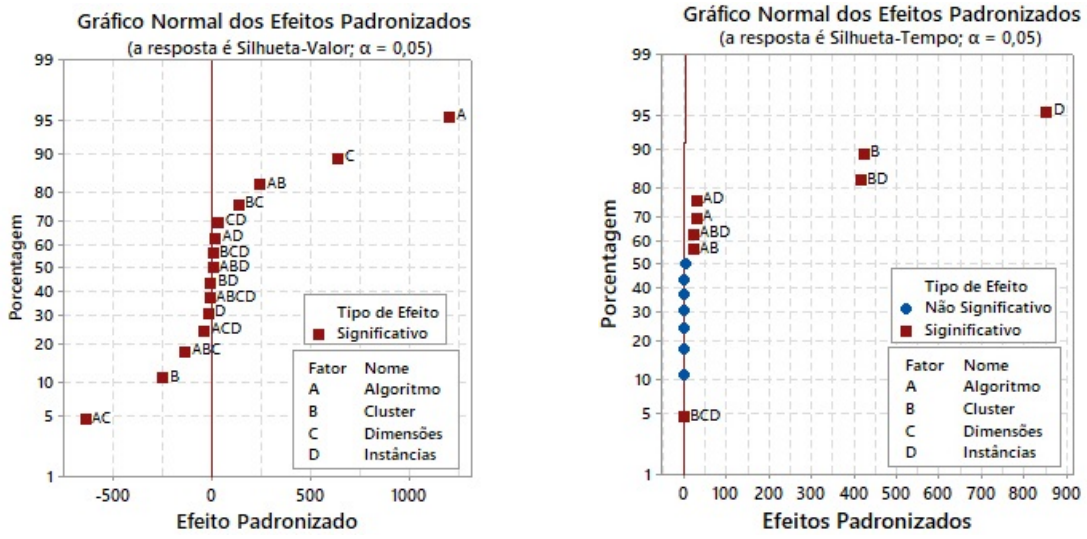
(a) (b)

Figura 5.4: Gráficos Normais: Influência dos fatores em Stress-Valor e Stress-Tempo.



(a) (b)

Figura 5.5: Gráficos de Pareto: Influência dos fatores em Silhueta-Valor e Silhueta-Tempo.



(a) (b)

Figura 5.6: Gráficos Normais: Influência dos fatores em Silhueta-Valor e Silhueta-Tempo.

as instâncias no espaço de baixa de dimensão, corroborando com Silva, Rauber e Telea (2016).

Em contrapartida, nas Figuras 5.5 (b) e 5.6 (b) é possível perceber que o fator com maior efeito no tempo de cálculo da silhueta é o número de instâncias (D), sendo o número de dimensões (C), pouco significativo para esta variável de resposta. Isto é facilmente explicado observando-se sua formulação matemática:

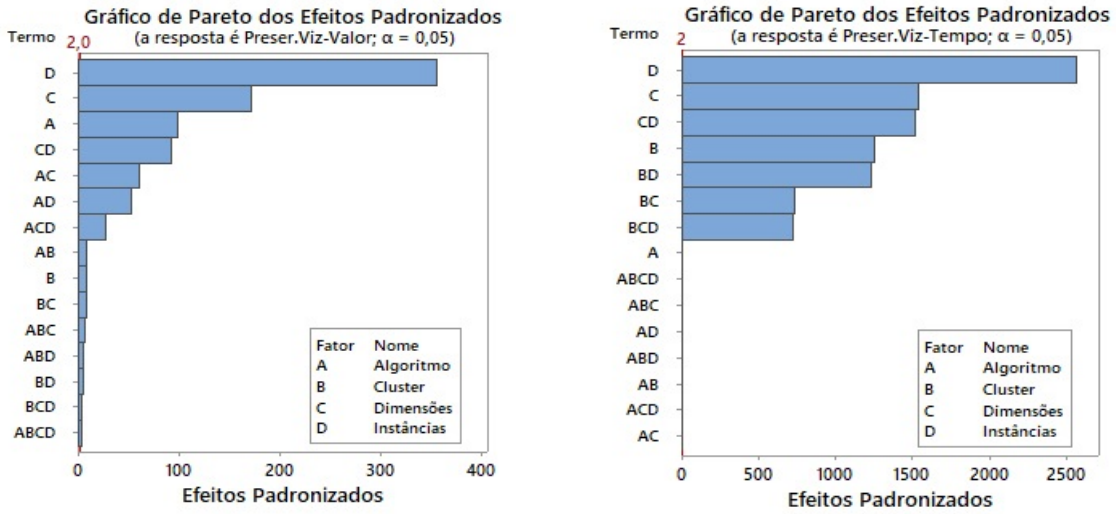
$$S = \frac{1}{n}\sum_{i=1}^{n}\frac{(b_i - a_i)}{max(a_i, b_i)} \tag{5·3}$$

Dado uma instância d_i, a sua coesão a_i é definida como a média das distâncias entre a mesma e as demais instâncias presentes no seu *cluster*. Por outro, lado a separação b_i é definida como a distância mínima entre d_i e todas as outras instâncias pertencentes aos demais *clusters*. Sendo assim, novamente, quanto maior o número de instâncias presentes na base de dados e por conseguinte em cada *cluster*, mais cálculos são necessários para calcular-se a silhueta da projeção, o que significa maior tempo de execução.

Além disto, a partir da Figura 5.6 (b) é possível verificar que sete fatores e/ou combinação de fatores (C, BC, CD, AC, ABC, ACD e ABCD) não possuem significância estatística para a variável de resposta Silhueta-Tempo, sendo todos relacionados ao número de dimensões (C). Em adição, o algoritmo (A) também não contribui em nada para este tempo do cálculo da silhueta.

5.3.4 Preservação de Vizinhança

Duas outras variáveis de respostas relacionadas a métricas de avaliação da projeção foram o Preser.Viz-Valor e Preser.Viz-Tempo.



(a) (b)

Figura 5.7: Gráficos de Pareto: Influência dos fatores em Preser.Viz-Valor e Preser.Viz-Tempo.

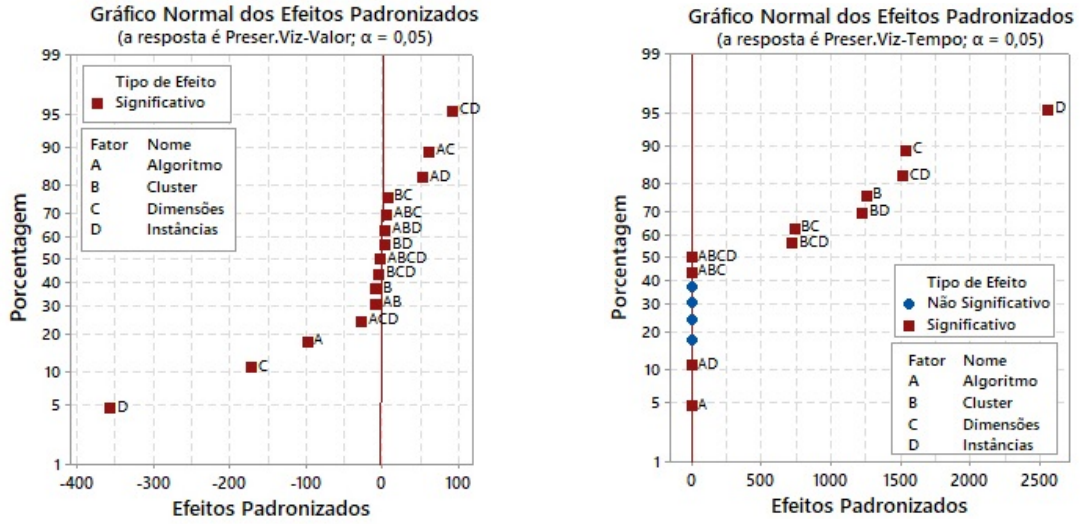
Nas Figuras 5.7 (a) e 5.8 (a), observa-se que a maior influência para o valor da preservação de vizinhança agora é advinda do fator Instâncias (D), estando este fator ao lado esquerdo da normal. A informação que é levada em consideração neste cálculo são os k vizinhos (instâncias) mais próximos de cada instância no espaço original e espaço projetado. Sendo assim o fator mais influenciador no valor da preservação de vizinhança são as próprias instâncias.

Já nas Figuras 5.7 (b) e 5.8 (b), este resultado permanece. O tempo de cálculo desta métrica é influenciado em maior escala pelo número de instâncias (D). Quanto maior o número de instâncias maior o tempo requerido.

5.3.5 Tempo Total por Experimento

A última variável a ser analisada é o Tempo-Total, que como já explicitado anteriormente é a soma dos tempos da projeção, cálculo do stress, silhueta e preservação de vizinhança.

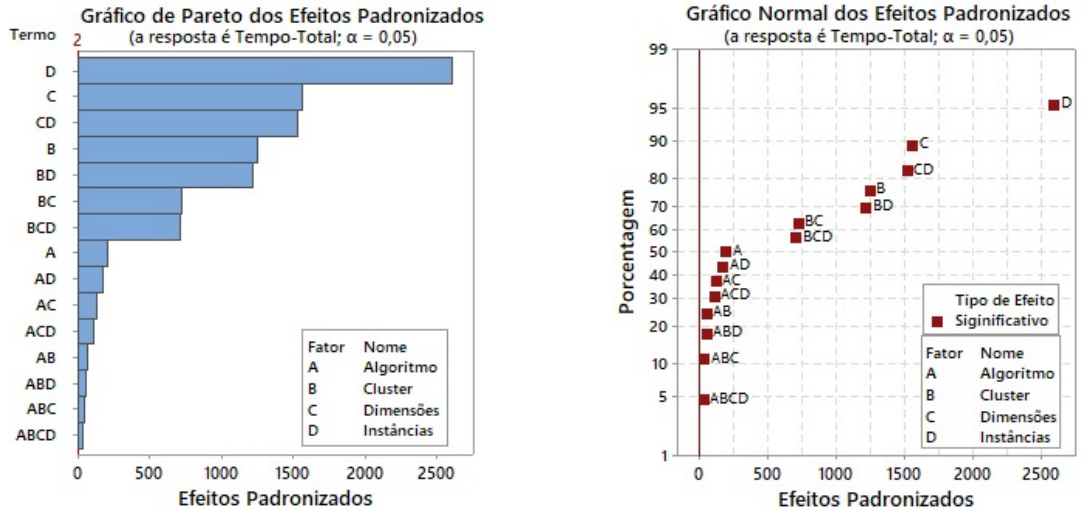
Ao observar as Figuras 5.9 (a) e 5.9 (b), é claro que o número de instâncias (D) é o fator de maior influência no tempo total do experimento, uma vez que todos os demais tempos são influenciados, em maior grau, pelo mesmo. Além disso, dois outros fatores são significativos estatisticamente: o número de dimensões (C) e em seguida o número de



(a) (b)

Figura 5.8: Gráficos Normais: Influência dos fatores em Preser.Viz-Valor e Preser.Viz-Tempo.

clusters (B). Por outro lado, o algoritmo (A), praticamente não tem influência no tempo total de projeção.



(a) (b)

Figura 5.9: Influência dos fatores em Tempo-Total.

5.4 DISCUSSÃO

A partir dos resultados apresentados na Seção 5.3, é possível definir algumas diretrizes para o processo de execução e avaliação de projeções multidimensionais.

Ficou evidente que o fator determinante para o tempo de projeção é justamente o algoritmo, o que está fortemente associado à sua complexidade. Para a grande maioria das projeções a complexidade é definida em termos de número de instâncias e/ou dimensões. Sendo assim, este fator deve ser levado em consideração na decisão sobre qual projeção utilizar e consecutivamente qual sua complexidade computacional, principalmente quando o número de instância ou dimensões é bem alto.

Um outro ponto a se analisar é a escolha de qual métrica de avaliação escolher. Como normalmente deseja-se saber a qualidade da projeção, principalmente no desenvolvimento de novas técnicas, aplica-se com frequência, alguma métrica de avaliação da técnica. A decisão de qual métrica utilizar depende do objetivo de quem analisa a projeção. Por exemplo, se o interesse é obter informações sobre a preservação de distâncias, utiliza-se o stress, se o intuito é avaliar a qualidade dos *clusters* utiliza-se a silhueta.

No entanto, apenas uma métrica de avaliação não é suficiente para emitir um parecer geral sobre a qualidade da projeção. Por exemplo, o stress, como discutido por Paulovich et al. (2008), não consegue assegurar a capacidade da técnica de projeção de preservar as relações de vizinhança, pois ela só considera relações de distância. Sendo assim, é comum empregar mais de uma métrica de avaliação. Porém, aplicar várias métricas requer tempo e processamento, os quais, em muitas ocasiões, não são recursos tão disponíveis, sendo estes fatores limitantes.

Como visto na Seção 5.3 deste capítulo, para todas as métricas o fator mais influente em seus tempos de cálculo foi o número de instâncias, seguido do número de dimensões. Apenas no caso da silhueta o 2º fator mais relevante foi o número de *clusters*, o que faz sentido pois a silhueta considera informações de localidade dos grupos. Desta forma, chega-se à seguinte conclusão: a escolha de uma métrica em detrimento de outra vai também estar baseada, dentre outros fatores, na configuração da base de dados. Por exemplo, supondo um cenário em que não há tempo suficiente para aplicar as três métricas de avaliação citadas e que a base de dados a ser testada possui milhares de instâncias e dimensões é de se esperar, de acordo com os experimentos realizados e a literatura previamente citada, que a projeção tenha um baixo valor de stress. Desta forma, o cálculo do stress para este cenário pode ser descartado, tendo a opção de investir em outra métrica como a preservação de vizinhança, por exemplo. Porém, a escolha não deve nem pode ser arbitrária, mas estar baseada no objetivo primordial da avaliação.

Por fim, a partir dos conhecimentos adquiridos no Capítulo 4 e nas análises realizadas neste capítulo, pode-se definir as seguintes diretrizes:

- **Escolha da técnica de projeção:** Deve estar baseada nas características dos dados de entrada, se será uma matriz de distâncias ou pontos, se os dados possuem relações lineares ou não, dentre outros quesitos, mas também no número de instâncias e/ou dimensões, pois são quantias que determinarão o tempo de execução da técnica por meio de sua complexidade.

- **Escolha da métrica de avaliação:** Deve ser levado em consideração o que deseja-se verificar no espaço projetado com relação ao espaço original: as relações entre as instâncias (distâncias e vizinhança) ou entre os grupos (*clusters*). Deve-se estar ciente de que a dimensionalidade pode afetar diretamente o valor do stress e silhueta. Além disso, novamente, o tempo de execução de cada métrica será determinado principalmente pelo número de instâncias (em maior ênfase) e dimensões no caso de stress e preservação de vizinhança e número de instâncias e grupos no caso da silhueta. Caso haja limitação de tempo para a execução dos experimentos, a decisão de qual métrica escolher fica a cargo do analisador, sempre observando as características da configuração da base de dados de entrada.

5.5 CONSIDERAÇÕES FINAIS

Neste capítulo foram descritos os experimentos realizados com base no modelo do fatorial completo 2^k evidenciando toda a metodologia e execução, além de apresentar todo o processo de escolha dos k fatores utilizados no modelo. Nos experimentos quatro fatores ($k = 4$) foram definidos: o algoritmo, stress, silhueta e preservação de vizinhança, sendo estes três últimos métricas de avaliação de projeções multidimensionais. Para cada fator foram definidos dois níveis (seguindo o modelo 2^k), onde no caso das métricas foram seus valores e tempos de execução e no caso do algoritmo duas projeções multidimensionais: LAMP e LoCH. Além disso, para o processo de projeção, utilizou-se diferentes configurações de bases de dados, variando o número de instâncias, dimensões e grupos.

Por fim, com base nos experimentos realizados, foi possível apresentar algumas diretrizes no processo de análise e avaliação de projeções multidimensionais e suas métricas. Ficou evidente que o número de instâncias, seguido do número de dimensões têm maior influência no tempo de execução da técnica de projeção e das métricas stress e preservação de vizinhança. Somente no caso da silhueta o 2º fator mais influenciador é o número de grupos. Além disso, foi possível perceber que o aumento da dimensionalidade proporcionou menores valores de stress e maiores valores de silhueta.

Tais conclusões têm grande relevância, visto que serve de guia para o processo de análise de projeções multidimensionais e suas métricas de avaliação. A importância destas diretrizes é mais facilmente vista quando o tempo é, por exemplo, um fator limitador à execução dos experimentos. De posse de tais conclusões é possível poupar tempo e esforço tendo ciência da complexidade da projeção multidimensional escolhida e como a configuração de cada base de dados utilizada irá afetar no valor de cada uma das três métricas de avaliação, bem como em seus tempos de execução.

Capítulo 6

CONCLUSÕES

Este Trabalho de Conclusão de Curso apresentou uma nova abordagem para avaliação de projeções multidimensionais considerando diferentes configurações de bases de dados.

Como primeiro passo realizou-se uma Revisão Sistemática da Literatura sobre a avaliação de projeções multidimensionais, o que possibilitou entender qual é a motivação do uso destas métricas, bem como perceber as técnicas mais utilizadas na literatura. Todo o conhecimento adquirido na Revisão Sistemática, foi aplicado na realização, condução e análises dos experimentos.

No Capítulo 3 algumas técnicas mais tradicionais de visualização de dados multidimensionais foram abordadas expondo suas características e limitações, servindo estas de motivação para o estudo das técnicas de projeção multidimensional, no Capítulo 4, percebendo assim mais facilmente suas vantagens.

A grande contribuição deste trabalho é a definição de algumas diretrizes no processo de análise de projeções multidimensionais e suas métricas de avaliação. Isto foi possível por meio das análises dos resultados experimentais descritos no Capítulo 5. A partir de diferentes configurações de bases de dados, onde a variação se dava em número de instâncias, dimensões e grupos, percebeu-se que o número de instâncias é o fator mais relevante no tempo de execução das métricas de avaliação. Além disso, o número de instâncias e dimensões foram fatores que influenciaram diretamente nos valores das métricas, além da própria técnica de projeção.

Nos experimentos foram utilizadas bases sintéticas, onde os dados possuíam um pequeno desvio padrão, tornando as instâncias muito similares, e uma variação muito pequena do número de grupos (3 ou 5). Como possibilidade de trabalhos futuros destacam-se: *i*) modificar o número de grupos, o desvio padrão e as demais características das bases, a fim de observar melhor os efeitos de cada fator no processo de projeção e cálculo das métricas, *ii*) considerar outras características das projeções multidimensionais como diferentes funções de distância, por exemplo, e por fim, *iii*) considerar dados esparsos como aqueles advindos de representações textuais, por exemplo.

REFERÊNCIAS BIBLIOGRÁFICAS

BABAEE, M.; DATCU, M.; RIGOLL, G. Assessment of dimensionality reduction based on communication channel model; application to immersive information visualization. In: *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*. [S.l.: s.n.], 2013. p. 1–6.

BÖHM, C.; KRIEGEL, H.-P. Dynamically optimizing high-dimensional index structures. In: SPRINGER. *International Conference on Extending Database Technology*. [S.l.], 2000. p. 36–50.

CHAN, W. W.-Y. *A survey on Multivariate Data Visualization*. Dissertação (Mestrado) — Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, jun 2016.

CHEN, C.-h.; HÄRDLE, W. K.; UNWIN, A. *Handbook of data visualization*. [S.l.]: Springer Science & Business Media, 2007.

CHERNOFF, H. The use of faces to represent points in k-dimensional space graphically. *Journal of the American statistical Association*, Taylor & Francis Group, v. 68, n. 342, p. 361–368, 1973.

DIAS, M. M. et al. Visualization techniques: Which is the most appropriate in the process of knowledge discovery in data base? In: *Advances in Data Mining Knowledge Discovery and Applications*. [S.l.]: InTech, 2012.

ELER, D. M. et al. Simplified stress and simplified silhouette coefficient to a faster quality evaluation of multidimensional projection techniques and feature spaces. In: IEEE. *Information Visualisation (iV), 2015 19th International Conference on*. [S.l.], 2015. p. 133–139.

FADEL, S. G. et al. Loch: A neighborhood-based multidimensional projection technique for high-dimensional sparse spaces. *Neurocomputing*, Elsevier, v. 150, p. 546–556, 2015.

FANG, H.-r.; SAKELLARIDI, S.; SAAD, Y. Multilevel nonlinear dimensionality reduction for manifold learning. Technical report, Minnesota Supercomputer Institute, University of Minnesota, 2009.

GRIPARIS, A.; FAUR, D.; DATCU, M. A dimensionality reduction approach for the visualization of the cluster space: A trustworthiness evaluation. In: *International Geoscience and Remote Sensing Symposium (IGARSS)*. [S.l.: s.n.], 2016. v. 2016-November, p. 2917–2920.

GRIPARIS, A.; FAUR, D.; DATCU, M. A dimensionality reduction approach to support visual data mining: Co-ranking-based evaluation. In: IEEE. *Communications (COMM), 2016 International Conference on*. [S.l.], 2016. p. 391–394.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011.

HINTON, G. E.; ROWEIS, S. T. Stochastic neighbor embedding. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2003. p. 857–864.

INSELBERG, A.; DIMSDALE, B. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: *Proceedings of the First IEEE Conference on Visualization: Visualization '90*. [S.l.: s.n.], 1990. p. 361–378.

JAIN, R. *The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling*. [S.l.]: John Wiley & Sons, 1990.

JOIA, P. et al. Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics*, IEEE, v. 17, n. 12, p. 2563–2571, 2011.

JOLLIFFE, I. Principal component analysis. In: *International encyclopedia of statistical science*. [S.l.]: Springer, 2011. p. 1094–1096.

KEIM, D. A. Pixel-oriented visualization techniques for exploring very large data bases. *Journal of Computational and Graphical Statistics*, v. 5, n. 1, p. 58–77, mar 1996.

KEIM, D. A. Visual exploration of large data sets. *Communications of the ACM*, ACM, v. 44, n. 8, p. 38–44, 2001.

KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele University*, v. 33, n. 2004, p. 1–26, 2004.

LEE, J. A.; VERLEYSEN, M. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, Elsevier, v. 72, n. 7-9, p. 1431–1443, 2009.

MAATEN, L. v. d.; HINTON, G. Visualizing data using t-sne. *Journal of machine learning research*, v. 9, n. Nov, p. 2579–2605, 2008.

MAATEN, L. V. D.; POSTMA, E.; HERIK, J. Van den. Dimensionality reduction: a comparative. *J Mach Learn Res*, v. 10, p. 66–71, 2009.

MARTIN-MERINO, M.; MUNOZ, A. A new sammon algorithm for sparse data visualization. In: IEEE. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. [S.l.], 2004. v. 1, p. 477–481.

MARTINS, R. M. et al. Visual analysis of dimensionality reduction quality for parameterized projections. *Computers & Graphics*, Elsevier, v. 41, p. 26–42, 2014.

MARTINS, R. M. et al. Explaining neighborhood preservation for multidimensional projections. *EG UK Computer Graphics and Visual Computing*, University College London, 2015.

MAZZA, R. *Introduction to information visualization*. [S.l.]: Springer Science & Business Media, 2009.

MENG, D.; LEUNG, Y.; XU, Z. A new quality assessment criterion for nonlinear dimensionality reduction. *Neurocomputing*, Elsevier, v. 74, n. 6, p. 941–948, 2011.

MOTTA, R. et al. Graph-based measures to assist user assessment of multidimensional projections. *Neurocomputing*, Elsevier, v. 150, p. 583–598, 2015.

NEVES, T. T. et al. Análise visual utilizando projeções multidimensionais. *Revista de Informática Teórica e Aplicada-RITA*, Instituto de Informática da UFRGS, v. 22, n. 2, p. 258–288, 2015.

NONATO, L. G.; AUPETIT, M. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE transactions on visualization and computer graphics*, IEEE, 2018.

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

PAGLIOSA, P. et al. Projection inspector: Assessment and synthesis of multidimensional projections. *Neurocomputing*, v. 150, n. PB, p. 599–610, 2015.

PAI, M. et al. Systematic reviews and meta-analyses: an illustrated, step-by-step guide. *The National Medical Journal of India*, v. 17, n. 2, p. 86–95, 2004.

PATRO, A. *Pixel Oriented Visualization in XmdvTool*. Dissertação (Mestrado) — WORCESTER POLYTECHNIC INSTITUTE, aug 2004.

PAULOVICH, F. V. et al. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, v. 14, n. 3, p. 564–575, 2008.

PAULOVICH, F. V.; SILVA, C. T.; NONATO, L. G. Two-phase mapping for projecting massive data sets. *IEEE Transactions on Visualization and Computer Graphics*, IEEE, v. 16, n. 6, p. 1281–1290, 2010.

PETTICREW, M. Systematic reviews from astronomy to zoology: myths and misconceptions. *Bmj*, British Medical Journal Publishing Group, v. 322, n. 7278, p. 98–101, 2001.

RABELO, E. et al. Information visualization: Which is the most appropriate technique to represent data mining results? In: *International Conference on Computational Intelligence for Modelling, Control and Automation*. [S.l.: s.n.], 2008. p. 1228–1233.

RAO, R.; CARD, S. K. *The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+Context Visualization for Tabular Information*. [S.l.]: ACM, 1994. 318–322 p.

RIECK, B.; LEITTE, H. Agreement analysis of quality measures for dimensionality reduction. *Topological Methods in Data Analysis and Visualization*, Springer, p. 103–117, 2015.

RIECK, B.; LEITTE, H. Persistent homology for the evaluation of dimensionality reduction schemes. *Computer Graphics Forum*, v. 34, n. 3, p. 431–440, 2015.

ROWEIS, S. T.; SAUL, L. K. Nonlinear dimensionality reduction by locally linear embedding. *science*, American Association for the Advancement of Science, v. 290, n. 5500, p. 2323–2326, 2000.

SCHUBERT, E. et al. A framework for clustering uncertain data. *PVLDB*, v. 8, n. 12, p. 1976–1979, 2015.

SILVA, R. R. da; RAUBER, P. E.; TELEA, A. C. Beyond the third dimension: Visualizing high-dimensional data with projections. *Computing in Science & Engineering*, IEEE, v. 18, n. 5, p. 98–107, 2016.

STURROCK, K.; ROCHA, J. A multidimensional scaling stress evaluation table. *Field methods*, Sage Publications Sage CA: Thousand Oaks, CA, v. 12, n. 1, p. 49–60, 2000.

TEJADA, E.; MINGHIM, R.; NONATO, L. G. On improved projection techniques to support visual exploration of multi-dimensional data sets. *Information Visualization*, SAGE Publications Sage UK: London, England, v. 2, n. 4, p. 218–231, 2003.

TENENBAUM, J. B.; SILVA, V. D.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. *science*, American Association for the Advancement of Science, v. 290, n. 5500, p. 2319–2323, 2000.

TORGERSON, W. S. Multidimensional scaling: I. theory and method. *Psychometrika*, Springer, v. 17, n. 4, p. 401–419, 1952.

TSAI, F. S.; CHAN, K. L. Dimensionality reduction techniques for data exploration. In: *2007 6th International Conference on Information, Communications Signal Processing*. [S.l.: s.n.], 2007. p. 1–5.

UMAN, L. S. Systematic reviews and meta-analyses. *Canadian Academy of Child and Adolescent Psychiatry*, v. 20, n. 1, p. 57–59, 2011.

YANG, L. Distance-preserving dimensionality reduction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 1, n. 5, p. 369–380, 2011.