



UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Tecnologia

Miguel Angelo Mena Póvoa

**Visualização de múltiplas curvas de risco de campos e
poços como auxílio na seleção de modelos representativos
de reservatórios de petróleo**

Limeira
2024

Miguel Angelo Mena Póvoa

Visualização de múltiplas curvas de risco de campos e poços como auxílio na seleção de modelos representativos de reservatórios de petróleo

Dissertação apresentada à Faculdade de Tecnologia da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Tecnologia, na área de Sistemas de Informação e Comunicação.

Orientador: Prof. Dr. Celmar Guimarães da Silva

Este trabalho corresponde à versão final da Dissertação defendida por Miguel Angelo Mena Póvoa e orientada pelo Prof. Dr. Celmar Guimarães da Silva.

Limeira
2024

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Faculdade de Tecnologia
Felipe de Souza Bueno - CRB 8/8577

P869v Póvoa, Miguel Angelo Mena, 1981-
Visualização de múltiplas curvas de risco de campos e poços como auxílio na seleção de modelos representativos de reservatórios de petróleo / Miguel Angelo Mena Póvoa. – Limeira, SP : [s.n.], 2024.

Orientador: Celmar Guimarães da Silva.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Tecnologia.

1. Visualização da informação. 2. Reservatórios de petróleo. I. Silva, Celmar Guimarães da, 1978-. II. Universidade Estadual de Campinas. Faculdade de Tecnologia. III. Título.

Informações Complementares

Título em outro idioma: Visualization of multiple risk curves of fields and wells as an aid in the selection of representative models of oil reservoirs

Palavras-chave em inglês:

Information visualization

Petroleum reservoirs

Área de concentração: Sistemas de Informação e Comunicação

Titulação: Mestre em Tecnologia

Banca examinadora:

Celmar Guimarães da Silva [Orientador]

Maria Beatriz Alves de Sousa Santos

Jorge Luis Poco-Medina

Data de defesa: 21-02-2024

Programa de Pós-Graduação: Tecnologia

Identificação e informações acadêmicas do(a) aluno(a)

- ORCID do autor: <https://orcid.org/0000-0002-4519-2152>

- Currículo Lattes do autor: <https://lattes.cnpq.br/8670341324098267>

FOLHA DE APROVAÇÃO

Abaixo se apresentam os membros da comissão julgadora da sessão pública de defesa de dissertação para o Título de Mestre em Tecnologia na área de concentração Sistemas de Informação e Comunicação, a que se submeteu o aluno Miguel Angelo Mena Póvoa, em 21 de fevereiro de 2024 na Faculdade de Tecnologia – FT/UNICAMP, em Limeira/SP.

Prof. Dr. Celmar Guimarães da Silva

Presidente da Comissão Julgadora

Profa. Dra. Maria Beatriz Alves de Sousa Santos

Departamento de Electrónica, Telecomunicações e Informática/Universidade de Aveiro, Portugal

Prof. Dr. Jorge Luis Poco-Medina

Escola de Matemática Aplicada/Fundação Getúlio Vargas, Brasil

Ata da defesa, assinada pelos membros da Comissão Examinadora, encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-graduação da Faculdade de Tecnologia.

Dedico esse trabalho aos meus pais Eluzicléa Maria de Oliveira Mena e Antônio Adriano Póvoa (*in memorian*), com todo o meu amor e gratidão.

*Se as coisas são inatingíveis... ora!
Não é motivo para não querê-las...
Que tristes os caminhos, se não fora
A presença distante das estrelas!*

(Mário Quintana)

Agradecimentos

Agradeço primeiro a Deus por ter me mantido na trilha certa durante este projeto de pesquisa com saúde e forças para chegar até o final.

Agradeço à minha esposa Isis Fontes Póvoa e ao meu filho João Miguel Fontes Póvoa por compreenderem as várias horas em que estive ausente por conta do desenvolvimento deste trabalho, e pelo profundo amor, me apoiando em todos os momentos.

Agradeço ao meu orientador, o Prof. Dr. Celmar Guimarães da Silva, pelas valiosas contribuições dadas durante todo o processo e por aceitar conduzir o meu trabalho de pesquisa.

Agradecimentos a todos os meus professores da Faculdade de Tecnologia da UNICAMP pela excelência da qualidade técnica de cada um.

Agradeço ao meu irmão Marcus Vinícius Mena Póvoa pelo carinho e atenção que sempre me deu, além das revisões cuidadosas das versões preliminares desse trabalho.

Agradeço ao meu amigo Isaias Ferreira, que esteve sempre ao meu lado, pela amizade incondicional e pelo apoio demonstrado ao longo de todo o tempo em que me dediquei a este trabalho.

Também agradeço ao meu amigo Renato Pinheiro de Souza que sempre me ajudou com sua vasta experiência desde o início deste projeto de pesquisa.

Agradeço o apoio do Centro de Inovação em Produção de Energia (EPIC), sediado na Universidade Estadual de Campinas (UNICAMP) e patrocinado pela Equinor Brasil e pela Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP (2017/15736-3).

Agradeço o apoio da ANP (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis) por meio da regulamentação da política de P&D.

Agradecimentos são estendidos ao Centro de Estudos de Energia e Petróleo (CEPETRO), Faculdade de Engenharia Mecânica (FEM) e Faculdade de Tecnologia (FT).

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Resumo

A simulação computacional de modelos de reservatórios é uma ferramenta indispensável nas metodologias para desenvolvimento e gestão de reservatórios de petróleo. Engenheiros de petróleo criam centenas de modelos para representarem as incertezas de um determinado reservatório de petróleo. No entanto, simular o comportamento de todo esse conjunto de modelos muitas vezes é demorado e, às vezes, inviável para algumas etapas desta metodologia. A seleção de modelos representativos (MRs) é uma estratégia para contornar esta limitação. Um conjunto de MRs, por ser um subconjunto dos modelos inicialmente considerados, pode ser utilizado nas etapas de simulação em substituição ao seu conjunto de origem, visando reduzir o tempo total de processamento computacional na execução de simulações. Apesar de utilizar abordagens automatizadas para seleção de MRs, esta escolha deve ser criteriosamente assistida para assegurar que estes modelos realmente capturem as características relevantes do conjunto principal de modelos, conforme as expectativas técnicas dos analistas de petróleo. Auxiliar nesta escolha inclui fornecer maneiras fáceis de analisar semelhanças entre pares de curvas de risco referentes a cada variável de saída definida na estratégia de exploração dos modelos em análise. O uso de técnicas interativas de visualização para aprimorar o processo de seleção de MRs, em relação às curvas de risco, pode colaborar para que se faça simulações com modelos que efetivamente representem o reservatório sob análise, auxiliando os analistas a obter uma melhor previsão do comportamento do reservatório. Neste contexto, este trabalho procura explorar recursos de visualização interativa de dados que se mostrem úteis para apoiar a análise de dados de probabilidade acumulada de riscos atualmente utilizados para auxiliar engenheiros de petróleo na escolha de MRs em processo de desenvolvimento e gerenciamento de reservatórios de petróleo. A metodologia proposta inclui: (i) a análise de visualizações propostas pela literatura para esse propósito; (ii) o entendimento das necessidades dos analistas da área de petróleo para tomadas de decisão referentes a essas escolhas; e (iii) a proposta, implementação e testes de visualizações que procurem melhorar a análise em questão. Dessa forma, desenvolveu-se um sistema de visualização interativo para apresentar uma visão geral da representatividade e da variabilidade de modelos, usando *heatmaps* e projeção multidimensional, além de um gráfico para resumir múltiplas curvas de risco. Esse sistema foi avaliado por meio de um teste com usuários, o qual revelou bons indícios da sua utilidade: ajudar analistas de petróleo a efetuarem melhores escolhas de conjuntos de MRs.

Abstract

Computer simulation of reservoir models is an indispensable tool in methodologies for developing and managing oil reservoirs. Oil engineers create hundreds of models to represent the uncertainties of a given oil reservoir. However, simulating the behavior of this entire set of models is often time-consuming and sometimes unfeasible for some stages of this methodology. The selection of representative models (RMs) is a strategy to overcome this limitation. A set of RMs, being a subset of the models initially considered, can be used in the simulation stages to replace their original set, in order to reduce the total computational processing time when running simulations. Despite using automated approaches for selecting MRs, this choice must be carefully assisted to ensure that these models really do capture the relevant characteristics of the main set of models, as per the technical expectations of oil analysts. Assisting in this choice includes providing easy ways to analyze similarities between pairs of risk curves referring to each output variable defined in the exploration strategy of the models under analysis. The use of interactive visualization techniques to improve the process of selecting MRs in relation to risk curves can help simulations to be carried out with models that effectively represent the reservoir under analysis, helping analysts obtain a better prediction of the reservoir's behavior. In this context, this work seeks to explore interactive data visualization resources that prove useful to support the analysis of cumulative probability risk data currently used to assist petroleum engineers in choosing MRs in the process of developing and managing oil reservoirs. The proposed methodology includes: (i) an analysis of the visualizations proposed in the literature for this purpose; (ii) an understanding of the needs of oil analysts when making decisions regarding these choices; and (iii) the proposal, implementation and testing of visualizations that seek to improve the analysis in question. In this way, an interactive visualization system was developed to present an overview of the representativeness and variability of models, using heatmaps and multidimensional projection, as well as a graph to summarize multiple risk curves. This system was evaluated through a user test, which revealed good indications of its usefulness: helping oil analysts to make better choices of MR sets.

Lista de Figuras

2.1	Painel interativo para análise exploratória do conjunto de dados <i>MovieLens</i> . . .	26
2.2	Definições dos elementos visuais marcas.	30
2.3	Quantidade de pessoas por cor ou raça, e escolaridade.	31
2.4	Gráfico de dispersão com quatro variáveis.	33
2.5	Matriz de dispersão de quatro dimensões.	34
2.6	MDS representado semelhanças entre elementos químicos.	35
2.7	Visualizações de seis mil dígitos manuscritos do conjunto de dados MNIST . .	36
2.8	Gráfico de linhas com três variáveis quantitativas (normalizadas).	37
2.9	Gráfico de coordenadas paralelas com doze variáveis.	38
2.10	<i>Heatmap</i> por gênero de filmes agrupados por década.	39
2.11	Representação de <i>heatmaps</i> utilizando técnica de reordenação de matrizes . .	39
2.12	Desenvolvimento e gerenciamento de campo em fluxo fechado.	45
2.13	Curva de risco da variável de saída <i>NPV</i>	48
2.14	Seleção manual de nove MRs com base nos <i>cross-plots</i> das variáveis de saída. .	49
2.15	Mapas 3D do caso de referência OLYMPUS.	50
2.16	Cálculo do $rm^f(s)$ para variável de saída <i>NPV</i>	55
2.17	Linhas cinza representando o cálculo da diferença entre curvas de risco para a variável de saída <i>NPV</i>	56
2.18	Dois CCDFs da variável de saída <i>NPV</i>	58
2.19	Gráfico com dez curvas de risco combinadas (sobrepostas) para a variável de saída <i>NPV</i>	59
2.20	Gráficos com curvas de risco justapostas.	60
2.21	Gráficos de curvas de risco justapostas e sobrepostas.	61
2.22	Gráficos de <i>cross-plot</i>	62
2.23	Gráfico de diferença de atributo-nível.	63
3.1	Representações visuais do RMViewer	68
3.2	Visualização de 471 modelos agrupados de acordo com suas características de produção.	69
3.3	Matriz de comparação de erros de soluções calculadas para reservatórios de petróleo.	70
3.4	Visualização de MRs selecionados via <i>stacking plot</i>	71
3.5	Visualização da saturação de óleo a partir de 214 modelos de reservatórios. . .	72
3.6	Exemplo do teste não-paramétrico proposto por Kolmogorov-Smirnov.	73
3.7	Resultado de agrupamento para 100 modelos geológicos, com seis agrupamentos para o fator de recuperação de óleo.	74
4.1	Extrato de um exemplo do conjunto de dados de entrada.	78

4.2	Arquivo de parâmetros salvo pela interface gráfica do RMFinder, armazenando a configuração de execução definida pelo usuário.	78
4.3	Exemplo da interface gráfica para definição de parâmetros do RMFinder. . . .	80
4.4	Arquivo de saída do RMFinder para um conjunto de dez MRs.	81
4.5	O ciclo de projeto proposto à área de InfoViz aplicada.	83
5.1	Projeção multidimensional das curvas de risco	89
5.2	<i>Heatmap</i> reordenável de soluções por variáveis	90
5.3	<i>Heatmap</i> reordenável de soluções por modelos	91
5.4	Curvas de risco com envoltório	92
6.1	Visão geral do sistema de visualização interativo	99
6.2	Representações visuais disponíveis no sistema de visualização interativo . . .	100
6.3	Recursos interativos do gráfico de projeção multidimensional de soluções . . .	103
6.4	Recursos interativos do <i>heatmap</i> reordenável de variáveis por soluções	106
6.5	Recursos interativos do <i>heatmap</i> de modelos utilizados por solução	108
6.6	Seleção das marcas dos gráficos para serem detalhadas as curvas de risco associadas a elas	110
6.7	Recursos interativos do gráfico curvas de risco sumarizadas por envoltório . .	112
6.8	Visão geral das curvas de risco sumarizadas por envoltório	114
6.9	Recursos interativos adicionais do gráfico curvas de risco sumarizadas por envoltório	115
6.10	Arquivo de saída do RMFinder para uma solução com três MRs.	117
6.11	Exemplo do cálculo da distribuição empírica $F(r)$ de uma solução com três MRs da variável de saída <i>NPV</i>	118
6.12	Recursos interativos do gráfico curvas de risco justapostas e sobrepostas . . .	119
6.13	Visão geral das curvas de risco justapostas e sobrepostas	121
6.14	Visão geral das curvas de risco sumarizadas por envoltório como alternativa à abordagem híbrida de justaposição e sobreposição	122
6.15	Recursos interativos adicionais do gráfico curvas de risco justapostas e sobrepostas	124
6.15	Recursos interativos adicionais do gráfico curvas de risco justapostas e sobrepostas	125
6.16	Recursos interativos do gráfico curvas de risco do conjunto de todos os modelos	126
6.17	Recursos interativos adicionais do gráfico curvas de risco do conjunto de todos os modelos	127
6.18	Barras de botões exibidas no sistema de visualização interativo	128
6.19	Passeio pelos recursos interativos e representações visuais do sistema de visualização.	130
6.20	Filtros para selecionar soluções com conjuntos de até nove MRs.	132
6.21	Caso de uso do <i>heatmap</i> reordenável de variáveis por soluções	133
6.22	Caso de uso do <i>heatmap</i> de modelos utilizados por solução	134
6.23	Seleção das células do <i>heatmap</i> considerando somente os modelos não utilizados em nenhuma solução	135
6.24	Caso de uso do gráfico de projeção multidimensional de curvas de risco	136
6.25	Seleção de soluções no gráfico projeção multidimensional de curvas de risco .	137
6.26	Caso de uso do gráfico curvas de risco sumarizadas por envoltório.	139
6.27	Caso de uso do gráfico curvas de risco justapostas e sobrepostas	141
6.27	Caso de uso do gráfico curvas de risco justapostas e sobrepostas	142

6.27	Caso de uso do gráfico curvas de risco justapostas e sobrepostas	143
6.28	Gráficos de visão geral destacando a solução 6 e execução 9	145
7.1	Análise dos resultados da avaliação do sistema de visualização interativo. . . .	151
7.2	Gráfico de projeção multidimensional de curvas de risco com apenas uma solução.	154
G.1	Modelo Entidade-Relacionamento	195

Lista de Tabelas

3.1	Trabalhos relacionados.	66
3.2	Comparação entre os trabalhos relacionados.	75
6.1	Cálculo da distribuição empírica $F(r)$ de uma solução com três MRs da variável de saída NPV	117
7.1	Perguntas para a etapa de teste remoto do sistema de visualização interativo .	150
7.2	Comparação entre a literatura atual e a presente pesquisa.	160
B.1	Artigos resultantes da revisão sistemática.	171
H.1	Parâmetros do algoritmo de projeção multidimensional t-SNE	196
I.1	Parâmetros do algoritmo de agrupamento <i>X-means</i>	197

Lista de Símbolos

P_r	Estratégia de produção.
s	Representa um cenário.
x	A união de um cenário s e uma estratégia de produção P_r (denominado caso).
τ	Conjunto de todos os modelos.
A	Conjunto de atributos.
a	Representa um atributo do conjunto A .
n_a	Atributo-nível.
l	O nível de atributo.
$p_{(a,l)}$	Probabilidade do atributo-nível (sendo l o nível de atributo a).
P_{10}	Percentil 10.
P_{50}	Percentil 50.
P_{90}	Percentil 90.
acc^f	Probabilidade acumulada.
$p(r)$	Probabilidade individual do elemento de um conjunto menor contido em um conjunto maior.
f	Valor da variável de saída.
NPV	<i>Net Present Value</i> (Valor Presente Líquido).
Np	<i>Cumulative Oil Production</i> (Produção Cumulativa de Óleo).
ORF	<i>Oil Recovery Factor</i> (Fator de Recuperação do Óleo).
Wp	<i>Cumulative Water Production</i> (Produção Cumulativa de Água).
τ'	Conjunto de modelos representativos.
S	Conjunto de soluções do RMFinder, no qual representa um conjunto de modelos representativos
F_{cross}	Função objetivo <i>cross-plot</i> .
F_{risk}	Função objetivo curva de risco.
F_{atr}	Função objetivo atributo-nível.
$F_{penalty}$	Função penalidade.
F	Função multi-objetivo.
q_i	Pesos atribuídos pelo usuário no RMFinder para aumentar ou diminuir a importância dos componentes de F .
$F_{f,g}^{cross}$	Distorção de erro quadrático do <i>cross-plot</i> $f \times g$.

q_i^c	Pesos atribuídos pelo usuário no RMFinder para indicar quais cross-plots são mais importantes.
$d_{f,g}$	Soma das distâncias euclidianas quadradas entre cada cenário e o MR mais próximo.
q_i^r	Pesos atribuídos pelo usuário no RMFinder para indicar as curvas de risco são mais importantes.
Q	Soma dos pesos q_i^r para todas as variáveis de saída f .
m	Quantidade de variáveis de saída.
P	Conjunto de probabilidades dos MRs.
$\Delta(P)$	Função distância baseada no conjunto de probabilidades P dos MRs.
T	Conjunto de cenários.
rm^f	Representação do MR r de cada cenário de T .
θ	Retorno da função penalidade que corresponde ao número de atributos-nível não encontrados nos MRs.
EMV	<i>Estimated Monetary Value</i> (Valor Monetário Estimado).
$Winj$	<i>Water injected, cumulative</i> (Água injetada, cumulativa).
OIP	<i>Oil in Place</i> (Óleo no lugar).
WIP	<i>Water in Place</i> (Água no lugar).
$WEIIN1$ a $WEIIN7$	<i>Well economic indices</i> (Índices econômicos do poço) para poços injetores de 1 a 7.
$WEIPR1$ a $WEIPR9$	<i>Well economic indices</i> (Índices econômicos do poço) para poços produtores de 1 a 9.
E	Conjunto de execuções do RMFinder
S	Conjunto de soluções do RMFinder
Δ^f	Curva de risco (ou CCDF) de uma variável de saída
ξ	Conjunto de curvas de riscos utilizado para o cálculo da matriz de dissimilaridade
D	Matriz de dissimilaridade
d	Medida de distância
$H^{n \times m}$	Matriz numérica do <i>heatmap</i> variáveis por soluções
u_{f_i}	Função dos limites superior de curvas de riscos associadas a uma variável de saída
l_{f_i}	Função dos limites inferior de curvas de riscos associadas a uma variável de saída
$b^{f_i}(x)$	Função de contorno (limite superior e inferior) de curvas de riscos associadas a uma variável de saída
$F^{f_i}(r)$	Função distribuição empírica das curvas de risco
$B^{n \times m}$	Matriz binário do <i>heatmap</i> modelos utilizados por solução

Lista de Abreviaturas e Siglas

AG	Algoritmos Genéticos
AWS	Amazon Web Services
BIC	Bayesian Information Criterion
CAAE	Certificado de Apresentação de Apreciação Ética
CCDF	<i>Complementary Cumulative Distribution Function</i>
CEP	Comitê de Ética em Pesquisa
CEPETRO	Centro de Estudos de Energia e Petróleo
CEPs	Comitês de Ética em Pesquisa
CGVCVIP	<i>Computer Graphics, Visualization, Computer Vision and Image Processing</i>
CONEP	Comissão Nacional de Ética em Pesquisa
CSV	<i>Comma Separated Values</i>
EC2	Amazon Elastic Compute
EPIC	<i>Energy Production Innovation Center</i>
ETL	Extração, Transformação e Carga
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
FT	Faculdade de Tecnologia
IHC	Interação Humano-Computador
InfoVis	<i>Information Visualization</i>
INPI	Instituto Nacional da Propriedade Industrial
K-S	Kolmogorov-Smirnov
LLE	<i>Locally Linear Embedding</i>
MDS	<i>Multidimensional Scaling</i>
MER	Modelo Entidade-Relacionamento
MNIST	<i>Modified National Institute of Standards and Technology</i>

MR	Modelos Representativos
ORM	Object-Relational Mappers
PCA	<i>Principal Component Analysis</i>
PNAD	Pesquisa Nacional por Amostra de Domicílios
PNG	<i>Portable Network Graphics</i>
SEIS	Sistemas de Informação e Engenharia de <i>Software</i>
SGBD	Sistema de Gerenciamento de Banco de Dados
t-SNE	<i>t-Stochastic Neighbour Embedding</i>
TCLE	Termo de Consentimento Livre e Esclarecido
UNICAMP	Universidade Estadual de Campinas
UNISIM	Grupo de Pesquisa em Simulação e Gerenciamento de Reservatórios

Sumário

1	Introdução	21
1.1	Motivação	23
1.2	Problema de pesquisa	23
1.3	Objetivo	23
1.4	Estrutura do trabalho	24
2	Referenciais teóricos	25
2.1	Visualização de informação	25
2.1.1	Caracterização de usuários	27
2.1.2	Caracterização de dados	28
2.1.3	Estruturas visuais e mapeamento visual	29
2.1.4	Dados multidimensionais e multivariados	32
2.1.5	Interação	40
2.1.6	Avaliação	41
2.2	Engenharia de petróleo	42
2.2.1	Metodologia de 12 passos de Schiozer et al.	44
2.2.2	Curvas de risco	47
2.2.3	Modelos representativos	48
2.2.4	Caso de estudo	50
2.3	RMFinder	50
2.3.1	Modelagem matemática	51
2.3.2	Visualizações do RMFinder	57
2.4	Considerações finais	63
3	Trabalhos relacionados	64
3.1	Revisão da literatura	65
3.2	Estado da arte	67
3.3	Considerações finais	75
4	Metodologia	76
4.1	Conjunto de dados	76
4.1.1	Dados de entrada	77
4.1.2	Arquivo de parâmetro	78
4.1.3	Conjunto de dados de saída	80
4.2	Ferramentas	81
4.3	Método de trabalho	82
4.4	Considerações finais	84

5	Entrevista e protótipo estático	85
5.1	Preparação das entrevistas	86
5.2	Preparação dos protótipos estáticos	87
5.2.1	Projeção multidimensional de curvas de risco	88
5.2.2	<i>Heatmap</i> reordenável de variáveis por soluções	89
5.2.3	<i>Heatmap</i> reordenável de modelos utilizados por solução	91
5.2.4	Curvas de risco sumarizadas por envoltório	92
5.3	Análise dos resultados das entrevistas	93
5.3.1	Avaliação da entrevista semiestruturada	93
5.3.2	Avaliação do protótipo estático	94
5.4	Considerações finais	97
6	Desenvolvimento do sistema de visualização	98
6.1	Visualizações interativas propostas	100
6.1.1	Projeção multidimensional de curvas de risco	101
6.1.2	<i>Heatmap</i> reordenável de variáveis por soluções	104
6.1.3	<i>Heatmap</i> reordenável de modelos utilizados por solução	107
6.1.4	Detalhamento das curvas de risco	109
6.1.5	Curvas de risco sumarizadas por envoltório	111
6.1.6	Curvas de risco justapostas e sobrepostas	116
6.1.7	Curvas de risco do conjunto de todos os modelos	125
6.1.8	Outros recursos interativos	127
6.2	Exemplos de caso de uso	130
6.3	Considerações finais	145
7	Avaliação do sistema de visualização	147
7.1	Preparação da avaliação	147
7.2	Análise dos resultados da avaliação	149
7.3	Limites e possibilidades de continuidade da pesquisa	155
7.4	Relacionamento entre as tarefas do usuário e as abordagens técnicas	156
7.5	Comparação entre os trabalhos relacionados	159
7.6	Considerações finais	160
8	Conclusão	162
	Referências bibliográficas	164
A	Processo de revisão sistemática	169
B	Artigos resultantes da revisão sistemática	171
C	Termo de consentimento livre e esclarecido (etapa de entrevistas)	172
D	Perguntas para guiar a entrevista semiestruturada	178
E	Termo de consentimento livre e esclarecido (etapa teste remoto)	180
F	Formulário com questionários sobre a ferramenta	187
G	Modelo Entidade-Relacionamento	195

H	t-SNE	196
I	<i>X-means</i>	197

Capítulo 1

Introdução

O cenário mundial de produção de energia vem procurando cada vez mais a diversificação de fontes de energia, caminhando para alternativas mais limpas, como a geração de energia a partir dos raios solares e dos ventos. Contudo, apesar disso, a matriz energética mundial ainda é fortemente dependente das reservas de petróleo existentes ao redor do planeta. Desse modo, continua sendo um assunto relevante compreender como se dá o processo de exploração de petróleo, e como esse processo pode ser aprimorado, em especial no que diz respeito às tomadas de decisão necessárias ao longo do processo.

As decisões de desenvolvimento e gerenciamento de reservatórios de petróleo envolvem riscos devido a diversas incertezas físicas, operacionais e econômicas (SCHIOZER et al., 2019; MIRZAEI-PAIAMAN; SANTOS; SCHIOZER, 2021). As incertezas físicas estão associadas às reservas recuperáveis dos poços de petróleo e às características de vazão. Os riscos operacionais estão relacionados à disponibilidade do sistema de produção. Por sua vez, as incertezas econômicas são atribuídas ao preço do petróleo, investimentos e despesas operacionais. Além dessas incertezas, o desenvolvimento e gerenciamento de reservatórios de petróleo também envolve definir uma estratégia de produção, que consiste no sequenciamento de perfuração, abertura, injeção e posicionamento de poços, entre outras atividades.

Por conta destas incertezas, o processo de tomada de decisão na definição de uma estratégia de produção em campos de petróleo torna-se um problema complexo, sendo necessário analisar inúmeros cenários (esse conceito será formalmente explicado na Seção 2.2). Idealmente, todas as combinações possíveis das características incertas do

reservatório sob análise devem ser consideradas e podem levar a um número exponencial de cenários a serem analisados (MEIRA et al., 2020).

Somado a isto, a simulação do comportamento do campo de petróleo para cada estratégia de produção possível e para cada cenário sob análise é inviável em situações reais, pois cada simulação pode exigir muitas horas de recurso computacional (até dias), para ser concluída. Portanto, técnicas de redução de cenários são altamente desejáveis (embora não obrigatórias), para que os tomadores de decisão possam trabalhar com um conjunto viável de modelos, que seja representativo de todo o problema (MEIRA et al., 2017).

Neste contexto, a metodologia e o *software* RMFinder (MEIRA et al., 2016) usam uma abordagem baseada em otimização, proposta para identificar automaticamente, segundo as preferências do analista de petróleo, subconjuntos de modelos de campos petrolíferos, chamados de Modelos Representativos (MR), que representem um conjunto de modelos originalmente fornecido. A função matemática a ser otimizada pelo RMFinder foi desenvolvida considerando a representatividade de um determinado subconjunto de modelos em relação a três tipos de recortes de conjuntos de dados: (i) gráficos de dispersão (*cross-plots*) que combinam pares de variáveis relevantes; (ii) às curvas de risco associadas a essas variáveis; e (iii) à distribuição de probabilidade (em níveis predefinidos) de cada atributo incerto do problema (MEIRA et al., 2016). O RMFinder é utilizado pelo Grupo de Pesquisa em Simulação e Gerenciamento de Reservatórios (UNISIM) do Centro de Estudos de Energia e Petróleo (CEPETRO) da Universidade Estadual de Campinas (UNICAMP); e esse projeto de mestrado é apoiado pelo *Energy Production Innovation Center* (EPIC)¹, iniciativa conjunta entre UNICAMP, Equinor² e Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Parte das saídas produzidas pelo RMFinder, para ilustrar como são os conjuntos de MRs escolhidos, é um conjunto de gráficos de curvas de risco, cada um associado a uma variável do problema. Cada gráfico de curvas de risco apresenta duas ou mais curvas de risco, sendo cada uma delas uma função de distribuição cumulativa complementar (*Complementary Cumulative Distribution Function* – CCDF) relativa a um determinado conjunto de modelos. Uma dessas curvas refere-se ao conjunto original de modelos. Cada uma das demais curvas dizem respeito a um conjunto de MRs selecionado pelo RMFinder (SILVA et al., 2020).

¹O EPIC, resultou de uma parceria entre UNICAMP, Equinor e FAPESP, com o princípio de que os recursos energéticos naturais de um país devem ser utilizados da maneira mais sustentável possível.

²A Equinor é uma empresa global de energia, com sede na Noruega e operações em mais de trinta países. No Brasil está presente há duas décadas, com foco em exploração e produção de óleo e gás, e em energias renováveis.

1.1 Motivação

Apesar de ser possível selecionar um conjunto de MRs de forma automatizada, essa escolha deve ser criteriosamente assistida para assegurar que esses modelos realmente capturem as características relevantes do conjunto principal de modelos, com relação às expectativas técnicas dos analistas de petróleo (MEIRA et al., 2016). Considerando que gráficos já vêm sendo utilizados nesse processo, considera-se que visualizações interativas podem facilitar ainda mais esse processo, fornecendo maneiras mais facilitadas de comparar curvas de risco e, portanto, de possibilitar melhores escolhas de conjuntos de modelos representativos. Uma melhor escolha desses modelos pode contribuir para simulações que representem mais fielmente os possíveis comportamentos do reservatório em estudo.

1.2 Problema de pesquisa

Esta pesquisa procura explorar recursos de visualização interativa de dados que se mostrem úteis para apoiar a análise de dados de probabilidade acumulada de riscos atualmente utilizados para auxiliar engenheiros de petróleo na escolha de MRs em processo de desenvolvimento e gerenciamento de reservatórios de petróleo.

1.3 Objetivo

Este projeto tem por objetivo geral propor e/ou aprimorar visualizações interativas de múltiplas curvas de risco no intuito de ajudar analistas de petróleo a efetuarem melhores escolhas de conjuntos de MRs de reservatórios de petróleo em cenários com múltiplas soluções e variáveis.

Como objetivo específico, procura-se focar na proposição de melhorias para as visualizações atualmente utilizadas pelo *software* RMFinder, bem como na proposição de visualizações de caráter mais geral, atualmente ausentes nesse sistema. Essas visualizações devem ser capazes de possibilitar comparações de mais alto nível de abstração, permitindo comparar cenários com diversas variáveis e soluções.

1.4 Estrutura do trabalho

O conteúdo deste trabalho está organizado da seguinte forma: no Capítulo 2 é apresentado o referencial teórico, contendo os principais conceitos que envolvem esse trabalho, e no Capítulo 3 são elencados os trabalhos relacionados com uma discussão sobre o estado da arte. Após isso, no Capítulo 4, detalhadamente, é apresentada a metodologia proposta. Nos Capítulos 5, 6 e 7 constam os resultados obtidos, descrevendo a execução das etapas descritas no Capítulo 4. No Capítulo 5 é apresentado o estudo dos usuários e suas tarefas, realizado por meio de entrevistas e apoiado por um protótipo estático. No Capítulo 6, é abordada a implementação do sistema interativo de visualização, bem como o seu desenvolvimento. Em seguida, no Capítulo 7, é detalhado o teste desse sistema com usuário e uma análise dos resultados dessa avaliação. No Capítulo 8 é realizada uma conclusão sobre este trabalho.

No Apêndice A encontra-se a especificação dos parâmetros para a revisão sistemática efetuada e, no Apêndice B, os trabalhos acadêmicos identificados nesse processo. Os Apêndices C, D, E e F contém os documentos utilizados na entrevista e no teste remoto realizados com o usuário, e o modelo entidade-relacionamento, criado para armazenar os dados utilizados nas visualizações, no Apêndice G. Por fim, nos Apêndices I e H, constam os parâmetros dos algoritmos utilizados neste trabalho.

Capítulo 2

Referenciais teóricos

Este Capítulo 2 tem o intuito de realizar uma revisão das pesquisas e discussões feitas por outros autores sobre os temas abordados neste trabalho. Sendo assim, na Seção 2.1 consta os principais conceitos da área de Visualização de Informação. Na Seção 2.2 é abordada a análise de decisão relacionada ao desenvolvimento e gerenciamento de campos petrolíferos. Por fim, na Seção 2.3 é detalhada a metodologia e o *software* RMFinder.

2.1 Visualização de informação

O atual estágio da evolução tecnológica, com uma aceleração acentuada na última década, tem produzido um volume cada vez maior de informações. Elas fazem parte das atividades rotineiras de muitas pessoas, por exemplo: *e-mails* e mensagens instantâneas que chegam em computadores e dispositivos móveis a todo momento; interação entre pessoas e organizações interligadas via redes sociais; monitoramento das variações de índices econômicos no cenário global; conteúdos publicitários divulgados em diversos canais de comunicação; entre outros. Por outro lado, essa “tempestade de informações” (ou, como Mazza (2009) se refere, a “poluição da informação”) trouxe uma frustração muito grande quando não se consegue analisar, interpretar ou mesmo compreender essas informações com intuito de extrair conhecimentos a partir delas. Segundo Silva (2014), em parte isso ocorre porque, apesar de cada dado poder ser analisado individualmente, a quantidade de dados e as nossas limitações cognitivas não nos permitem armazenar e processar esses dados mentalmente na totalidade.

É neste contexto que a área de Visualização de Informação (*Information Visualization* – InfoVis) vem sendo desenvolvida, com o intuito de representar graficamente conjuntos de

dados de modo a amplificar a cognição do usuário (MAZZA, 2009). Para isso, são utilizadas diversas técnicas, geralmente interativas, em que os dados são representados visualmente para o usuário. Durante a construção dessas representações visuais, vários conceitos devem ser considerados, sempre visando facilitar que o usuário extraia conhecimento a partir de dados e informações (CARD; MACKINLAY; SHNEIDERMAN, 1999).

Como exemplo, a Figura 2.1 apresenta uma tela de uma ferramenta de visualização contendo informações do conjunto de dados *MovieLens* (<https://grouplens.org/datasets/movielens>), um serviço de recomendação de filmes. A mesma tela permite verificar várias informações, como, por exemplo, a evolução dos gêneros de filmes nos últimos 120 anos, os comentários dos usuários sobre gênero de filmes específicos, as relações entre os gêneros, os filmes mais populares e com melhores avaliações das últimas décadas.

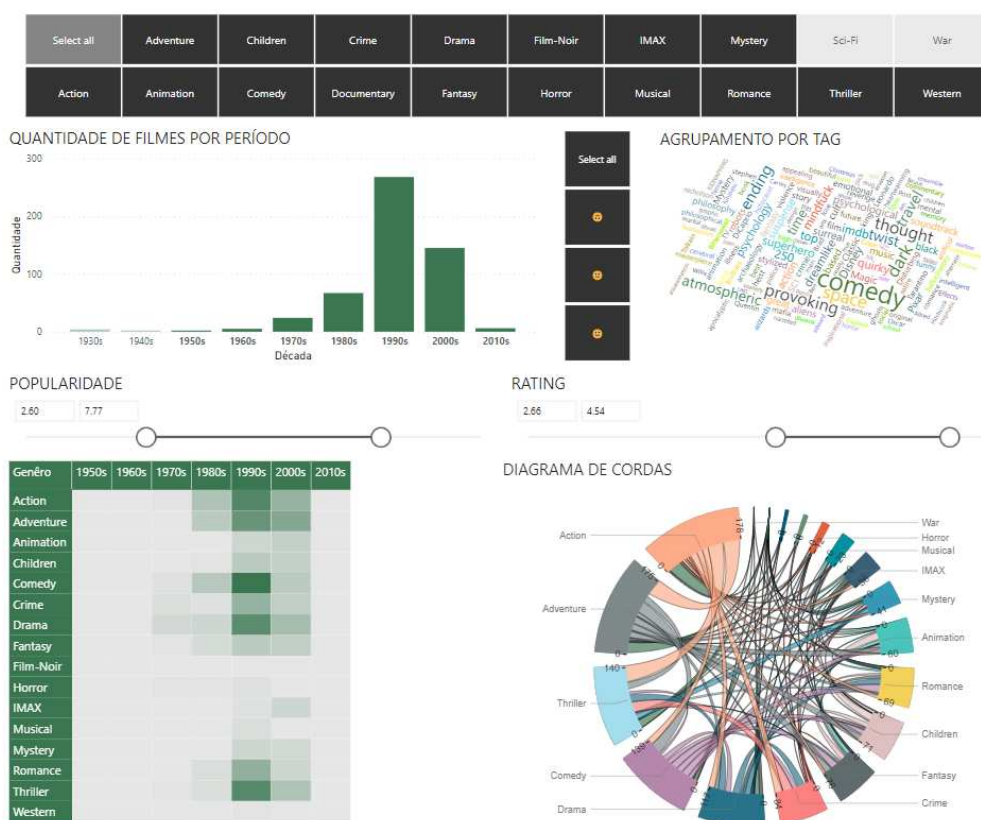


Figura 2.1: Painel interativo para análise exploratória do conjunto de dados *MovieLens*.

Fonte: Projeto realizado em disciplina de InfoVis da UNICAMP.

Nesse sentido, a presente Seção 2.1 tem por objetivo introduzir a área de InfoVis e os aspectos dela relevante para este trabalho. Em virtude disso, o conteúdo da Seção 2.1 está organizado como segue. Nas Seções 2.1.1 e 2.1.2 são abordadas a caracterização de usuários e

a caracterização de dados, respectivamente, visando guiar o processo de escolha de técnicas de visualização a partir desses elementos. Esse processo (denominado mapeamento visual) e suas respectivas estruturas visuais são detalhados na Seção 2.1.3. Entre as opções elencadas na Seção 2.1.3 será dado um enfoque maior às estruturas visuais de dados multidimensionais na Seção 2.1.4. Na Seção 2.1.5 constam técnicas de interação aplicadas a visualização. Por fim, na Seção 2.1.6 são relacionados os possíveis procedimentos de avaliação utilizados na área de InfoVis.

2.1.1 Caracterização de usuários

Segundo Spence (2014), o usuário que interpreta uma visualização é uma parte interessada extremamente complexa, cujas características exercem influência considerável sobre o desenho e a eficácia de uma representação. Embora possam ser aprimoradas por meio de treinamento, as habilidades perceptivas e cognitivas de cada ser humano são basicamente invariantes. Os usuários de uma representação devem ser considerados como pessoas com características, motivações, necessidades, capacidades e preferências próprias.

Como outros sistemas de *software* interativos, soluções baseadas em visualização precisam apresentar **usabilidade**¹; algumas das características embutidas neste conceito são: efetividade, eficiência, satisfação de usuário, facilidade de aprendizagem, flexibilidade, facilidade de lembrar, facilidade de uso, e tolerância a erros (SILVA, 2014). Como parte da estratégia de desenho de uma visualização que seja centrado no usuário, Kulyk et al. (2017) propõem que se deva responder às seguintes perguntas, com o intuito de construir um modelo sobre os usuários-alvo da aplicação:

- **Quem serão os usuários da visualização?** Aqui devem ser levantados os perfis dos usuários, com suas habilidades e/ou deficiências físicas e cognitivas, bem como aspectos culturais;
- **Que dados serão vistos?** Conhecer isto auxilia a compreender melhor o modelo mental do usuário; e

¹Segundo Kulyk et al. (2017), o termo usabilidade é uma propriedade, e não uma entidade independente de um sistema. Além disso, está relacionado com a utilização bem sucedida de um sistema.

- **Que tarefas podem ser feitas com os dados?** Entrevistas, questionários, observação etnográfica, protótipos, *workshops* participativos e demonstração de tarefas pelo usuário são técnicas úteis para auxiliar a responder a esta pergunta.

2.1.2 Caracterização de dados

Entender as características do conjunto de dados a ser representado também é de suma relevância, pois essas podem servir de guia para a escolha ou rejeição de técnicas de visualização a ser utilizadas em uma determinada situação (SILVA et al., 2020). Nesta Seção 2.1.2 constam algumas das principais caracterizações de dados a serem consideradas antes de propor uma técnica de visualização sob um conjunto de dados específico.

Os conjuntos de dados podem ser obtidos de forma manual (coleta de dados pelo próprio indivíduo) ou automatizada (p. ex. *log* de servidores, dados de sensores, transações de banco de dados, etc.) com origem em uma ou mais fontes. Esses dados, denominados por Card, MacKinlay e Shneiderman (1999) de “**dados brutos**”, normalmente não estão disponíveis em uma estrutura² adequada para serem utilizados diretamente em visualizações. Eles precisam ser transformados em estruturas, chamadas por Card, MacKinlay e Shneiderman (1999) de **tabelas de dados**, que evidenciem os **registros** (tuplas) e as **variáveis** (dimensões) a serem apresentados por uma estrutura visual (conceito também apresentado por Card, MacKinlay e Shneiderman (1999) e descrito na Seção 2.1.3).

Outro procedimento relevante é definir uma série de metadados³ referentes a classificações de variáveis da tabela de dados. Com relação a domínio, as variáveis podem ser classificadas semanticamente em (CARD; MACKINLAY; SHNEIDERMAN, 1999; MAZZA, 2009; SPENCE, 2014):

- **Nominais** (ou **categóricas**): conjunto cujos elementos não possuem uma ordem específica, p. ex. lista de nomes de pessoas;
- **Ordinais**: conjunto cujos elementos apresentam necessariamente uma relação de ordem entre si, p. ex. classificações em faixas etárias; e
- **Quantitativas** (ou **numéricas**): apresenta um escopo numérico, sendo possível aplicar operações aritméticas sobre ela, p. ex. tamanho de pessoas.

²Uma estrutura de dados é uma coleção tanto de valores (e seus relacionamentos) quanto de operações (sobre os valores e estruturas decorrentes).

³Metadados são “dados que fornecem informações sobre outros dados”, mas não o conteúdo dos dados.

Outra classificação de variáveis se diz respeito à **dependência funcional** entre elas. Segundo Card, MacKinlay e Shneiderman (1999) e Mazza (2009), esta classificação pode ser:

- **Entrada:** seus valores não dependem dos valores de nenhuma outra variável; e
- **Saída:** seus valores são determinados pelos valores das variáveis de entrada.

De acordo com Silva (2014), um terceiro tipo de metadado relevante é a **estrutura** dos dados da tabela de dados. Essa estrutura pode ser linear, temporal, espacial, hierárquica ou em rede:

- **Linear:** os dados não apresentam nenhuma relação explícita entre os elementos das tuplas da tabela de dados;
- **Temporal:** como o nome sugere, os dados variam ao longo do tempo;
- **Espacial:** assume que os dados têm um atributo espaço (pontos, linhas e regiões) ou espaço-temporal implícito ou explícito (WARD; GRINSTEIN; KEIM, 2015); e
- **Hierárquica** ou em **rede:** são tuplas relacionadas entre si na tabela de dados, com relações de árvore (p.ex., diretório de um sistema de arquivos) ou grafo (p.ex., conexões entre pessoas em redes sociais), respectivamente.

2.1.3 Estruturas visuais e mapeamento visual

Tendo sido escolhido e devidamente classificado o conjunto de dados de interesse (Seção 2.1.2), e compreendidas as necessidades da representação com foco no usuário-alvo (Seção 2.1.1), é necessário escolher como esses dados podem ser apresentados. De acordo com Card, MacKinlay e Shneiderman (1999), um conjunto de dados pode ser representado por um conjunto de elementos visuais (chamado de **estrutura visual**). Uma estrutura visual é composta de (CARD; MACKINLAY; SHNEIDERMAN, 1999):

- **Substrato espacial:** define as dimensões no espaço físico onde uma marca é criada dentro da representação. O substrato espacial pode ser definido em eixos (p.ex., no espaço cartesiano o substrato espacial corresponde aos eixos X e Y);
- **Elemento gráfico:** são todas as marcas que aparecem no espaço, como p.ex., pontos, linhas, superfícies, volumes, entre outros (Figura 2.2); e

- **Propriedade gráfica:** são propriedades dos elementos gráficos para os quais a retina do olho humano é sensível, como p. ex. cores, tamanhos, formas, texturas, orientação, entre outros (Figura 2.2).

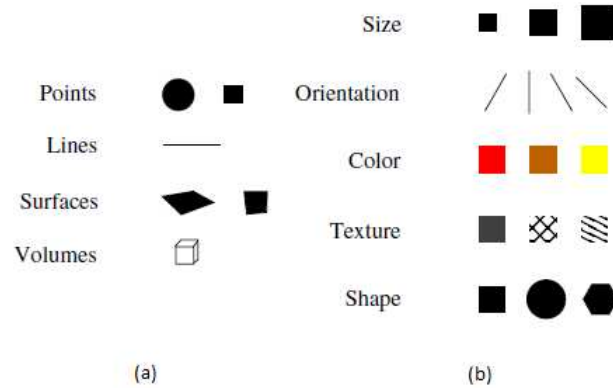


Figura 2.2: Definições dos elementos visuais marcas sob os aspectos de (a) elemento gráfico e (b) propriedade gráfica.
Fonte: (MAZZA, 2009).

Marcas podem ainda estar explicitamente relacionadas entre si, em especial por três grupos de técnicas (CARD; MACKINLAY; SHNEIDERMAN, 1999); são eles:

- **Conexão**, onde linhas unem marcas relacionadas (p. ex. tradicionais diagramas vértice-aresta para representação de grafos ou árvores);
- **Encapsulamento**, onde uma marca contém outras, o que é muito útil na apresentação de hierarquias ou relacionamentos entre conjuntos (p. ex. a técnica *Treemap*);
- **Adjacência**, onde duas marcas colocadas uma ao lado da outra indicam uma relação de hierarquia entre elas (p. ex.. as técnicas *Icicle* (KRUSKAL; LANDWEHR, 1983; ZHENG; SADLO, 2021) e *Sunburst* (STASKO; ZHANG, 2000; ZHENG; SADLO, 2021)).

Para uma estrutura visual representar os dados de uma tabela de dados, cada variável dessa tabela pode ser associada a uma propriedade gráfica ou espacial, processo esse chamado de **mapeamento visual** (SILVA, 2014). De preferência, o mapeamento visual deve reproduzir as dependências funcionais existentes na tabela de dados, repetindo as categorias desses dados. Como exemplo, o gráfico de barras da Figura 2.3 apresenta um mapeamento visual referente aos dados de escolaridade da população brasileira do conjunto de dados do instituto de Pesquisa Nacional por Amostra de Domicílios (PNAD)⁴:

⁴A PNAD é um conjunto de dados para apoiar no desenvolvimento e análises de políticas públicas diversas com informações da população brasileira.

- Eixo X: valor categórico, nível de instrução;
- Eixo Y: valor quantitativo agregado via soma para estabelecer quantidade de pessoas;
- Marca: grupo de pessoas de mesma cor ou raça;
- Cor: cor ou raça; e
- Luminosidade: marca está (ou não) selecionada.

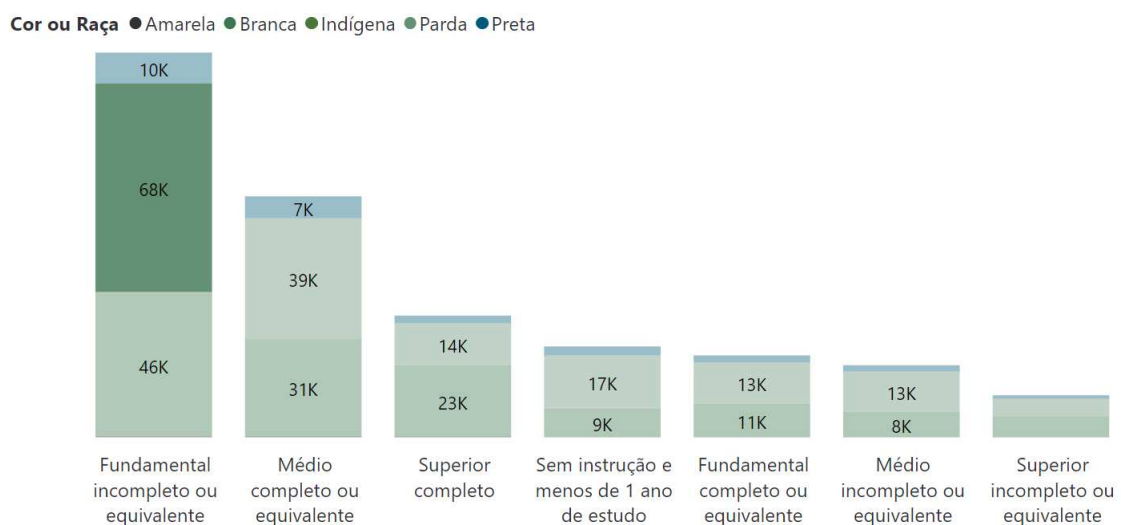


Figura 2.3: Quantidade de pessoas por cor ou raça, e escolaridade.
 Fonte: Projeto realizado em disciplina de InfoViz da UNICAMP.

Gráficos comumente apresentam eixos que definem uma métrica segundo a qual o espaço será utilizado. Eixos podem ser classificados como nominais, ordinais ou quantitativos, classificação essa que é intrinsecamente associada às características dos dados (Seção 2.1.2) e do tipo de gráfico que os usa. Por exemplo, gráficos de barras (Figura 2.3) necessariamente possuem variáveis quantitativas no Eixo Y, e no Eixo X normalmente nominais. Por outro lado, os gráficos de dispersão admitem irrestritamente eixos X e Y com variáveis nominais, ordinais ou quantitativas.

A escolha do mapeamento visual adequado é um dos pontos centrais na elaboração de uma boa solução de visualização de dados (SILVA, 2014). Essa escolha, segundo Spence (2014), pode ser dividida em quatro tarefas relacionadas a elementos visuais: associar marcas, diferenciar marcas, perceber a ordem entre elas, e perceber proporcionalidade. Para isto, Spence (2014) considerou as seguintes propriedades, chamadas de **propriedades retiniais**: tamanho, luminosidade, textura, cor, orientação e forma.

- **Posição e tamanho** são propriedades efetivas para representar dados de variáveis independentemente da sua classificação;
- **Luminosidade** é efetivo para representar dados de variáveis ordinais; e
- **Cor, textura, forma e orientação** são efetivos para permitir a diferenciação de marcas de dados nominais.

Por fim, diferentes classificações de estruturas visuais consideram a dimensionalidade dos dados e sua estruturação (MAZZA, 2009; WARD; GRINSTEIN; KEIM, 2015; SPENCE, 2014). Essas classificações podem ser listadas em 5 tipos: dados de baixa dimensionalidade, dados multidimensionais (Seção 2.1.4), dados em rede (grafos) ou hierárquicos (árvores), textos e dados temporais.

2.1.4 Dados multidimensionais e multivariados

Dados multidimensionais são dados que possuem mais do que três dimensões. Por sua vez, dados multivariados são coleções de dados nos quais muitos atributos (geralmente mais de quatro) mudam em relação a um ou mais atributos de entrada (ou independentes, Seção 2.1.2) (MAZZA, 2009). Nessa Seção 2.1.4, são apresentadas quatro técnicas (ou estruturas visuais) mais comuns para representação visual de dados multivariados ou multidimensionais (WARD; GRINSTEIN; KEIM, 2015); são elas:

- Técnicas baseadas em ponto;
- Técnicas baseadas em linha;
- Técnicas baseadas em região; e
- Combinações de técnicas.

Técnicas baseadas em ponto

De acordo com Ward, Grinstein e Keim (2015), as técnicas baseadas em ponto projetam registros de n dimensões (ou pontos de k dimensões), em um espaço de k dimensões, i.e., para cada registro, uma marca é desenhada e associada a pontos em k dimensões. Essa técnica pode ser encontrada em algumas representações visuais como gráficos de dispersão, matrizes de gráficos de dispersão e métodos baseados em força.

Gráficos de dispersão representam por um ponto (marca) os valores de dois de seus atributos nos eixos X e Y . Os demais atributos podem ser representados por outras propriedades gráficas, como cores e tamanhos da marca (Figura 2.4).

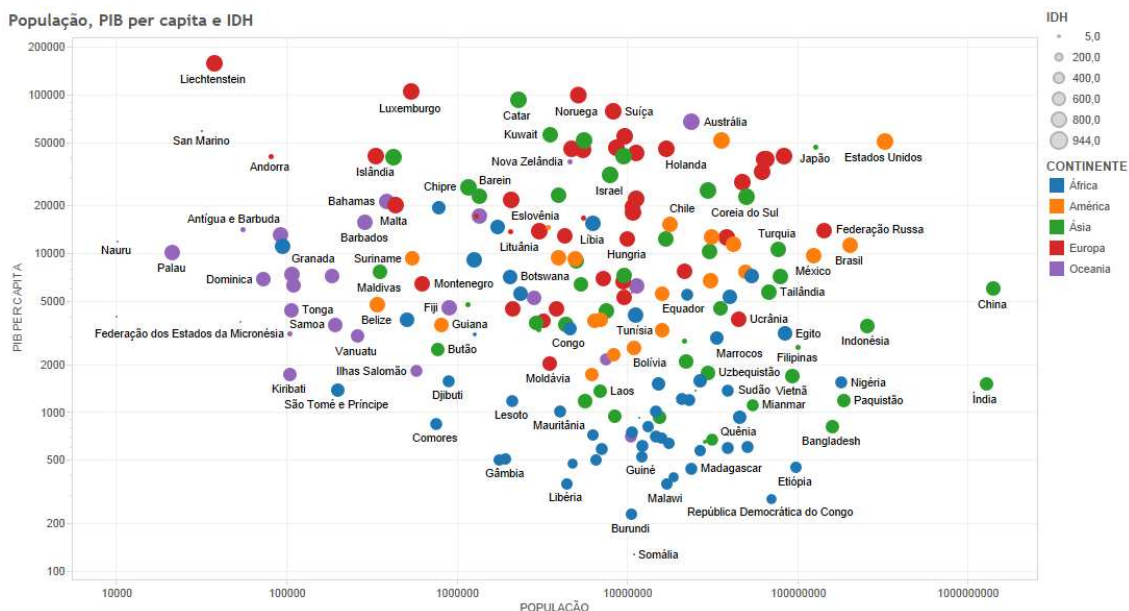


Figura 2.4: Gráfico de dispersão com quatro dimensões (nesse caso, população, PIB per capita, IDH e continente).

Fonte: (SILVA, 2014)

Outra possibilidade é utilizar matrizes de gráficos de dispersão (Figura 2.5), em que cada uma das possíveis combinações de variáveis duas a duas geram um gráfico de dispersão. Nessas matrizes, interatividade permite verificar simultaneamente em todos os gráficos os valores de uma mesma tupla ou conjunto de tuplas.

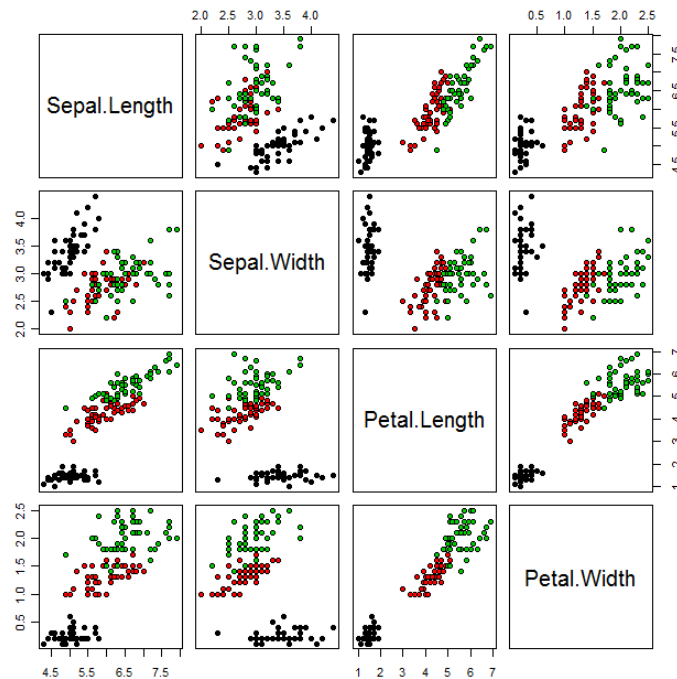


Figura 2.5: Matriz de gráfico de dispersão de quatro dimensões relativo à largura e comprimento de pétalas e sépalas de flores (conhecido como *Iris dataset*).

Fonte: Dados originais da pesquisa.

Os métodos baseados em força usam estratégia de redução da quantidade de dimensões (ou redução de dimensionalidade), buscando preservar nos pontos k -dimensionais os relacionamentos pré-existentes nos pontos n -dimensionais originais, sendo $k \ll n$ e k normalmente é igual a dois (WARD; GRINSTEIN; KEIM, 2015). Uma das classes de técnicas mais conhecidas para isso é a de escalonamento multidimensional (*Multidimensional Scaling* – MDS). Como exemplo, a Figura 2.6 apresenta a correlação existente entre o período dos elementos químicos e suas características quantitativas. Nela, cada marca representa um elemento da tabela periódica. As cores das marcas indicam a que período da tabela periódica elas pertencem. A proximidade entre as marcas reflete a semelhança entre propriedades dos elementos, como eletronegatividade, densidade, ponto de fusão, ebulição, etc.

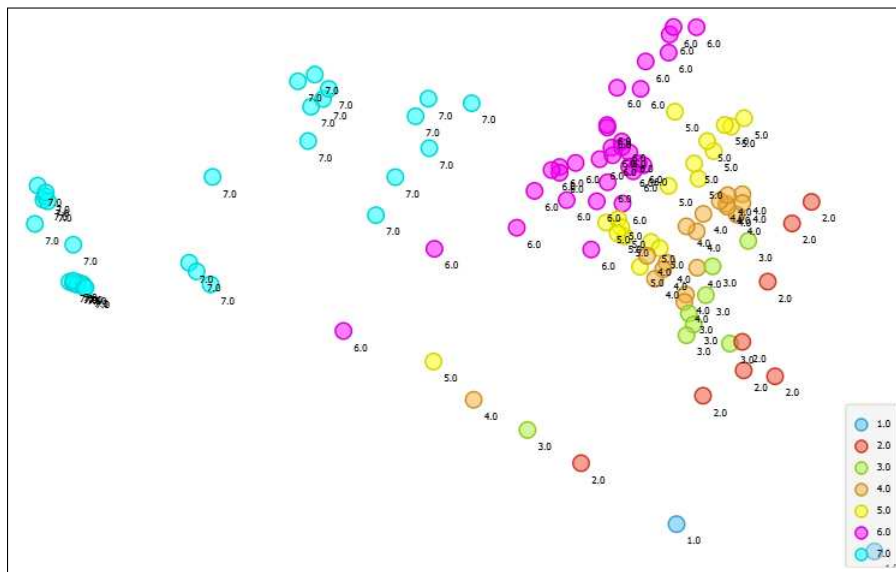


Figura 2.6: MDS representado semelhanças entre elementos químicos.

Fonte: Projeto realizado em disciplina de InfoVis da UNICAMP.

Além do MDS, existem outras técnicas de redução de dimensionalidade, as quais podem auxiliar a projetar registros de n dimensões, em um espaço de k dimensões, onde $k \ll n$. Em Nonato e Aupetit (2019) são elencadas recomendações para a escolha de uma das diversas técnicas de projeção multidimensional e também são fornecidos resultados com a análise de cada algoritmo estudado. Uma das abordagens citadas em Nonato e Aupetit (2019) é o *t-Stochastic Neighbour Embedding* (t-SNE).

O t-SNE é uma técnica de projeção multidimensional robusta na criação de uma única visualização. Ela revela a estrutura dos dados em várias escalas diferentes, tornando-o robusto contra a concentração de padrões (MAATEN; HINTON, 2008). Na Figura 2.7 é demonstrado um dos resultados de Maaten e Hinton (2008), comparando o t-SNE com outras técnicas de projeção multidimensional (Sammon, Isomap e *Locally Linear Embedding* – LLE), usando o conjunto de dados *Modified National Institute of Standards and Technology* (MNIST)⁵. Na Figura 2.7 (a), observa-se que a abordagem t-SNE (a), embora contenha alguns pontos assinalados com a classe errada, obteve uma melhor distinção do grupo de cada dígito manuscrito do que as outras abordagens utilizadas na Figura 2.7 (b), (c) e (d).

⁵O banco de dados MNIST é um banco de dados de dígitos manuscritos (de zero a nove) sendo comumente usado para treinar vários sistemas de processamento de imagens.

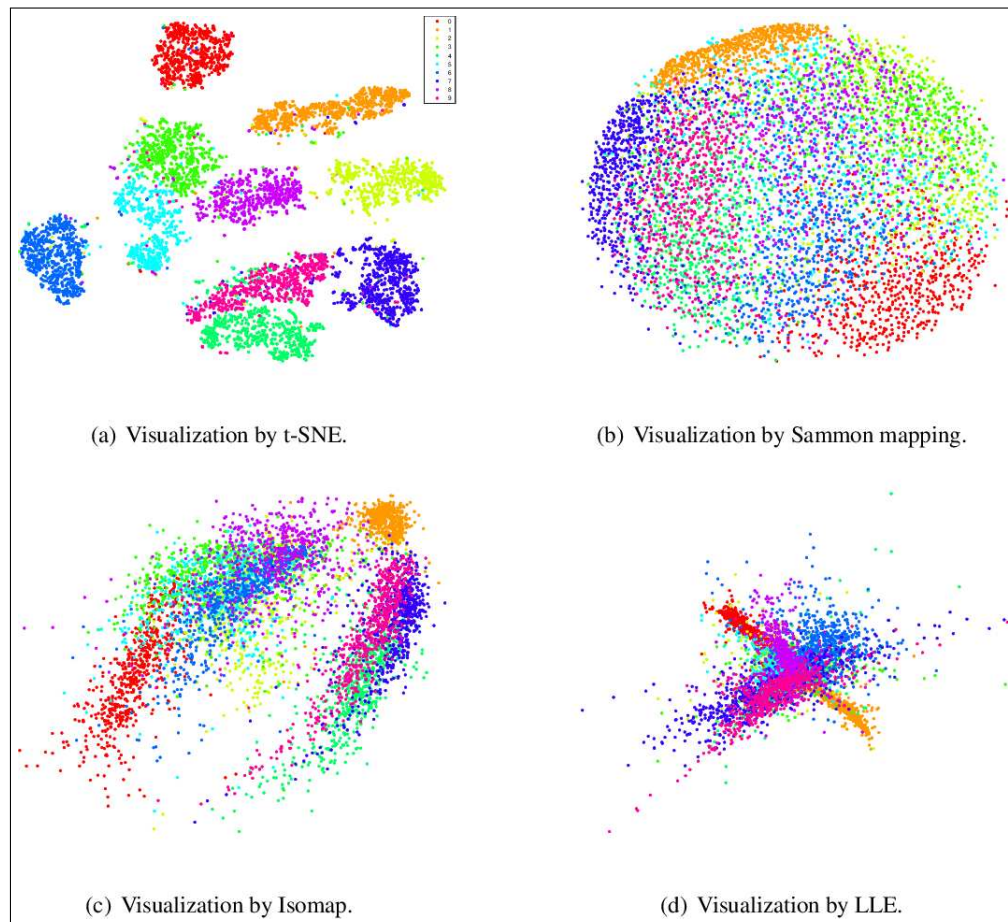


Figura 2.7: Visualizações de seis mil dígitos manuscritos do conjunto de dados MNIST comparando as abordagens de projeções multidimensional t-SNE, Sammon, Isomap e LLE. Fonte: (MAATEN; HINTON, 2008).

Na Figura 2.7 utilizou-se as classes dos dígitos manuscritos para determinar as cores de cada ponto no gráfico. Diferenciar as classes (ou rótulos) das marcas (p. ex., por cores) após a execução de uma técnica de projeção multidimensional pode ser interessante para ajudar os usuários a identificar possíveis agrupamentos. No entanto, nem sempre os dados estarão rotulados no conjunto de dados a ser analisado. Uma alternativa é utilizar uma técnica de *clustering* associada à projeção multidimensional, para prover a rotulação necessária. Dentre diversas técnicas de *clustering*, o *X-means* (PELLEG; MOORE, 2000) é um algoritmo de *clustering* derivado do *K-means* (MACQUEEN, 1967; JAIN; MURTY; FLYNN, 1999). Em vez de receber como entrada o número de *clusters* k a ser retornado, o *X-means* calcula o número ideal de *clusters* para um determinado conjunto de dados usando o Bayesian Information Criterion (BIC) para decidir a melhor divisão de *clusters*. Essa atribuição do número ideal de *clusters* pode ser extremamente útil no contexto de visualização de informação, evitando que o usuário precise selecionar o número de *clusters* desejado.

Técnicas baseadas em linha

Nos métodos baseados em linha, os pontos correspondentes a um registro (ou dimensão específica) estão ligados entre si com linhas retas ou curvas (WARD; GRINSTEIN; KEIM, 2015). O objetivo desse método é reforçar relacionamentos entre valores de dados, e tornar perceptíveis certas características dos dados, por meio de inclinações, curvaturas, cruzamentos, etc. Essa técnica tem duas representações visuais principais: gráficos de linhas e gráficos de coordenadas paralelas.

Um gráfico de linha é uma técnica de visualização bidimensional e univariada, onde o eixo X representa a faixa de valores de uma variável e o eixo Y representa alguns registros ordenados do conjunto de dados. Técnicas como esta podem ser estendidas a dados multivariados, por justaposição ou superposição de linhas em um mesmo gráfico (WARD; GRINSTEIN; KEIM, 2015). A Figura 2.8 apresenta um exemplo de gráfico de linhas sobrepostas com dados de PIB per capita, população e IDH de cada país.

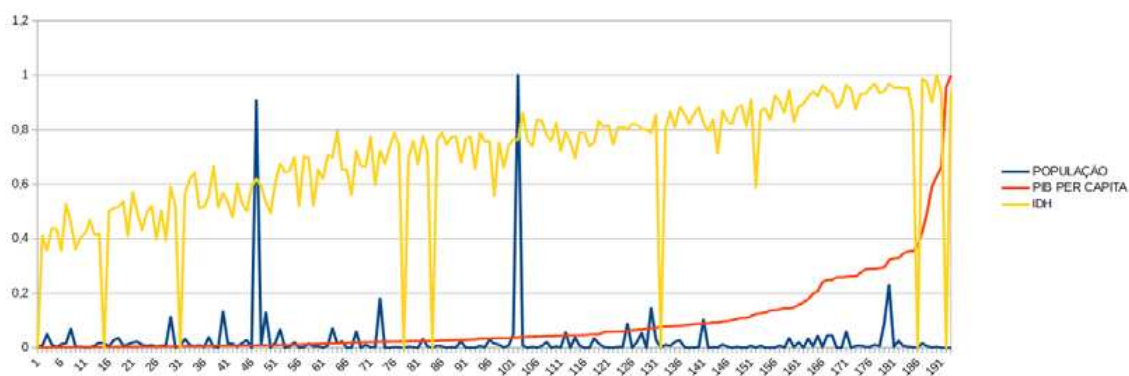


Figura 2.8: Gráfico de linhas com três variáveis quantitativas (normalizadas), ordenadas por PIB per capita

Fonte: (SILVA, 2014)

Os gráficos de coordenadas paralelas, por sua vez, são projetados para cenários multidimensionais. Seus eixos, ao invés de serem ortogonais, são paralelos entre si, com linhas verticais ou horizontais uniformemente espaçadas, representando uma ordenação particular das dimensões (WARD; GRINSTEIN; KEIM, 2015). Linhas poligonais indicam os valores de cada registro do conjunto de dados em cada eixo. A Figura 2.9 usa essa técnica para representar a magnitude e padrão dos valores numéricos para cada tipo de elemento químico da tabela periódica.

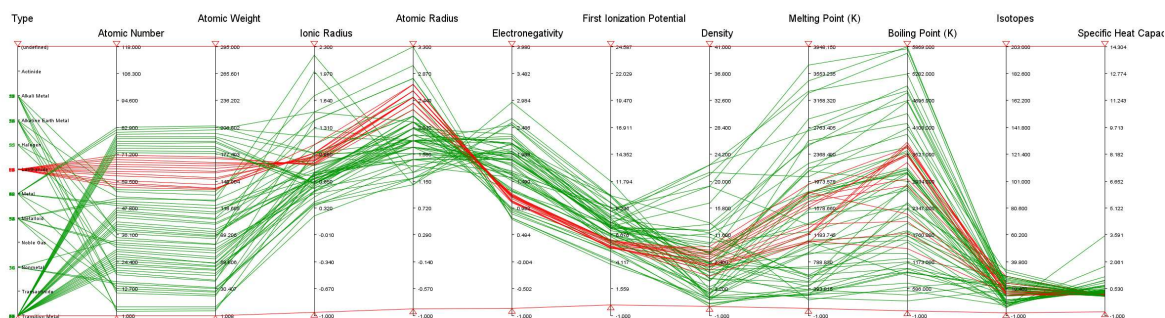


Figura 2.9: Gráfico de coordenadas paralelas com doze variáveis de propriedade dos elementos químicos.

Fonte: Projeto realizado em disciplina de InfoVis da UNICAMP.

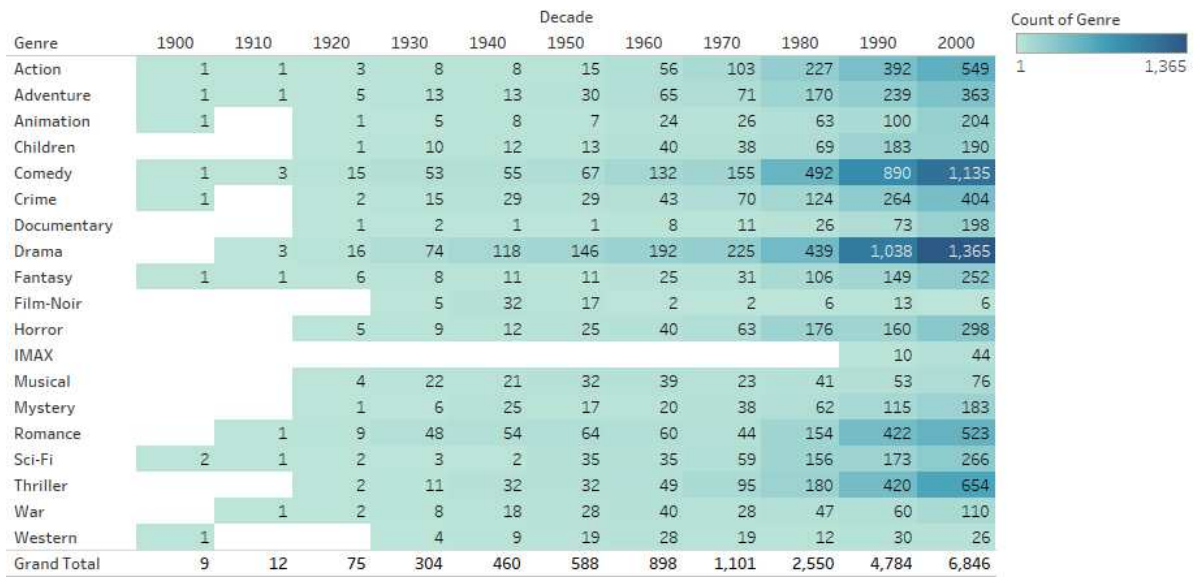
Técnicas baseadas em região

De acordo com Ward, Grinstein e Keim (2015), nas técnicas baseadas em região, os polígonos preenchidos são usados para transmitir valores, baseados em seu tamanho, forma, cor, ou outros atributos. Em especial, as representações mais comuns dessas técnicas, são:

- Gráficos de barras e histogramas;
- *Displays* tabulares (*heatmaps*, *TableLens*, matrizes reordenáveis); e
- Empilhamento dimensional.

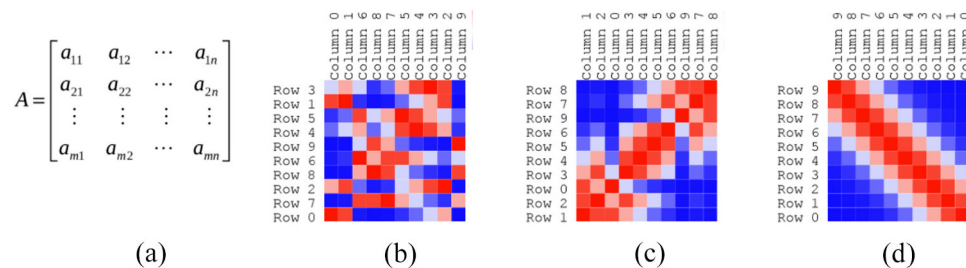
Em especial, a Figura 2.10 apresenta um *heatmap* contendo informações do conjunto de dados *MovieLens* (serviço de recomendação de filme). Nesse *heatmap*, as colunas são dispostas pelo agrupamento do ano por década, as linhas representam o gênero, e as células o total de filme, utilizando a cor para identificar a quantidade relativa (quanto mais escura, maior a quantidade).

Genre per Decade

Figura 2.10: *Heatmap* por gênero de filmes agrupados por década.

Fonte: Projeto realizado em disciplina de InfoVis da UNICAMP.

Para evidenciar informações ocultas em visualizações baseadas em matrizes, como os *heatmaps*, é possível em visualização de informação utilizar técnicas de reordenação de matrizes (BARONI; SILVA, 2022). A Figura 2.11 ilustra como diferentes arranjos da mesma matriz numérica podem exibir padrões diferentes (se existem em qualquer permutação possível da matriz), quando apresentados como um *heatmap*.

Figura 2.11: (a) Representação genérica de uma matriz; (b, c e d) representações gráficas baseadas em *heatmap*; (b) matriz embaralhada; e (c e d) matriz (b) reordenada de diferentes maneiras.

Fonte: (BARONI; SILVA, 2022).

Combinações de técnicas

Além das técnicas baseadas em pontos, linhas ou regiões, há um número de técnicas híbridas que combinam características de duas ou mais dessas classes (WARD; GRINSTEIN; KEIM, 2015). Dois grupos de técnicas desse tipo são:

1. *Glyphs* e ícones: representação visual de um registro de uma tabela de dados, em que uma entidade gráfica e seus atributos são controlados por um ou mais atributos de dados desse registro;
2. Técnicas orientadas a *pixel*: mapeiam cada valor da tabela de dados para *pixels* individuais e criam um polígono para representar cada variável dos dados.

2.1.5 Interação

Todas as técnicas apresentadas nas Seções 2.1.3 e 2.1.4 encontram o mesmo problema físico quando uma certa quantidade de dados é excedida: a falta de espaço para representar todos os dados destinados à visualização. Em virtude disso, ferramentas interativas podem ser extremamente úteis para endereçar esse problema. Segundo Mazza (2009), os sistemas de visualização de informações parecem ser mais úteis quando um usuário pode modificar os dados de entrada, alterar o mapeamento visual ou manipular a visualização gerada. As interações facilitam a exploração dos dados e podem revelar relações que poderiam permanecer ocultas em uma visão estática. Existem muitas classes de técnicas de interação (SPENCE, 2014; KEIM, 2002; WARD; GRINSTEIN; KEIM, 2015), as quais são controles do usuário para:

- **Navegação:** alterar a posição da tela e para escalar a visualização (o que é mapeado para a tela), como rotativo e *zoom*;
- **Seleção:** identificar um objeto, uma coleção de objetos, ou regiões de interesse a serem interesse de alguma operação, tais como destacando, apagando e modificando;
- **Filtragem:** reduzir o tamanho dos dados que estão sendo mapeados para a tela, seja eliminando registros, dimensões ou ambos;
- **Reconfiguração:** mudar como os dados são mapeados para entidades ou atributos gráficos, tais como reordenação dos dados ou leiautes, proporcionando assim uma maneira diferente de visualizar um subconjunto de dados;
- **Codificação:** alterar os atributos gráficos, tais como tamanho do ponto ou cor da linha, para potencialmente revelar características diferentes;
- **Conexão:** ligar diferentes visões ou objetos para mostrar itens relacionados;

- **Abstração e elaboração:** concentrar-se em um subconjunto de dados para obter detalhes (elaboração), ao mesmo tempo em que reduz o nível de detalhe (abstração) em outra parte dos dados.
- **Brushing e linking:** selecionar um ponto em uma visualização para destacar os pontos de dados correspondentes em outra visualização conectando visualizações heterogêneas; e
- **Híbrido:** combinar várias das técnicas acima em uma única técnica, por exemplo, aumentando o espaço de tela atribuído a uma ou mais áreas de foco para permitir que os usuários vejam os detalhes, enquanto mostram as outras áreas de dados em um espaço menor, para preservar o contexto.

O uso de técnicas interativas nos sistemas de visualização permite empregar o conceito *visual information seeking mantra* de Shneiderman (1996): primeiro visão geral; em seguida *zoom* e filtros; e, por fim, detalhes sob demanda. Esse conceito é especificamente útil para explorar e analisar dados complexos e/ou multidimensionais em que é impossível visualizar todos os aspectos importantes dos dados de uma só vez ou quando os dados contêm várias características (possivelmente dinâmicas) cuja compreensão é difícil em visualizações estáticas e não interativas.

2.1.6 Avaliação

O aspecto interativo (Seção 2.1.5) é desejado para a maioria das soluções de visualização; por isso, é importante conhecer as formas de avaliar essas soluções. A maioria dos métodos de avaliação deriva diretamente da área de Interação Humano-Computador (IHC), abordados de forma diferente na literatura (MAZZA, 2009; KERREN et al., 2008; WARD; GRINSTEIN; KEIM, 2015). Em geral, esses métodos podem ser classificados, segundo Kerren et al. (2008), como:

- **Qualitativos:** métodos que buscam alcançar uma compreensão mais rica, utilizando uma abordagem mais holística que considera a interação entre os fatores que influenciam as visualizações, seu desenvolvimento e seu uso. No centro dos métodos qualitativos está a habilidade e a sensibilidade com as quais os dados são coletados, que podem ser feitas de duas maneiras: observações e entrevistas (coleta de opiniões de participantes, grupo focal, e análise de casos de uso);

- **Quantitativos:** são aquelas metodologias em que a precisão é relativamente alta e em que algumas avaliações podem ser feitas sobre a possível generalização para uma população maior. Essas metodologias, incluem avaliação por diferentes conjuntos de heurísticas, percurso cognitivo, e revisão por pessoas experientes.

Segundo Kerren et al. (2008), esses métodos oferecem tanto vantagens quanto desvantagens. Uma parte importante nos trabalhos de InfoVis é selecionar os métodos mais apropriados para o seu escopo, seus propósitos, e seus usuários. Certamente, as avaliações fornecem as indicações necessárias para assegurar sistemas de visualização adequados ao uso e norteiam um ciclo de melhoria contínua.

De acordo com Carpendale (2008), apesar de ser importante escolher as perguntas certas para serem feitas na entrevista, ouvir ativamente o que o participante diz é a mais relevante de todas as habilidades de entrevista. A seguir, algumas recomendações de Carpendale (2008) sobre entrevistas:

- O entrevistador deve ter certeza de que compreendeu o que está sendo dito e que as descrições e explicações fornecidas são completas o bastante;
- O entrevistador deve conter a sua tendência de falar;
- Recomenda-se realizar pausas durante a entrevistas; às vezes, fazer anotações pode ser útil nesse tipo de situação;
- A entrevista deve ser moldada conforme as informações fornecidas pelo participante;
- O entrevistador deve evitar perguntas direcionadas. Uma das etapas fundamentais para minimizar o viés do pesquisador é redigir as perguntas com cautela, de modo a evitar qualquer tipo de insinuação de resposta; e
- O entrevistador deve fazer perguntas abertas.

2.2 Engenharia de petróleo

A análise de decisão relacionada ao desenvolvimento e gerenciamento de campos petrolíferos envolve riscos devido a várias incertezas, principalmente nos parâmetros de reservatório e fluido, modelo econômico, e disponibilidade operacional. O efeito dessas incertezas deve ser

combinado para estimar o risco envolvido nas decisões (SCHIOZER; SANTOS; DRUMOND, 2015). Esse processo não é simples de modelar devido ao tipo de decisões (projetos de longo prazo) e à complexidade do problema que envolve reservatório de petróleo, instalações de produção, avaliação econômica e técnicas estatísticas. O processo também pode ser muito demorado quando são utilizados modelos de alta fidelidade, e técnicas estatísticas específicas e avançadas.

Em particular, o desenvolvimento e o gerenciamento dos campos petrolíferos introduzem uma série de desafios tecnológicos que vão desde a aquisição de dados sísmicos, modelagem geológica e simulação computacional dos resultados ambientais e econômicos das estratégias adotadas, entre outros. Segundo Meira et al. (2016), no contexto computacional, vários processos de produção de petróleo podem ser modelados como problemas de otimização, nos quais uma multiplicidade de **variáveis incertas** (as variáveis de **saída** do sistema ⁶, i.e., valor presente líquido (*NPV*), fator de recuperação de óleo (*ORF*), produção acumulada de óleo (*Np*), produção acumulada de água (*Wp*), etc.) pode ser maximizada ou minimizada. Como os problemas de otimização resultantes estão longe de ser simples, eles raramente são problemas matemáticos bem definidos.

Dado o modelo geológico de um campo petrolífero, as empresas que possuem as licenças de exploração devem definir uma **estratégia de produção** a ser implementada nas próximas décadas de produção do campo. De acordo com Meira et al. (2017), essa estratégia de produção pode ser vista como a infraestrutura necessária de poços, tubulações e plataformas combinados com um cronograma que estabelece quando cada dispositivo começará a operar, e a localização exata de cada dispositivo no campo petrolífero. Uma estratégia de produção errada pode levar a graves perdas financeiras. Um dos principais desafios do processo de tomada de decisão nos campos petrolíferos é uma seleção adequada da estratégia de produção a ser implementada.

Para fazer essa seleção, engenheiros de petróleo dependem amplamente de simuladores. Eles reproduzem o desempenho do campo petrolífero com base em um modelo de simulação de reservatório e uma estratégia de produção, o qual é altamente dependente do modelo geológico. No entanto, esse modelo geológico contém variáveis incertas que podem afetar severamente a qualidade da estratégia de produção, criando um conjunto de cenários incertos. Neste contexto, de acordo com Meira et al. (2016), um **cenário** é definido como um

⁶Sistema é um modelo de simulação do reservatório e das instalações de produção, composto de variáveis de entrada e saída.

modelo geológico com valores fixos para todas as variáveis incertas e um **caso** x é a união de um cenário s e uma estratégia de produção P_r representada por $x = (s, P_r)$.

Quando uma determinada estratégia de produção é simulada, podem ser obtidos resultados diferentes para cada cenário distinto. Um conjunto de poços pode ser ideal para um cenário específico e inapropriado para outros, de modo que cada cenário geralmente tem sua própria estratégia ótima de produção. Portanto, o objetivo dos engenheiros de petróleo é obter uma estratégia de produção que maximize o retorno e reduza os riscos.

Com intuito de apoiar os engenheiros de petróleo a alcançarem esse objetivo, Schiozer, Santos e Drumond (2015) e Schiozer et al. (2019) propuseram uma metodologia com 12 passos a serem aplicados nos processos de análise de decisão relacionados ao desenvolvimento e gerenciamento de campos petrolíferos, integrando: simulação de reservatório, análise de risco, correspondência histórica, técnicas de redução de incerteza, modelos representativos e seleção de estratégia de produção sob incerteza (SANTOS; GASPAR; SCHIOZER, 2018).

A Seção 2.2.1 contém uma visão geral dessa metodologia. Após isso, nas Seções 2.2.2 e 2.2.3, são elencados os conceitos de curvas de risco e MRs, respectivamente. Por fim, na Seção 2.2.4, constam os detalhes do caso de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017), o qual é utilizado como referência para geração dos gráficos nesta seção.

2.2.1 Metodologia de 12 passos de Schiozer et al.

A metodologia de 12 passos de Schiozer, Santos e Drumond (2015) e Schiozer et al. (2019) é baseada no conceito de desenvolvimento e gerenciamento de fluxo fechado, conforme ilustrado na Figura 2.12.

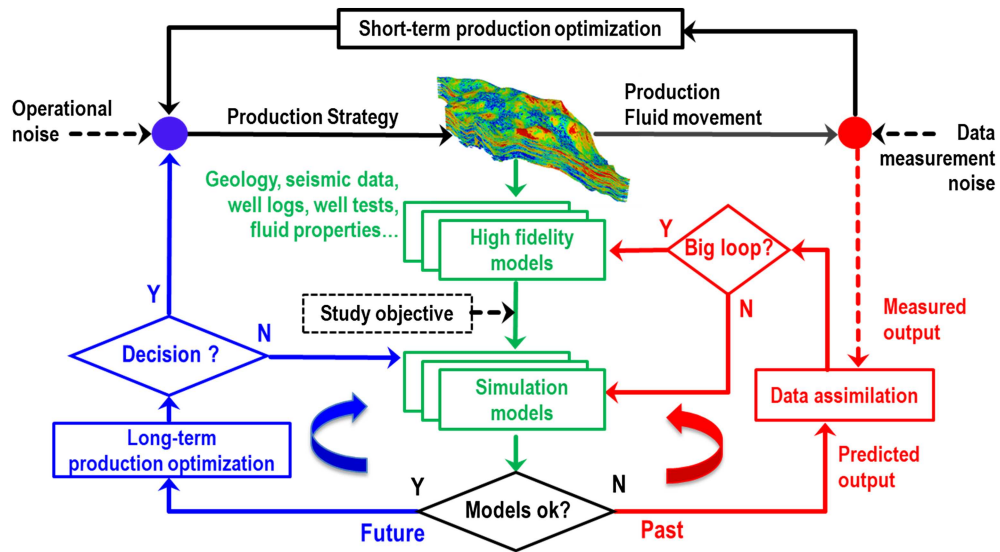


Figura 2.12: Desenvolvimento e gerenciamento de campo em fluxo fechado.

Fonte: (SCHIOZER et al., 2019)

De acordo com Schiozer et al. (2019), os principais componentes dessa metodologia, como se pode observar na Figura 2.12, são divididos em cores:

- Verde: coleta de todos os dados e incertezas, e construção de modelos;
- Vermelho: assimilação de dados, na qual todos os dados dinâmicos devem estar dentro de uma faixa de tolerância para selecionar modelos que serão usados na parte azul e;
- Azul: decisões baseadas em modelos, a longo prazo, sob incertezas;
- Preto: implementação de decisões de longo e curto prazo, definição do objetivo do estudo e seleção do tipo de estudo (análise de dados passados ou análise de decisões futuras).

A seguir são descritos, sucintamente, os 12 passos dessa metodologia:

Verde

1. Caracterização do reservatório sob incertezas (para construir modelos, desenvolver cenários e estimar probabilidades);
2. Construir e calibrar o modelo de simulação a partir de um caso base ⁷, chamado “Base 0”;

⁷O caso base é o cenário provável segundo a característica dos reservatórios.

Vermelho

3. Verificar inconsistências na “Base 0” e nos dados de poços a serem usados;
4. Gerar um conjunto de cenários considerando as incertezas dos reservatórios;
5. Explorar os dados e filtrar um subconjunto de cenários, selecionando um caso base para os próximos passos (“Base 1”);

Azul

6. Selecionar uma estratégia de produção para o caso base;
7. Estimar as curvas de risco para a estratégia de produção, considerando todos os possíveis cenários obtidos no passo 5;
8. Selecionar os MRs (conceito detalhado na Seção 2.2.3), reduzindo mais os cenários obtidos no passo 5;
9. Selecionar uma estratégia de produção para todos os MRs obtidos no passo 8;
10. Selecionar uma estratégia de produção sob incerteza, das obtidas no passo 9, incluindo as incertezas de reservatório, econômicas, operacionais e entre outras;
11. Aprimorar a estratégia obtida no passo 10 para melhor gerenciar as incertezas e melhorar a flexibilidade e a robustez; e

Preto

12. Gerar as curvas de risco finais e realizar a análise de decisão.

No passo 4, atributos incertos contínuos são discretizados durante o processo de tomada de decisão. Neste contexto, um atributo $a \in A$ pode ser discretizado em n_a níveis (conhecido como **atributo-nível**), e cada um desses níveis tem uma probabilidade $p_{a,l}$, associada a ele (sendo l o nível de atributo a). De acordo com Meira et al. (2020), dado que um determinado cenário é obtido com a definição de valores específicos para todos os seus atributos e se todos os atributos incertos forem discretizados, o número total de cenários em um determinado problema é dado pela Equação 2.1.

$$|\tau| = \prod_{a \in A} n_a \quad (2.1)$$

Na Equação 2.1, τ é o conjunto de cenários e A o conjunto de atributos do problema. O número de cenários aumenta exponencialmente, tanto com o número de atributos, quanto com o número de níveis. Um pequeno problema com dez atributos incertos, p. ex., discretizado em três níveis, resultaria em $|\tau| = 3^{10} = 59.049$ diferentes cenários para serem analisados.

No passo 5, usando dados de produção dos poços já no campo, o engenheiro de petróleo verifica a viabilidade de cada cenário do passo 4 e elimina os inviáveis, em um processo chamado assimilação de dados. Ou seja, aqueles cenários cujos dados de produção não são semelhantes aos do histórico de produção já conhecido, dentro de um limite de tolerância aceito, são descartados. A partir desse conjunto reduzido, um novo caso base (denominado “P50”) é construído.

2.2.2 Curvas de risco

No passo 7 da metodologia de 12 passos (Seção 2.2.1), são estimadas **curvas de risco**, considerando todos os cenários de risco obtidos no passo 5 e a estratégia de produção identificada no passo 6 para o caso base “P50”. As curvas de risco, que também são referidas como CCDF na literatura estatística, são construídas a partir das previsões de produção e do desempenho econômico de múltiplos cenários de reservatórios (SANTOS; SANTOS A; SCHIOZER, 2020). Dada uma estratégia de produção arbitrária P_r , as curvas de risco correspondem ao gráfico de cada variável de saída $f(s)$ em relação à probabilidade acumulada $acc^f : \mathbb{R} \rightarrow [0, 1]$, como dado na Equação 2.2 (MEIRA et al., 2020).

$$acc^f(s, \tau) = \frac{p(s, \tau)}{2} + \sum_{s' \in \tau : f(s') > f(s)} p(s', \tau) \quad (2.2)$$

Na Equação 2.2, τ é o conjunto de cenários, s um cenário, $f(s)$ o valor da variável de saída e $p(s, \tau)$ a probabilidade de cada cenário s em τ . Na Figura 2.13 é apresentado um exemplo de uma curva de risco da variável de saída NPV . Nessa curva de risco, cenários (pontos roxos) com valores baixos de NPV e valores altos de acc^{NPV} são denominados como pessimistas, enquanto cenários com valores altos de NPV e valores baixos de acc^{NPV} são denominados otimistas.

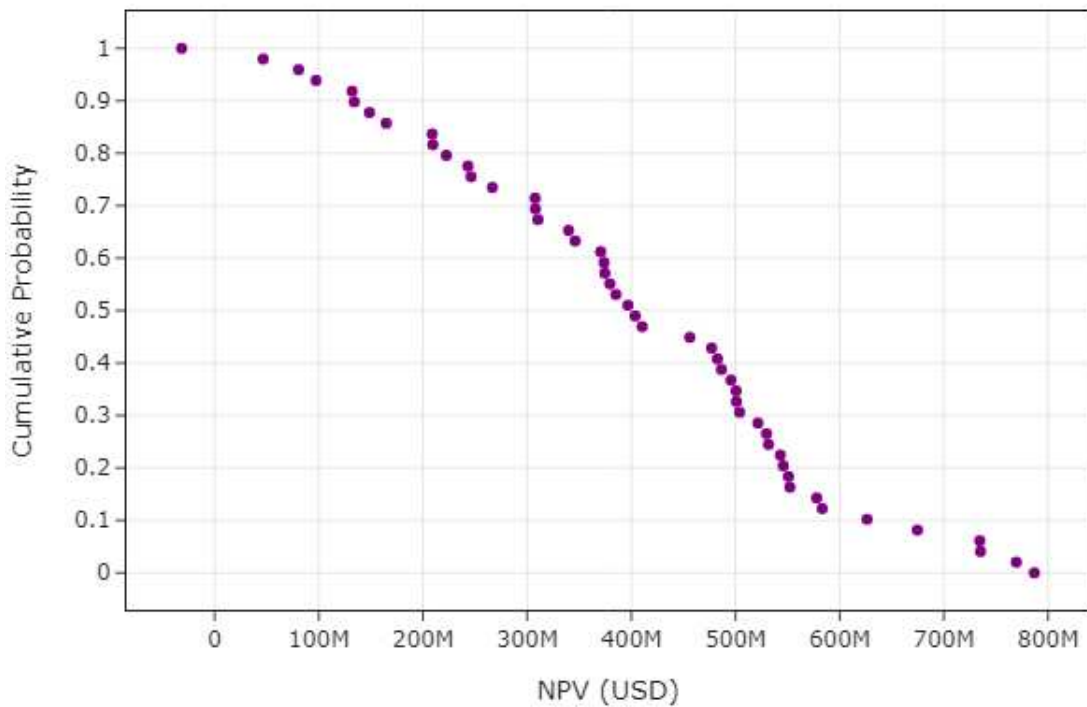


Figura 2.13: Curva de risco de uma variável de saída (nesse caso, NPV).

Fonte: Dados originais da pesquisa.

Observa-se que essas curvas de risco não refletem a avaliação final de risco da metodologia de 12 passos, mas fornecem insumos ao passo 8. As curvas de risco finais são geradas no passo 12 dessa metodologia, após MRs serem selecionados. Esses MRs devem representar adequadamente, entre outros aspectos, as curvas de risco originais (passo 7) de todas as variáveis de saída $f(s)$.

2.2.3 Modelos representativos

No passo 8 da metodologia de 12 passos (Seção 2.2.1), acontece uma nova redução no conjunto de cenários com a seleção de MRs. Esses MRs correspondem a um subconjunto dos cenários previamente identificados no passo 5, i.e., identificar os τ' cenários mais representativos (ou conjunto de MRs) dentro de um conjunto maior de cenários τ , onde $\{\tau' \subseteq \tau \wedge |\tau'| \ll |\tau|\}$. Segundo Santos et al. (2021), o conceito base de MR é que alguns modelos com características diferentes podem representar todo um sistema incerto de reservatório de petróleo.

O trabalho de Schiozer et al. (2004), pioneiro na área de engenharia de reservatórios, foi baseado na seleção manual de MR usando *cross-plots* para quatro variáveis de

saídas: NPV , ORF , Np e Wp . O critério de seleção foi escolher 9⁸ modelos com diferenças significativas na variável ORF que estivessem próximos aos percentis $P10$, $P50$ e $P90$ ou cenários (otimista, mais provável e pessimista, respectivamente) em relação à variável NPV (Figura 2.14). As distribuições Np e Wp , em relação ao NPV , foram utilizadas apenas para monitorar e assegurar uma ampla variação.

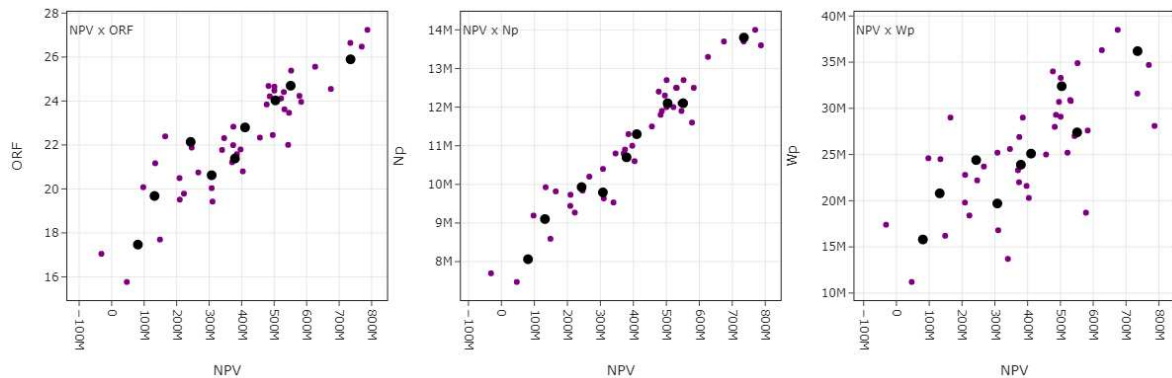


Figura 2.14: Seleção manual de nove MRs com base nos *cross-plots* das variáveis de saída (nesse caso, NPV , ORF , Np e Wp).

Fonte: Dados originais da pesquisa.

Com o número crescente de critérios na seleção de MRs, o processo manual tornou-se extremamente demorado e difícil de alcançar soluções satisfatórias. Isso despertou a necessidade de criar procedimentos automatizados para escolha de MRs, o que motivou o desenvolvimento de um método baseado em otimização implementado em um *software* chamado RMFinder (MEIRA et al., 2016).

Por fim, o interesse desse trabalho, segundo os objetivos assinalados na Seção 1.3, está restrito à seleção de MRs (passo 8 da metodologia de 12 passos) realizada de forma automatizada por meio do RMFinder. Contudo, pelo conceito de desenvolvimento e gerenciamento de campo em fluxo fechado (Figura 2.12), todos os componentes da metodologia de 12 passos estão interligados interativamente, devendo ser considerados na análise dos resultados desse trabalho.

⁸Embora grupos de pesquisa e empresas trabalhem com nove ou dez cenários para definir a estratégia de produção, o número ideal de cenários para cada problema ainda é uma questão em aberto na literatura.

2.2.4 Caso de estudo

Os dados utilizados para gerar os gráficos deste capítulo correspondem ao modelo de reservatório sintético OLYMPUS ⁹ (FONSECA; GEEL; LEEUWENBURGH, 2017). Esse modelo, inspirado em um campo de petróleo no Mar do Norte, foi desenvolvido com o propósito de ser utilizado como um caso de estudo para a otimização do desenvolvimento e gerenciamento de poços petrolíferos. Esse caso de estudo contém 50 imagens petrofísicas, cada uma delas representando um cenário distinto (os pontos roxos da Figura 2.14, p. ex.). Os seus mapas 3D para saturação de petróleo, permeabilidade e porosidade são ilustrados na Figura 2.15a-c, respectivamente.

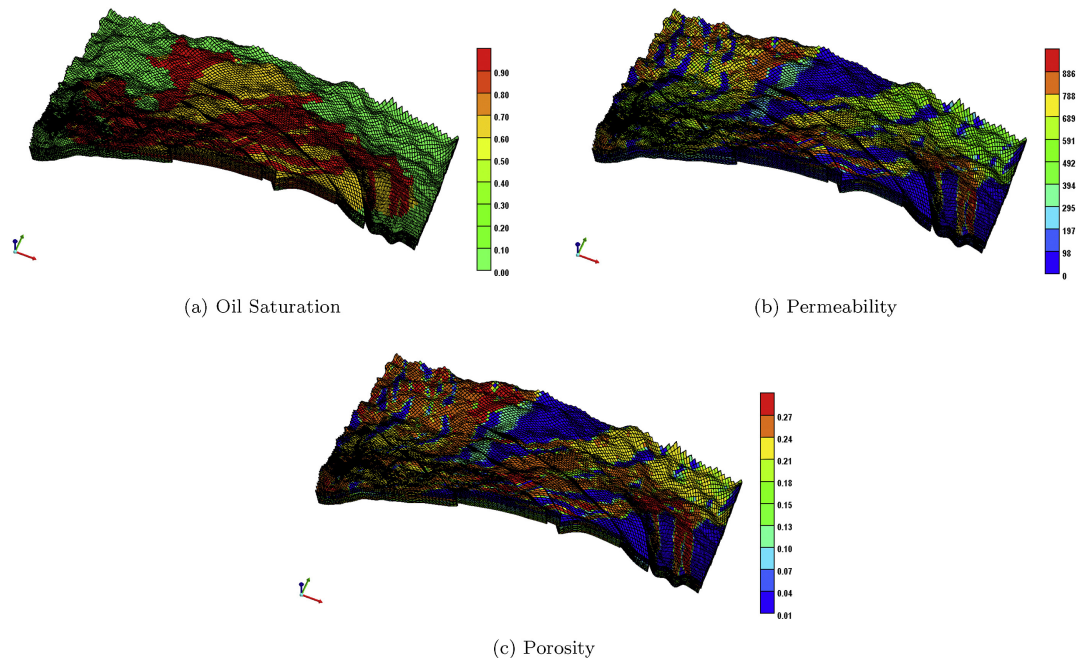


Figura 2.15: Mapas 3D do caso de referência OLYMPUS.
Fonte: (FONSECA; GEEL; LEEUWENBURGH, 2017)

2.3 RMFinder

O RMFinder é uma metodologia e um *software* para reduzir os cenários possíveis no processo de tomada de decisão de estratégias de produção de reservatórios de petróleo. Ele utiliza técnicas para selecionar automaticamente subconjuntos de cenários (ou seja, os chamados modelos representativos ou MRs) que representem as propriedades originais de um problema

⁹O caso de estudo OLYMPUS foi desenvolvido em um projeto conjunto entre TNO, Delft University of Technology, ENI, Statoil e Petrobras e colocado à disposição da comunidade científica

e possam ser aproveitados na definição da estratégia de produção, diminuindo assim o tempo computacional para realizar todas as simulações necessárias (MEIRA et al., 2016).

O RMFinder foi inicialmente proposto por Meira et al. (2016) e mais tarde foi sujeito a várias melhorias, levando ao lançamento do RMFinder 2.0 (MEIRA et al., 2017) e mais tarde do RMFinder 3.0¹⁰ (MEIRA et al., 2020), que incorpora e aprimora as características das demais versões. A presente Seção 2.3 ficará restrita a apresentar as funcionalidades e capacidades da versão 3.0.

O RMFinder está relacionado à metodologia de 12 passos para análise de decisão sobre reservatórios de petróleo proposta por Schiozer, Santos e Drumond (2015) e Schiozer et al. (2019), descrita na Seção 2.2.1. Em certos passos dessa metodologia (p.ex. análise de risco e assimilação de dados), centenas ou mesmo milhares de modelos são analisados. No entanto, em outros passos, o tomador de decisão deve comparar estratégias robustas e especializadas para obter mais informações sobre o problema e a flexibilidade das opções (SANTOS; GASPAR; SCHIOZER, 2018). Como a definição de estratégias especializadas tende a ser computacionalmente cara, essas estratégias são obtidas apenas para um conjunto de MRs (que é um subconjunto do conjunto principal de modelos). Então, o RMFinder auxilia na obtenção desses conjuntos de MRs, com o foco em automatizar o passo 8 dessa metodologia de 12 passos.

Em outras palavras, o RMFinder determina um conjunto de soluções $S = \{\tau'_1, \dots, \tau'_n\}$, ou simplesmente um conjunto de soluções $S = \{S_1, \dots, S_n\}$, onde cada S_i corresponde a um conjunto de MRs τ'_i , respectivamente, sendo $S_i \subseteq \tau$.

Na Seção 2.3.1 é detalhada a modelagem matemática da função multiobjetivo otimizada pelo RMFinder e na Seção 2.3.2 constam, como exemplo, as visualizações produzidas pelo RMFinder após uma execução.

2.3.1 Modelagem matemática

No passo 8 da metodologia de Schiozer et al. (2019), o analista de petróleo é responsável por definir o número desejado de MRs a serem selecionados dentre os que compõem o conjunto de modelos τ inicialmente informado ao RMFinder. Além disso, o analista deve também informar um conjunto de pesos, que equilibram a importância de diferentes aspectos que compõem

¹⁰O RMFinder 3.0 está registrado no Instituto Nacional da Propriedade Industrial (INPI) sob o processo de número BR512019002982-8.

o conceito de distorção de representatividade adotado pelo RMFinder (MEIRA et al., 2016). O conceito de distorção de representatividade de um dado conjunto τ' de MRs é definido por uma única função matemática composta por quatro componentes normalizados, i. e., os critérios (MEIRA et al., 2020):

- (i) $F_{cross}(\tau')$, responsável pela distorção da representação em relação às variáveis do problema;
- (ii) $F_{risk}(\tau')$, associado às curvas de risco do problema;
- (iii) $F_{atr}(\tau')$, relacionados à distribuição em níveis de atributos; e
- (iv) $F_{penalty}(\tau')$, que retorna o número de atributos-nível do problema que não estão presentes no conjunto de MRs.

Esses quatro componentes são combinados na função objetivo $F(\tau')$ por meio de uma soma ponderada com quatro pesos definidos pelo usuário (q_1 , q_2 , q_3 e q_4 , respectivamente), como apresentado na Equação 2.3. Neste contexto, o conjunto de MRs com maior representatividade é o que minimiza $F(\tau')$. Em outras palavras, a Equação 2.3 é a função multiobjetivo do problema de otimização solucionado pelo RMFinder (MEIRA et al., 2020).

$$F(\tau') = q_1 \cdot F_{cross}(\tau') + q_2 \cdot F_{risk}(\tau') + q_3 \cdot F_{atr}(\tau') + q_4 \cdot F_{penalty}(\tau') \quad (2.3)$$

Na Equação 2.3, $q_i \in \mathbb{R}^+$ são pesos que o usuário pode mudar para aumentar ou diminuir a importância de uma boa distribuição dos MRs. Segundo Meira et al. (2020), a Equação 2.3 revela duas das principais características do RMFinder: uma abordagem multicritérios para identificar modelos representativos, uma vez que combina quatro componentes em sua função objetivo principal; e uma abordagem interativa, pois o analista de petróleo tem um papel fundamental no processo via definição dos pesos q_i , i. e., trata-se de uma otimização interativa (SILVA et al., 2017). A cada vez que o analista altera esses pesos, ele deve solicitar uma nova execução do RMFinder, que por sua vez proverá um novo conjunto de resultados. Portanto, segundo Santos et al. (2021), pode-se considerar o uso do RMFinder como parte de um processo de otimização interativa, que visa conciliar a otimização provida e os elementos externos a ela que influenciam a decisão a ser tomada e as novas parametrizações dessa otimização, conforme experiência do analista no domínio do problema.

Função objetivo *cross-plot*

O primeiro componente da Equação 2.3 é a função *cross-plot* ($F_{cross}(\tau')$), que visa avaliar a distorção de representatividade do conjunto de MRs em relação às variáveis de saída do problema. O RMFinder considera quatro variáveis de saída (NPV , ORF , Np e Wp), que permitem a representação de cada cenário com um ponto em $\mathbb{R}^4$. De acordo com Meira et al. (2020), o conjunto de MRs idealmente pode ser representado por uma “nuvem” de pontos em $\mathbb{R}^4$. No entanto, o RMFinder foi desenvolvido para simular o processo manual de identificação de MRs, i. e., ao invés de considerar o espaço original em $\mathbb{R}^4$, ele trabalha com seis projeções de “nuvem” de pontos em $\mathbb{R}^2$, os chamados *cross-plot*. Essas projeções podem ser considerados como gráficos de dispersão com os pares de eixos $NPV \times ORF$, $NPV \times Np$, $NPV \times Wp$, $ORF \times Np$, $ORF \times Wp$ e $Wp \times Np$. Ou seja, esse conjunto de projeções é muito similar ao da técnica de matrizes de gráficos de dispersão (MAZZA, 2009), porém com apenas metade dos gráficos (p. ex., omitindo $B \times A$ enquanto mantém $A \times B$). Para cada *cross-plot*, o conjunto de MRs deve ser distribuído sobre a nuvem de pontos original de forma uniforme. Então, para cada conjunto de MRs τ' , $F_{cross}(\tau')$ é dado por:

$$F_{cross}(\tau') = q_1^c \cdot F_{NPV,ORF}^{cross}(\tau') + q_2^c \cdot F_{NPV,Np}^{cross}(\tau') + q_3^c \cdot F_{NPV,Wp}^{cross}(\tau') + q_4^c \cdot F_{ORF,Np}^{cross}(\tau') + q_5^c \cdot F_{ORF,Wp}^{cross}(\tau') + q_6^c \cdot F_{Wp,Np}^{cross}(\tau') \quad (2.4)$$

Na Equação 2.4, $q_i^c \in \mathbb{R}^+$ são pesos atribuídos pelo usuário para indicar quais *cross-plots* são mais importantes no processo e $F_{f,g}^{cross}(\tau')$ é a distorção de erro quadrático do *cross-plot* $f \times g$. Esta distorção de erro quadrático para um determinado *cross-plot* é representada pela soma das distâncias euclidianas quadradas entre cada cenário e o MR mais próximo, conforme descrito na Equação 2.5, dado um cenário s e um conjunto de cenários τ' cujo $\tau' \subseteq \tau$.

$$F_{f,g}^{cross}(\tau') = \sum_{s \in \tau'} d_{f,g}(s, \tau') \quad (2.5)$$

Função objetivo curva de risco

De acordo com Meira et al. (2020), o segundo termo da Equação 2.3 é a função objetivo das curvas de risco ($F_{risk}(\tau')$), dada pela Equação 2.6. O objetivo dessa equação é obter o conjunto

de probabilidades P que minimiza, simultaneamente, $\Delta^f(P)$ para todas as variáveis de saída $f \in (f_1, \dots, f_m)$.

$$F_{risk}(\tau') = \min_{\forall P} \left\{ \sum_{i=1}^m q_i^r \Delta^{f_i}(P) / Q \right\} \quad (2.6)$$

Na Equação 2.6, $q_i^r \in \mathbb{R}^+$ são pesos selecionados pelo analista de petróleo para aumentar ou diminuir a importância da distribuição dos MRs em uma curva de risco específica; $\{m \in \mathbb{N} \mid 1 \leq m \leq 50\}$ corresponde à quantidade de variáveis de saída otimizadas simultaneamente pelo RMFinder; $\Delta^f(P)$ é a função distância baseada no conjunto de probabilidades P dos MRs em τ' , descrita na Equação 2.7; e $Q = \sum_{i=1}^m q_i^r$. Considere na Equação 2.7 o conjunto de cenários $T = \{s_1, \dots, s_n\}$, onde cada cenário $s \in T$ tem uma probabilidade $1/n$, sendo $n = |T|$. Segundo Meira et al. (2016), também considere-se a função $f^* = (f_1, \dots, f_m)$, definida como $f^* : T \rightarrow \mathbb{R}^m$, dada uma variável de saída f .

$$\Delta^f(P) = \frac{1}{n} \sum_{s \in T} |f(s) - f(rm^f(s))| \quad (2.7)$$

Dado que o conjunto de MRs deve representar adequadamente o conjunto original de cenários T , então cada MR r representa um subconjunto de cenários em T . Portanto, é possível definir $rm^f(s)$ pela Equação 2.8 como sendo qual o MR r que representa um dado cenário s , considerando uma variável de saída f .

$$rm^f(s) = \left\{ r \mid r \in \tau' \wedge p_c^f(r, \tau') - \frac{p(r)}{2} \leq p_c^f(s, T) < p_c^f(r, \tau') + \frac{p(r)}{2} \right\} \quad (2.8)$$

Na Equação 2.8, $p(r)$ representa a probabilidade individual $p(r, \tau')$ e cada MR tem uma probabilidade cumulativa $p_c^f(r, \tau')$, dada pela Equação 2.2. Os mecanismos por trás do cálculo da Equação 2.8 são ilustrados na Figura 2.16 para a variável NPV, onde os pontos roxos representam os cenários originais em T , os pontos pretos representam os MRs e o tamanho das barras verticais indica a probabilidade individual de cada MR. Se todas as probabilidades envolvidas na Figura 2.16 forem consideradas, é possível observar que cenário $s \in T$ é representado por $r \in \tau'$, i. e., a probabilidade cumulativa de s está dentro do intervalo definido pela barra vertical de r .

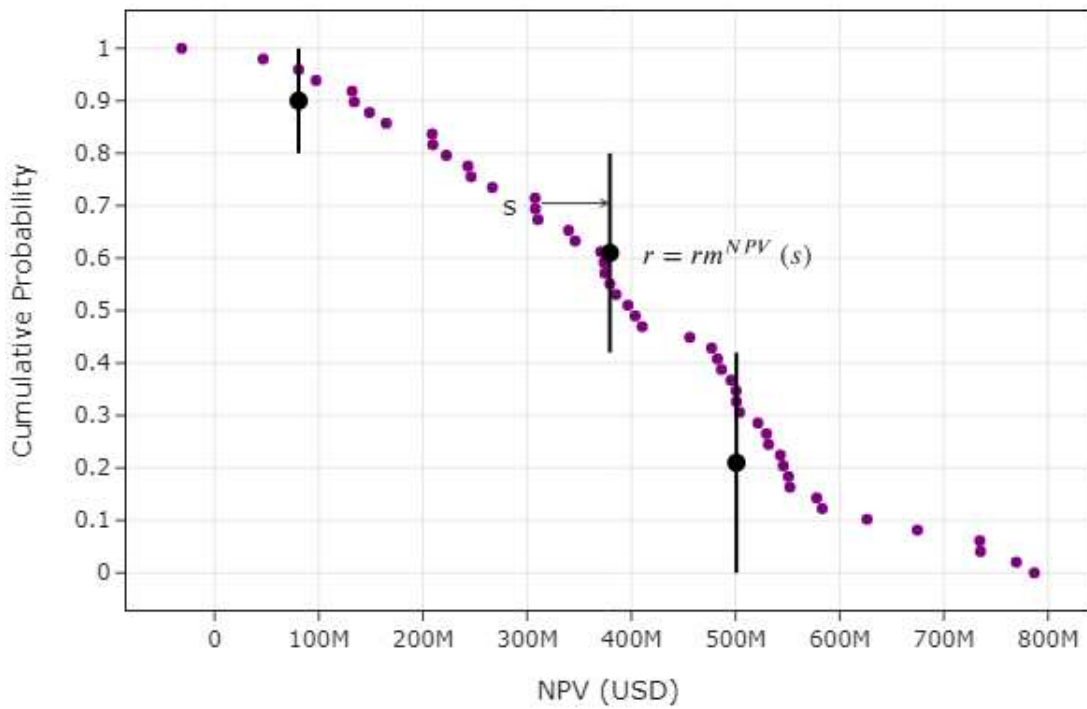


Figura 2.16: Cálculo de $rm^f(s)$ para variável de saída NPV .

Fonte: Dados originais da pesquisa.

Sendo assim, a diferença entre as curvas de risco geradas pelos modelos em T e pelos MRs pode ser estimada pela Equação 2.7. A lógica por trás da Equação 2.7 é ilustrada na Figura 2.17, que pode ajudar a elucidar o funcionamento do cálculo das distâncias mencionadas. $\Delta^f(P)$ é dado pela soma do tamanho das linhas horizontais cinzas associadas a cada $s \in T$ nessa figura, multiplicadas pela probabilidade $1/n$. Os valores das variáveis de saída $f^* = (f_1, \dots, f_m)$ são normalizados, de modo que a distância entre o primeiro e terceiro quartil torne-se 1.

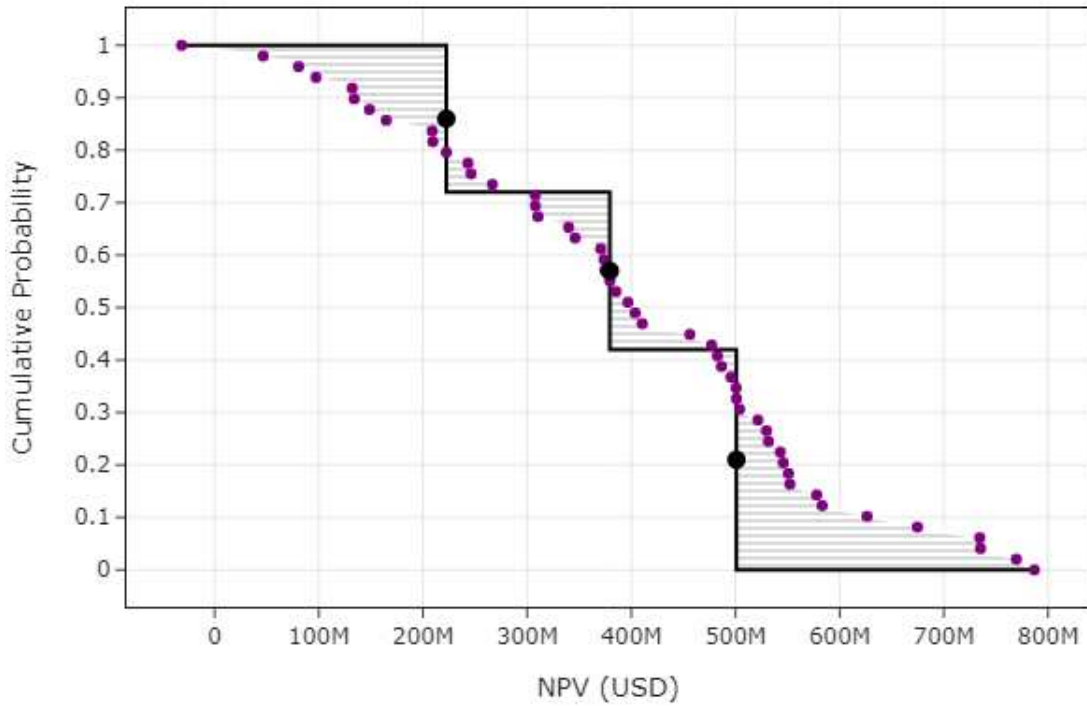


Figura 2.17: Linhas cinza representando o cálculo da diferença entre curvas de risco para a variável de saída *NPV*.

Fonte: Dados originais da pesquisa.

Função objetivo atributo-nível

O terceiro componente da Equação 2.3 é a função objetivo atributo-nível $F_{atr}(\tau')$, que representa a diferença da distribuição dos atributos-nível no conjunto de MRs τ' em comparação com o conjunto original de cenários τ (MEIRA et al., 2020). A função objetivo atributo-nível $F_{atr}(\tau')$ é detalhada na Equação 2.9.

$$F_{atr}(\tau') = \sum_{a \in A} \sum_{l=1}^{n_a} \left(p_{a,l}^{\tau} - p_{a,l}^{\tau'} \right)^2 \quad (2.9)$$

Na Equação 2.3, a é um atributo incerto do problema, n_a é o número de níveis em que a foi discretizado, $p_{a,l}^{\tau}$ é a probabilidade de nível l do atributo a no conjunto original τ e $p_{a,l}^{\tau'}$ é a probabilidade de nível l do atributo a no conjunto reduzido τ' . A intuição por trás desses cálculos é a necessidade de comparar distribuições de valores de cada atributo incerto. É necessário comparar as distribuições desses atributos existente nos modelos do conjunto original τ e a existente nos modelos representativos em τ' . É desejado que essas distribuições sejam o mais parecidas possível (para cada atributo), como uma das formas de garantir a representatividade dos modelos escolhidos.

Função penalidade

Por fim, o último componente da Equação 2.3, $F_{penalty}(\tau')$, penaliza conjuntos de MRs que não contêm pelo menos uma ocorrência de cada nível de atributo. A função $F_{penalty}(\tau')$ retorna θ , onde θ é o número de atributos-nível não encontrado nos MRs do conjunto τ' (MEIRA et al., 2020).

Otimização

Em resumo, a função objetiva utilizada pelo RMFinder considera, de forma conjunta:

- a) A representatividade de cada MR em projeções de pares de variáveis de saída;
- b) A representatividade de cada MR em curvas de risco baseadas em probabilidades acumuladas; e
- c) A representatividade do conjunto de MRs com relação à distribuição de valores de seus atributos incertos (incluindo penalidades por situações adversas que ocorrerem nessa distribuição).

Para obter o conjunto de MRs, o RMFinder utiliza meta-heurísticas para realizar a tarefa de otimização de $F(\tau')$. Em particular, o RMFinder adota o algoritmo de inspiração imunológica cob-aiNet[C]. Como os Algoritmos Genéticos (AG), o cob-aiNet[C] é bioinspirado, mas sua inspiração vem do sistema imunológico natural dos vertebrados em vez da teoria da seleção natural que norteia os AGs (MEIRA et al., 2016).

2.3.2 Visualizações do RMFinder

O RMFinder gera um relatório gráfico ao final da sua execução, visando ajudar os usuários a tomarem decisões para encontrarem as melhores soluções. Esse relatório é composto por três tipos de gráficos diferentes; são eles: gráficos de curva de risco, gráficos de *cross-plot* e o gráfico de atributo-nível.

Gráficos de curvas de risco

De acordo com Silva et al. (2020), a comparação de curvas de risco é uma abordagem usada para ajudar engenheiros de petróleo a selecionar um conjunto de MRs. O RMFinder gera um gráfico

de curva de risco para cada uma das variáveis da solução. Uma curva de risco é normalmente traçada em um gráfico de linhas, onde o eixo X representa uma variável de saída f , enquanto o eixo Y corresponde à probabilidade cumulativa de f (CCDF), dada pela Equação 2.2. Isso é feito com a curva de risco que representa modelos representativos. Por sua vez, quando se trata do conjunto original de modelos τ , marcas sem linhas (de forma similar a um gráfico de dispersão) são desenhadas para representar f e sua probabilidade. Na Figura 2.18 apresenta-se um exemplo de duas curvas de risco para a variável de saída NPV (uma curva com 50 modelos e outra com 10 modelos).

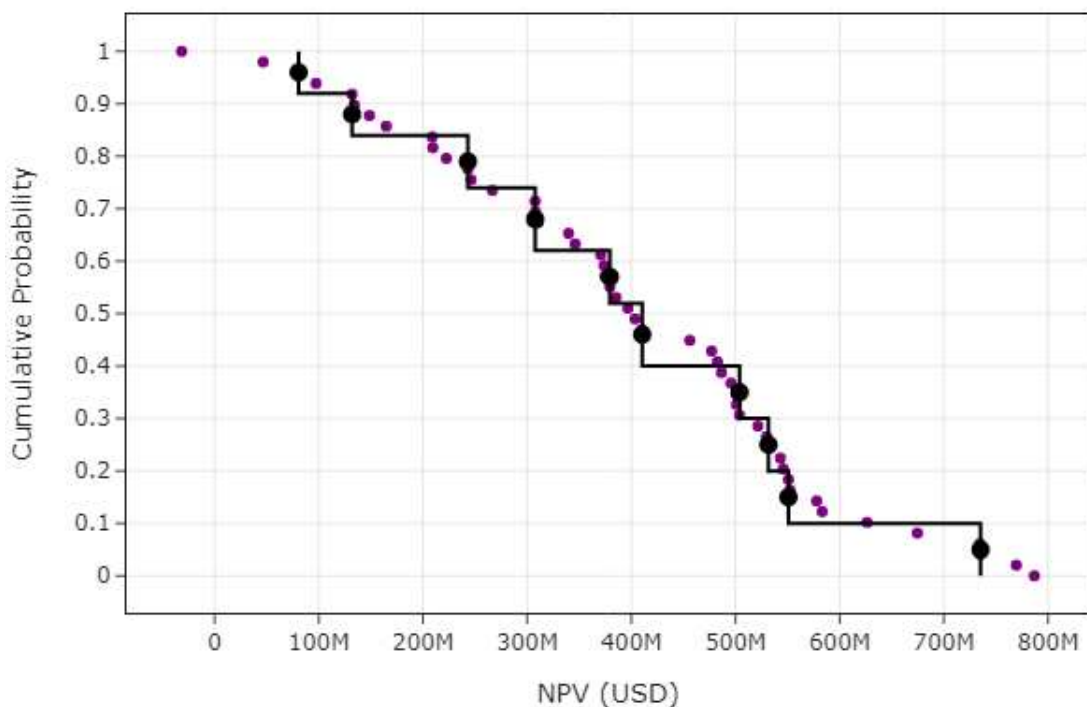


Figura 2.18: Dois CCDFs da variável de saída NPV , um para 50 modelos (pontos roxos) e outro para 10 modelos (linha preta).

Fonte: Dados originais da pesquisa.

Em Silva et al. (2020) são descritas quatro abordagens para a comparação de curvas de risco em um mesmo gráfico: superposição, sumarização, justaposição e abordagens híbridas.

A superposição (Seção 2.1.4) é a abordagem usada para gerar gráficos de curvas de risco combinadas, presente no RMFinder 3.0 (MEIRA et al., 2020). Na Figura 2.19 observa-se a sobreposição de um conjunto de dez curvas de risco para variável de saída NPV nas melhores soluções encontradas em uma única execução do RMFinder. Embora as linhas pretas estejam desordenadas (ou sobrepostas), esse gráfico apresenta uma visão geral das curvas de risco, e permite aos analistas comparar as suas semelhanças.

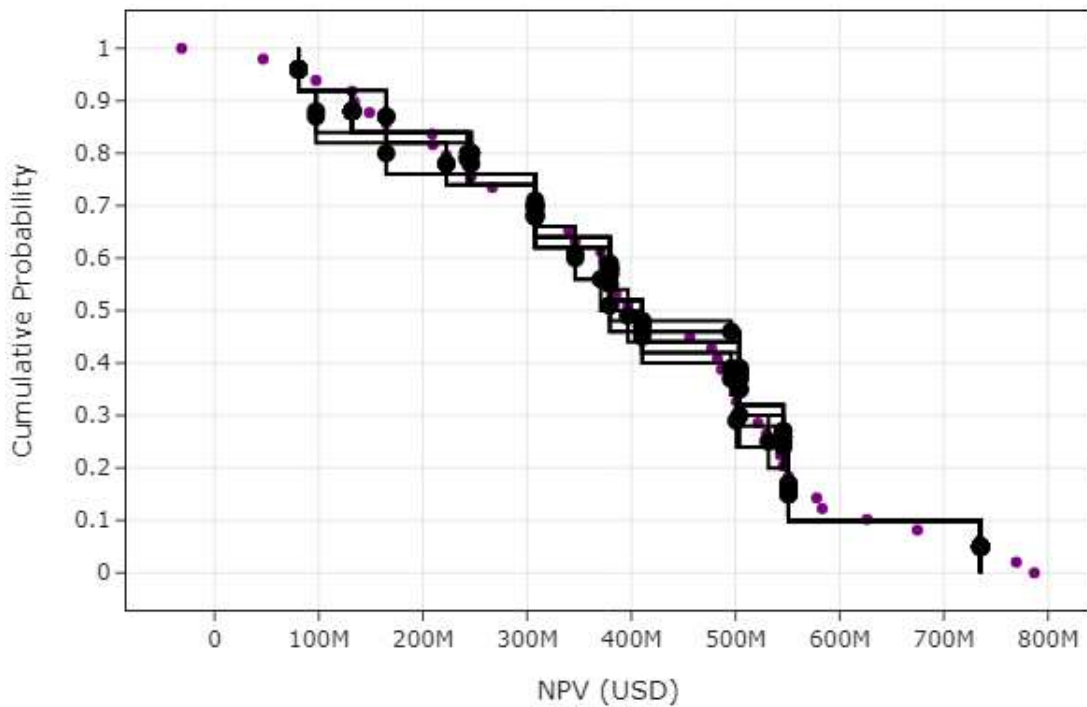


Figura 2.19: Gráfico com dez curvas de risco combinadas (sobrepostas) para a variável de saída *NPV*.

Fonte: Dados originais da pesquisa.

A abordagem de sumarização utiliza um gráfico de dispersão (Seção 2.1.4), em que cada marca representa uma curva de risco. Os eixos desse gráfico representam o risco e a medida de retorno. Santos, Gaspar e Schiozer (2017) mapeiam “Risco de *NPV*” para eixo *X*, “*EMV*” para eixo *Y*, e marcas para estratégias de produção, onde cada estratégia tem uma curva de risco para o *NPV*. A abordagem de sumarização é a única não usada no momento no RMFinder, dentre as mencionadas.

Uma abordagem de justaposição (Seção 2.1.4) coloca os gráficos de curvas de risco lado a lado em uma disposição matricial. Como exemplo, Meira et al. (2020) apresentam um gráfico com 25 pequenos gráficos de curvas de risco em uma matriz de 5×5 células. Cada um desses pequenos gráficos contém duas curvas de risco sobrepostas: um para todo o conjunto de modelos, e outra para os dez MRs. A Figura 2.20 exemplifica esses 25 gráficos de curvas de risco justapostos, uma para cada variável de saída, criados pelo RMFinder. Os pontos e linhas pretas, e os pontos roxos referem-se aos MRs e a todo o conjunto de modelos, respectivamente.

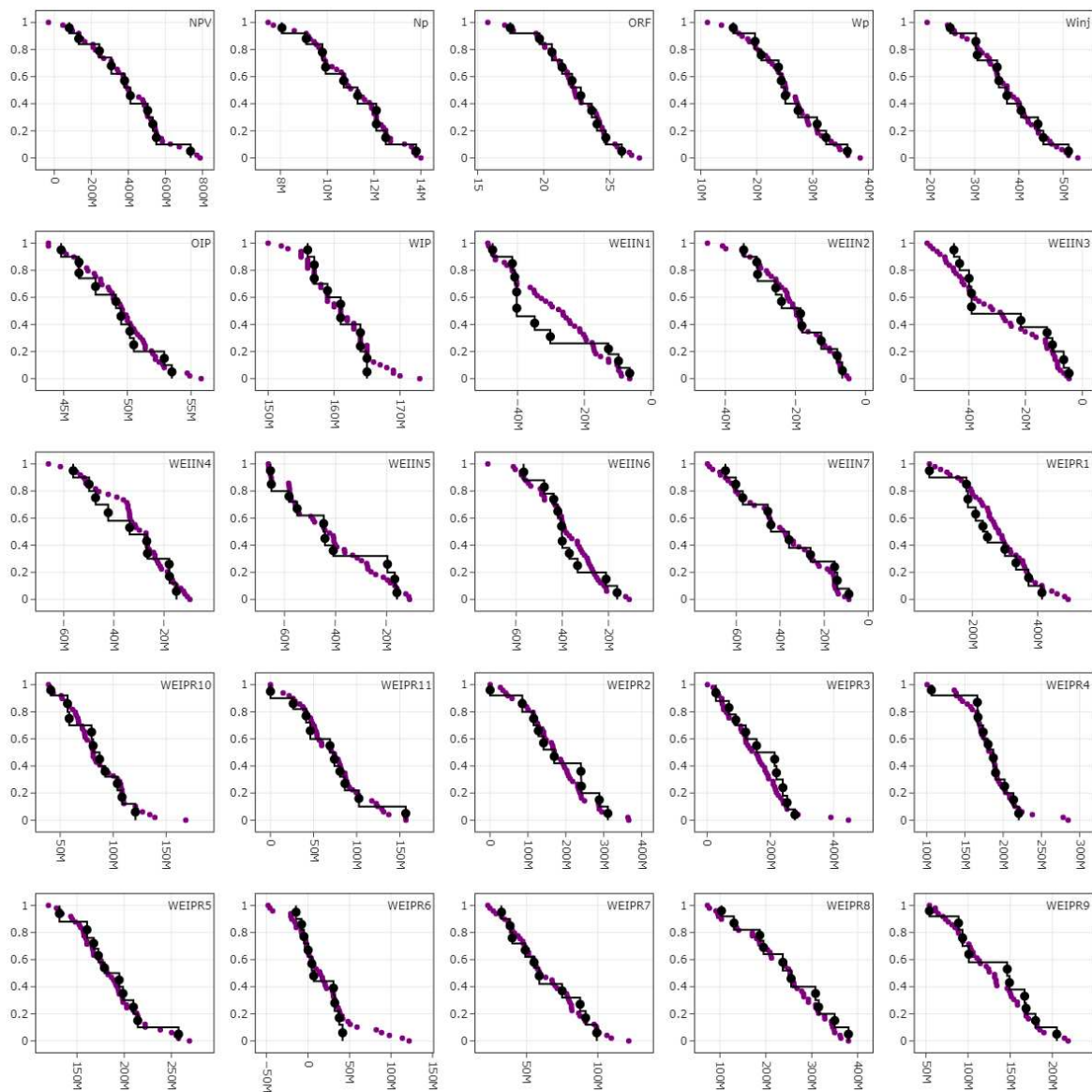


Figura 2.20: Gráficos com curvas de risco justapostas de 25 variáveis de saída. Os MRs e todo o conjunto de modelos são representados como pontos e linhas pretas, e pontos roxos, respectivamente.

Fonte: Dados originais da pesquisa.

Uma abordagem híbrida (Seção 2.1.4) é também proposta por Meira et al. (2020). Essa abordagem utiliza justaposição e sobreposição para criar uma matriz de todos os gráficos de curva de risco combinadas. A Figura 2.21 exemplifica a justaposição e sobreposição de 25 gráficos de risco, uma para cada variável de saída, para mostrar as melhores 10 soluções geradas pelo RMFinder. Os pontos e linhas pretas, e os pontos roxos são semelhantes aos da Figura 2.20, exceto que existem dez linhas pretas sobrepostas por gráfico, uma para cada solução.

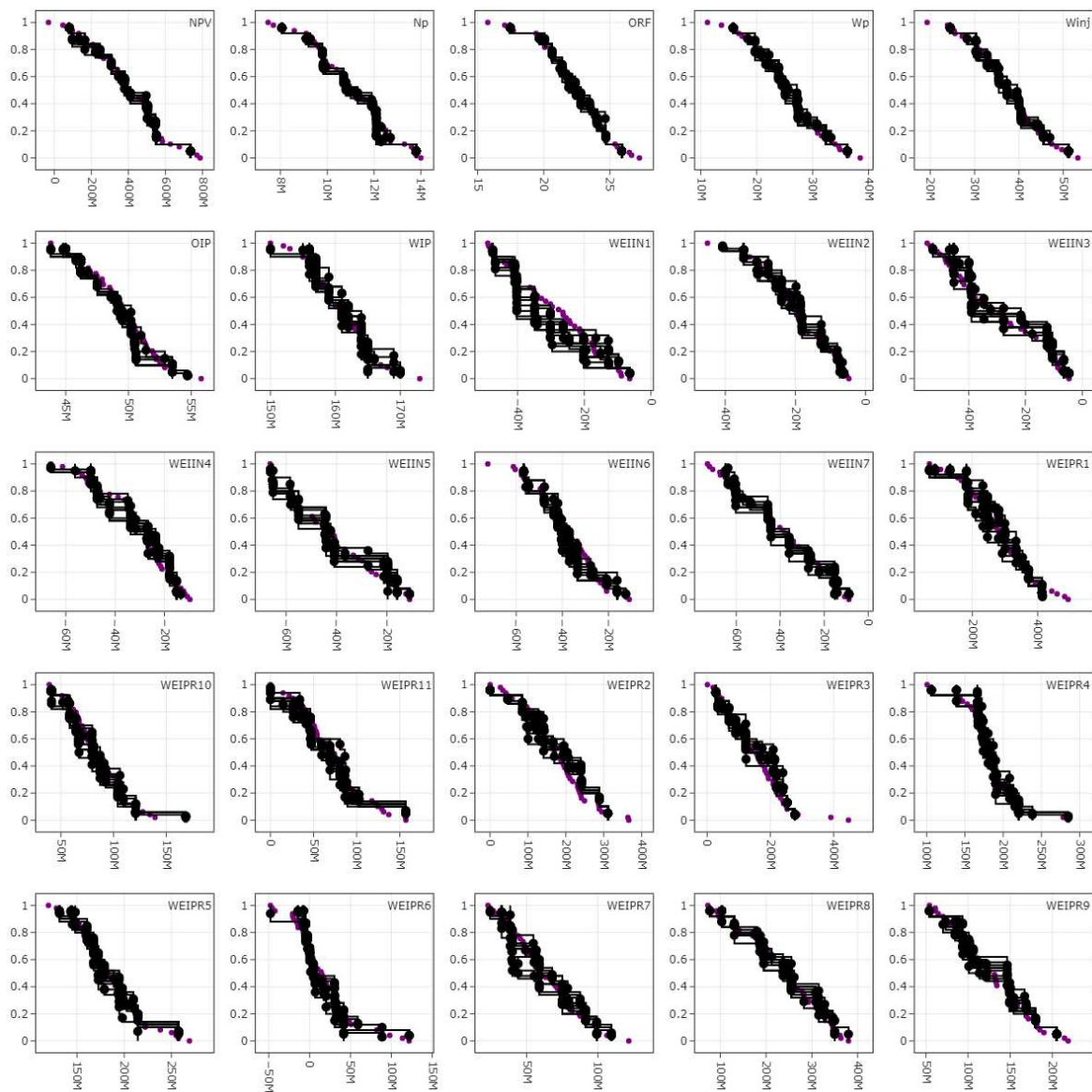


Figura 2.21: Gráficos de curvas de risco justapostas e sobrepostas de 25 variáveis de saída. Os MRs e as dez soluções, e todo o conjunto de modelos são representadas como pontos e linhas pretas, e pontos roxos, respectivamente.

Fonte: Dados originais da pesquisa.

Gráficos de *cross-plot*

Os *cross-plots* são gráficos de dispersão (Seção 2.1.4) e avaliam a representatividade do conjunto de MRs em relação às variáveis de saída, par a par (MEIRA et al., 2020). Entretanto, o RMFinder gera gráficos de *cross-plot* considerando apenas quatro variáveis de saída: *NPV*, *ORF*, *Np* e *Wp*. A Figura 2.22 representa os *cross-plots* dessas variáveis de saída e dez MRs de uma solução. Os pontos pretos e os pontos roxos referem-se aos MRs e a todo o conjunto de modelos, respectivamente.

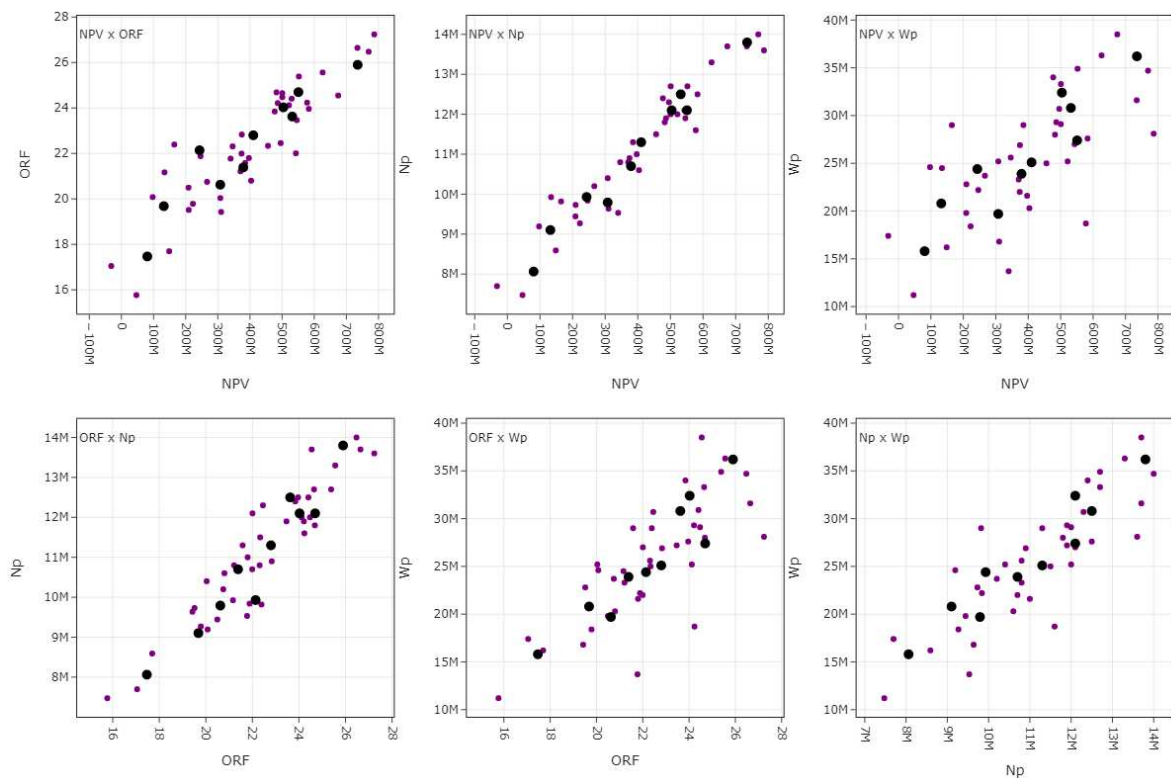


Figura 2.22: Gráficos de *cross-plot* de quatro variáveis de saída. Os MRs e todo o conjunto de modelos são representados como pontos pretos e pontos roxos, respectivamente.

Fonte: Dados originais da pesquisa.

Gráficos de diferença de atributo-nível

O gráfico de diferença de atributo-nível avalia o quão diferente é a distribuição dos níveis de atributo no conjunto de MRs em relação todo o conjunto de modelos por meio de um grupo de gráficos de barras (Seção 2.1.4), apresentado na Figura 2.23. Cada linha do gráfico representa um atributo, cada coluna um nível, e cada barra (região em azul) indica as diferenças de frequência para cada nível. Isso significa que quanto maior a barra, maior é a diferença de frequência entre o conjunto de MRs e todo o conjunto de modelos (MEIRA et al., 2016).

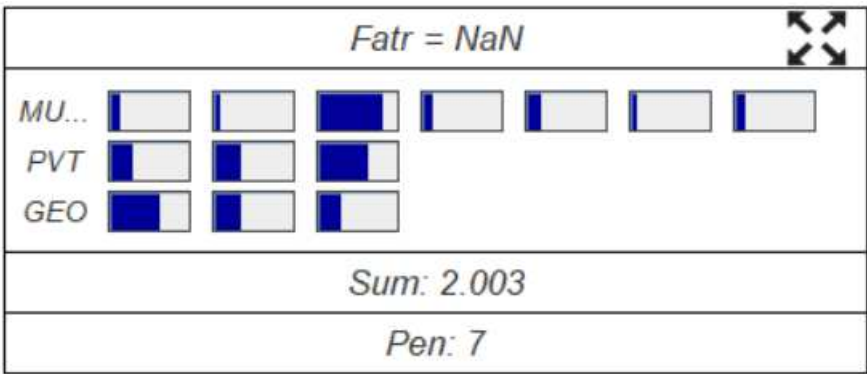


Figura 2.23: Gráfico de diferença de atributo-nível.
Fonte: (AMARAL, 2021).

2.4 Considerações finais

O Capítulo 2 apresentou o referencial teórico relacionado à pesquisa em questão. Foram apresentados conceitos sobre visualização de informação na Seção 2.1, dando destaque à análise multivariada. Em seguida, dissertou-se na Seção 2.2 sobre tomadas de decisões no desenvolvimento e gerenciamento de campos petrolíferos. A última Seção 2.3 abordou a metodologia e o *software* RMFinder, dando enfoque na modelagem matemática da função objetivo multicritério e nas visualizações gráficas geradas ao final de cada execução.

Capítulo 3

Trabalhos relacionados

Dada a importância de selecionar MRs nos processos de tomada de decisão nos campos petrolíferos, várias estratégias foram propostas nos últimos anos. Steagall e Schiozer (2001) concentraram-se em uma única curva de risco (no caso, para a variável incerta NPV) e propuseram que o conjunto de MRs fosse constituído por sete modelos, sendo dois próximos a $P90$, três próximos a $P50$ e dois próximos a $P10$ (ou modelos com probabilidade acumulada próxima a 90%, 50% e 10%, respectivamente). Schiozer et al. (2004) propuseram uma extensão da metodologia Steagall e Schiozer (2001), com o requisito adicional de selecionar os MRs próximos a estes percentis não apenas com relação ao NPV , mas também ao Np , Wp e ORF , com auxílio de gráficos de *cross-plots* (Seção 2.2.3). Com o número crescente de critérios nesse processo (detalhes na Seção 2.2), junto à exigência de alcançar soluções ainda mais satisfatória, novos trabalhos acadêmicos foram publicados para buscar alternativas a esses pontos. Nesse sentido, foram elaboradas soluções para automatizar a seleção de MRs com uso de técnica de otimização baseada em meta-heurísticas (MEIRA et al., 2016, 2017, 2020); e trabalhos de InfoVis aplicada com o intuito de ajudar os engenheiros de petróleo (ou tomadores de decisão) a melhor comparar e analisar as características de MRs.

Este Capítulo 3 tem por objetivo realizar uma breve discussão sobre os trabalhos relacionados à InfoVis aplicados na seleção de MRs. Em virtude disso, na Seção 3.1 consta a metodologia proposta para encontrar esses trabalhos e na Seção 3.2 eles são apresentados.

3.1 Revisão da literatura

A metodologia utilizada na revisão bibliográfica para encontrar as referências na literatura associadas a este trabalho, e também para possibilitar a reprodutibilidade por parte da comunidade científica, emprega seis etapas, adaptadas do framework PRISMA (MOHER et al., 2010); são elas:

1. Definição do tópico de pesquisa;
2. Definição dos conceitos envolvidos;
3. Seleção das fontes bibliográficas;
4. Definição dos termos de busca;
5. Definição da estratégia de pesquisa; e
6. Critérios de seleção;

O Apêndice A contém o resultado desse processo de revisão, enquanto o Apêndice B apresenta todos os trabalhos identificados a partir dele. Contudo, essa lista de referências do Apêndice B foi complementada por: trabalhos publicados pelo grupo de pesquisa UNISIM (no qual este trabalho está inserido), a partir de informações coletadas de Santos, Santos A e Schiozer (2020); projetos realizados pelo laboratório Sistemas de Informação e Engenharia de *Software* (SEIS) da Faculdade de Tecnologia (FT) da UNICAMP, do qual o autor e orientador desta dissertação fazem parte; e pesquisas realizadas no buscador *Google Scholar*¹ contendo as mesmas palavras-chave (ou termos) elencadas no Apêndice A. A lista final de publicações a serem discutidas na Seção 3.2 é apresentada na Tabela 3.1.

¹Google Scholar é um mecanismo de busca na *web* de livre acesso que indexa o texto completo ou metadados da literatura acadêmica por meio de uma série de formatos de publicação.

Tabela 3.1: Trabalhos relacionados.

Artigos
Easy Steps in Probabilistic Dynamic Modelling in FDP (Anupam e Tewari, 2021)
Visualização de informação aplicada à análise de resultados de processo de seleção de modelos representativos de reservatórios de petróleo (Amaral, 2021)
Using Reorderable Matrices to Compare Risk Curves of Representative Models in Oil Reservoir Development and Management Activities (Silva et al., 2020)
Developing a workflow to select representative reservoir models combining distance-based clustering and data assimilation for decision making process. (Mahjour et al., 2020)
Visualization of Ensembles of Oil Reservoir Models Based on Pixelization, Small Multiples and Reservoir Similarities (Silva, Santos e Schiozer, 2019)
Screening of Geological Uncertainty on Reservoir Dynamic Behavior with Statistical Learning Techniques (Barbiero et al., 2019)
An ROI visual-analytical approach for exploring uncertainty in reservoir models (Sahaf et al., 2019)
A visual analytics framework for exploring uncertainties in reservoir models (Sahaf et al., 2018)
A Cluster-Based Approach for Visualizing and Quantifying the Uncertainty in the Impacts of Uncertain Shale Barrier Configurations on SAGD Production (Zheng et al., 2018)
RMFinder 2.0: An Improved Interactive Multi-Criteria Scenario Reduction Methodology (Meira et al., 2017)
Use of representative models in the integration of risk analysis and production strategy definition (Schiozer et al., 2004)

3.2 Estado da arte

Entre as várias estratégias para auxiliar o processo de tomada de decisão em campos petrolíferos destaca-se o uso de técnicas da área de InfoVis propostas nos últimos anos. Meira et al. (2017) propuseram três técnicas de visualização para permitir que o usuário possa verificar a qualidade do conjunto de MRs gerado pelo RMFinder (Seção 2.3): um conjunto de gráficos de dispersão (Seção 2.1.4) para permitir a análise visual das correlações entre todos os pares de variáveis de saída; superposição, justaposição e abordagens híbridas para representar as curvas de risco (Seção 2.3.2); e gráficos de barras para discutir a semelhança entre distribuições de valores de atributos em níveis (Seção 2.3.2). Esse conjunto de visualizações define a representatividade gráfica da metodologia e *software* RMFinder. Amaral (2021) implementou uma ferramenta *web* (chamada de RMViewer) com diversas interações nessas visualizações (Seção 2.1.5), como: aplicar ou remover *zoom*, realizar *brushing*, selecionar um modelo específico, ou expandir um gráfico selecionado com detalhes dos modelos.

Para exemplificar, a Figura 3.1 apresenta os gráficos disponibilizados pela ferramenta RMViewer de Amaral (2021). A Figura 3.1a demonstra a relação entre duas variáveis de saída via gráfico de *cross-plot*, bem como os recursos interativos de *zoom-in* e *zoom-out*. Além disso, na Figura 3.1a, no canto superior direito de cada gráfico, encontra-se um botão para expandir essas visualizações para uma nova janela. A Figura 3.1b apresenta lado a lado três gráficos de curvas de risco (um para cada variável de saída), com recursos interativos similares ao gráfico de *cross-plot* apresentado na Figura 3.1a. Se um modelo for selecionado no gráfico de *cross-plot*, ou no gráfico de curva de risco, Figuras 3.1a e 3.1b, respectivamente, é realizado um *brushing*: uma linha vertical e outra horizontal com a cor vermelha são criadas nas coordenadas do modelo selecionado, destacando-o nos outros gráficos, conforme assinalado nas Figuras 3.1a e 3.1b. O *brushing* também é observado nos gráficos de diferenças de atributo-nível. As bordas das propriedades destacadas mudam para a cor verde e o fundo, de branco para azul-claro, com o objetivo de tornar o destaque mais evidente, conforme ilustrado na Figura 3.1c.

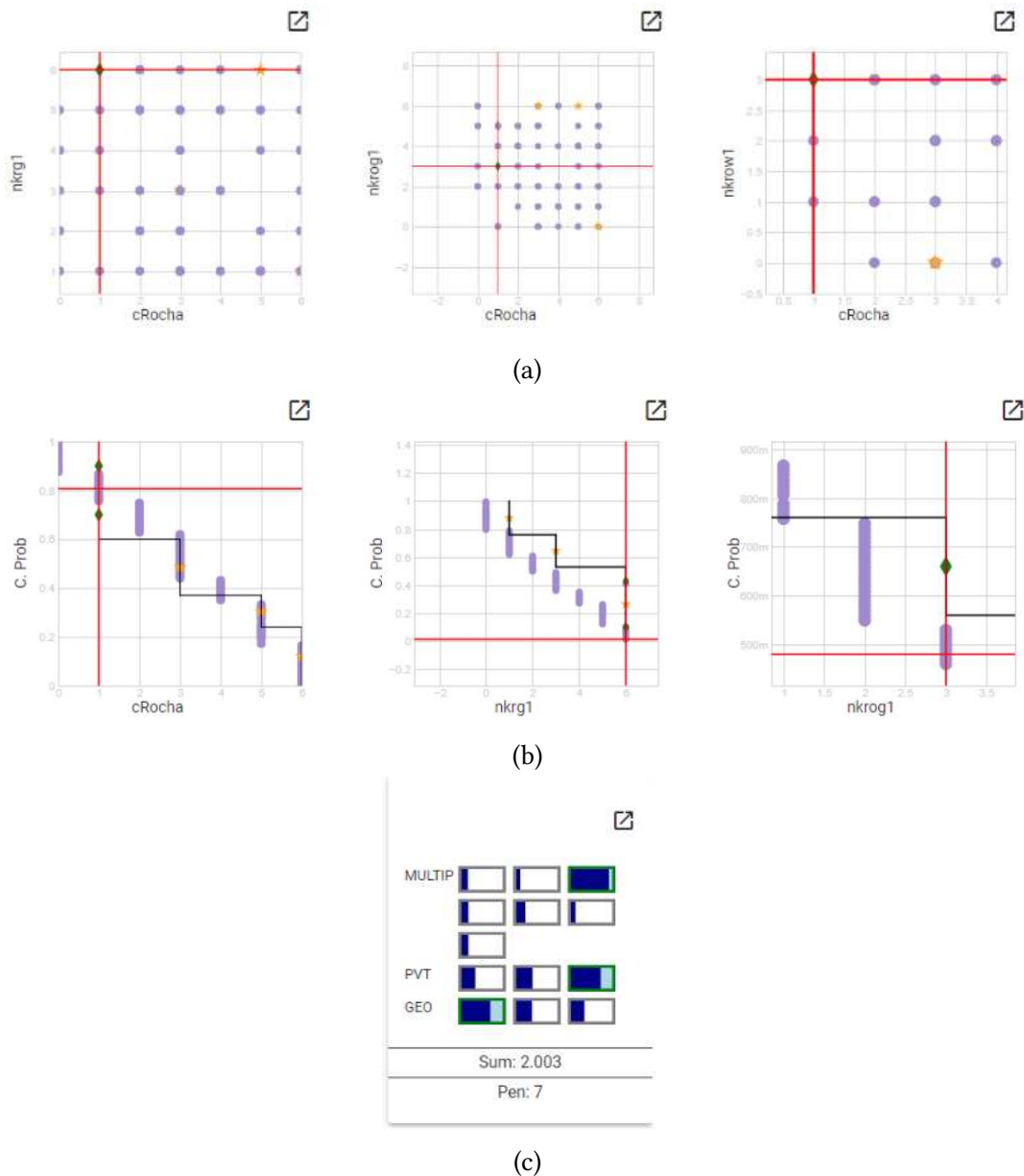


Figura 3.1: Representações visuais do RMViewer: (a) gráficos de *cross-plot*; (b) gráficos de curvas de risco; e (c) gráficos de diferença de atributo-nível.

Fonte: (AMARAL, 2021).

Por conta de o contexto do processo de tomada de decisão em campos petrolíferos geralmente envolver grandes conjuntos de dados multidimensionais (Seção 2.1.4), alguns trabalhos utilizam técnicas de redução de dimensionalidade (Seção 2.1.4) aplicadas à visualização. Anupam e Tewari (2021) propuseram duas técnicas de visualização multidimensional: *Principal Component Analysis* (PCA) e t-SNE em combinação com um agrupamento *K-means* para selecionar modelos representativos. Zheng et al. (2018) aplicam a técnica MDS (Seção 2.1.4) e análise de *clusters* para representar a influência das incertezas de

produção e selecionar MRs de grupos individuais, uma forma alternativa à seleção por intervalo de percentis de Schiozer et al. (2004).

Na Figura 3.2 é possível verificar a distribuição de 471 modelos de acordo com suas características de produção semelhantes, utilizando a técnica de projeção multidimensional MDS. De acordo com Zheng et al. (2018), a proximidade de alguns pontos na Figura 3.2 reforça a hipótese de que muitos modelos no conjunto de dados apresentam características de produção similares, podendo ser removidos em um processo de seleção de modelos mais representativos. Dessa forma, com o intuito de ajudar a selecionar esses modelos, Zheng et al. (2018) dividiram os modelos em seis grupos diferentes (uma cor para cada grupo) via algoritmo *K-means*, usando o coeficiente de silhueta² para determinar o número ideal de *clusters*. Ainda segundo os mesmos autores, esse gráfico pode ajudar a selecionar modelos ou cenários representativos de grupos individuais e melhorar o gerenciamento de reservatórios de petróleo quando há heterogeneidade entre eles.

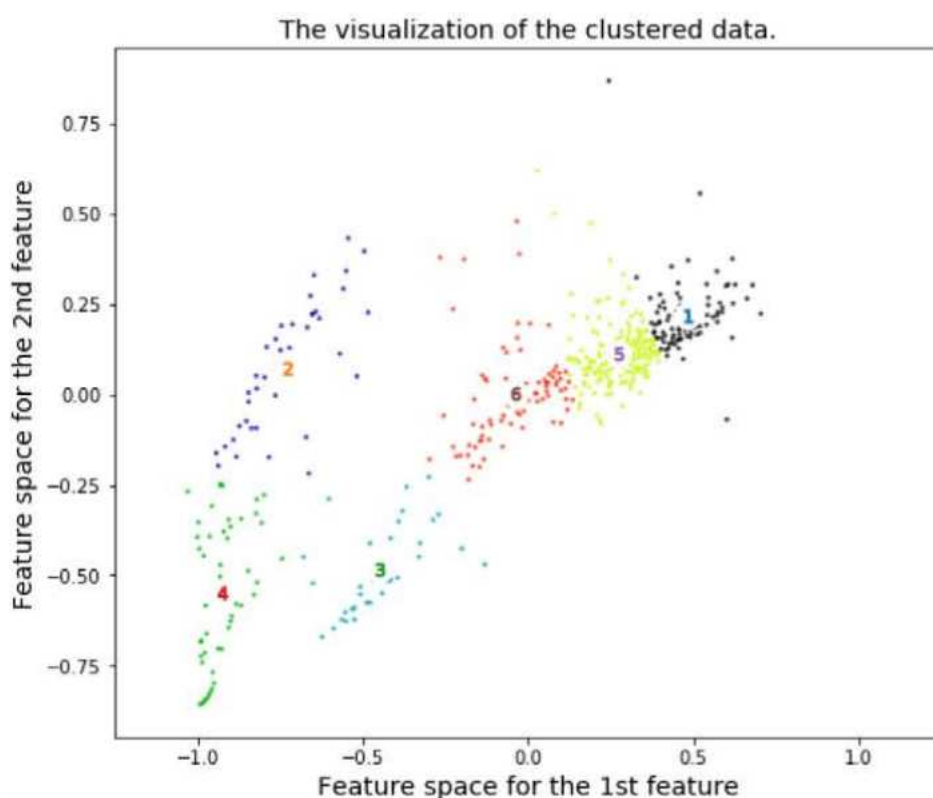


Figura 3.2: Visualização de 471 modelos de acordo com suas características de produção por meio dos algoritmos MDS e *K-means*.

Fonte: (ZHENG et al., 2018).

²O coeficiente da silhueta combina a ideia de separação entre *clusters* com a coesão interna de cada *cluster*, fornecendo uma visão abrangente da qualidade da clusterização.

Outra proposta destacada é o uso de técnicas de visualização baseadas em regiões (Seção 2.1.4) combinada com um agrupamento ou uma medida de dissimilaridade. Silva et al. (2020) propuseram representar a dissimilaridade dos pares de curvas de risco em vez das curvas propriamente ditas. Eles utilizaram matrizes reordenáveis, representadas como *heatmaps*, com objetivo de fornecer uma visão geral deste conjunto de dados que seja útil para os engenheiros tomarem suas decisões. A proposta considera como entrada um conjunto de soluções (onde cada solução é um conjunto de MRs), um conjunto de variáveis de saída, e medidas da diferença entre curvas de risco para cada variável e cada solução. A reordenação se mostrou um recurso útil para evidenciar padrões, como ilustrado na Figura 3.3.

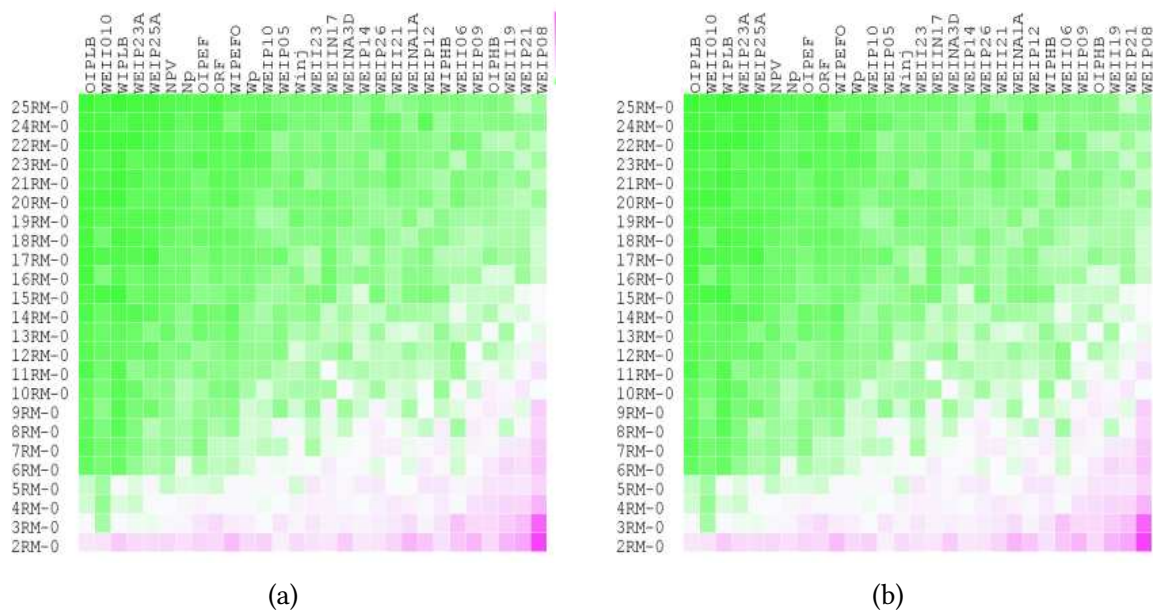


Figura 3.3: Matriz de comparação de erros de soluções calculadas para os seguintes reservatórios: (a) UNISIM-I-D (GASPAR et al., 2015) e (b) UNISIM-II-D (CORREIA et al., 2015).

Fonte: (SILVA et al., 2020)

Barbiero et al. (2019) simularam um conjunto de modelos com uma estratégia de produção pré-determinada, e as respostas dinâmicas foram agrupadas em classes de resultados por meio de algoritmos de *clustering*. Essas classes foram visualizadas em um *stacking plot* (Figura 3.4). Em função da entrada geológica, os parâmetros mais influentes são identificados via a ordenação dos eixos no gráfico, e classificados com uso de árvore de decisão (também representada no gráfico). Finalmente, com esses elementos visuais à disposição, um conjunto de MRs é selecionado a partir dos cenários geológicos.

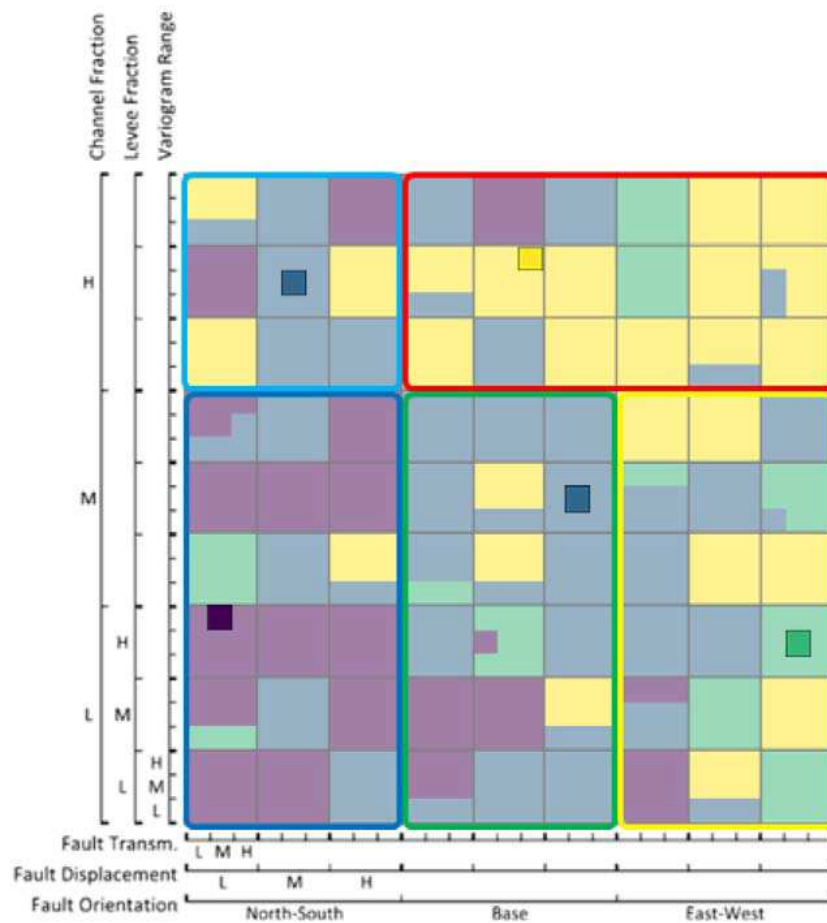


Figura 3.4: Visualização de MRs selecionados via *stacking plot*.
Fonte: (BARBIERO et al., 2019).

Silva et al. (2019) utilizaram técnicas de visualização baseadas em regiões (*heatmaps*) combinadas a técnicas orientadas a *pixel* (Seção 2.1.4) representadas de forma agrupada em uma visualização de *small multiples*. A abordagem dos *small multiples* estabelece os *heatmaps* dos modelos de reservatório 2D (com características de saturação de petróleo, permeabilidade e porosidade, Figura 2.15a-c, respectivamente) lado a lado em uma grade. A abordagem de pixelização reúne vários modelos em uma única imagem. Nela, cada célula (x, y) (“macro célula”) é subdividida de modo a conter todas as células (x, y) desses modelos (“micro células”), conforme é ilustrado na Figura 3.5. Ambas as abordagens exibem seus elementos agrupados por meio do algoritmo de *clustering* chamado de *X-means*. Esses agrupamentos podem ajudar os analistas a identificar semelhanças e MRs nesse conjunto de modelos. As matrizes de distância utilizadas no algoritmo *X-means* são baseadas na distância euclidiana de modelos para uma determinada propriedade, ou baseadas na distância euclidiana de vetores de características de modelos 2D originados dos modelos iniciais.

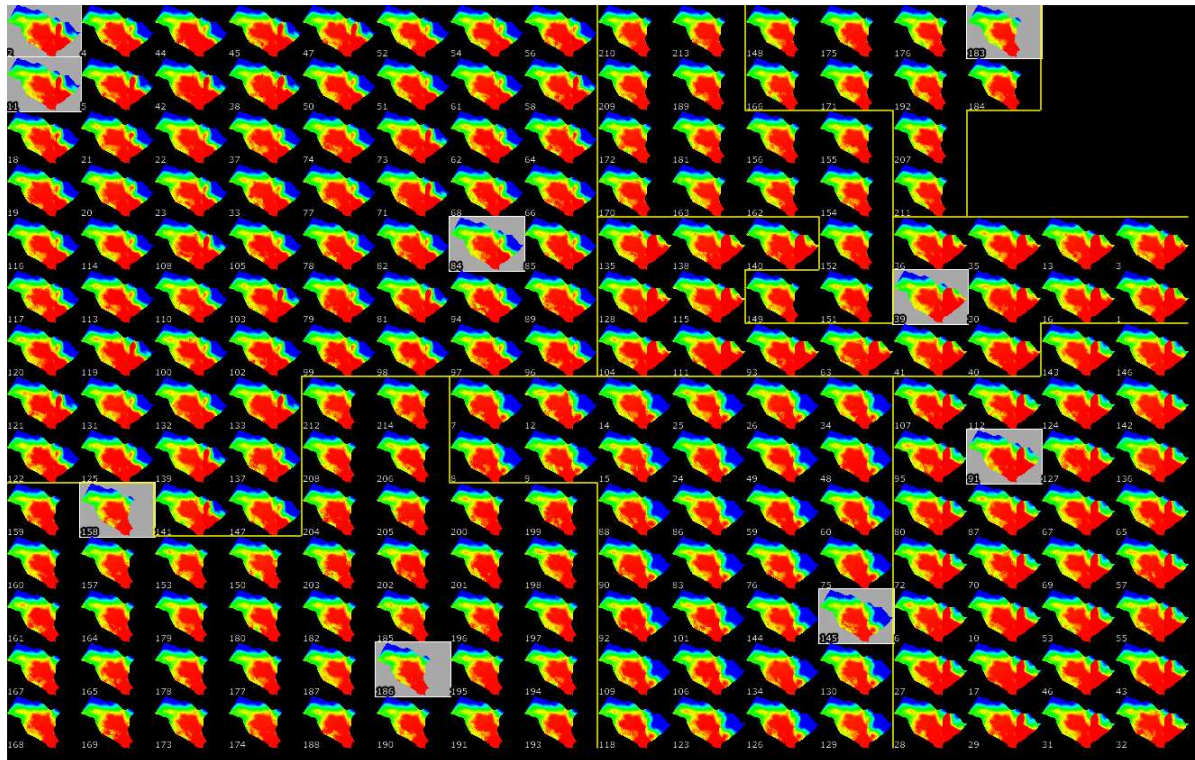


Figura 3.5: Visualização da saturação de óleo a partir de 214 modelos de reservatórios (caso de uso UNISIM-I-D de Gaspar et al. (2015)), com base na técnica de *small multiples*, com 7 conjuntos de modelos. Os MRs são destacados.

Fonte: (SILVA, 2021).

Mahjour et al. (2020) propuseram que o intervalo de incerteza dos modelos de reservatório pode ser avaliado usando as curvas de risco (CCDF) a partir do teste não paramétrico proposto por Kolmogorov-Smirnov (K-S). O teste K-S é baseado na diferença absoluta máxima entre os CCDFs observados para dois conjuntos de dados (Figura 3.6); embora Mahjour et al. (2020) não propuseram nenhuma visualização (ou agrupamento) baseada no teste K-S, ele pode ser utilizado na comparação de pares curvas de risco como complemento (ou alternativa) às medidas de distância propostas em Silva et al. (2019) e ao cálculo da Equação 2.7 de Meira et al. (2020) (ilustrada na Figura 2.17).

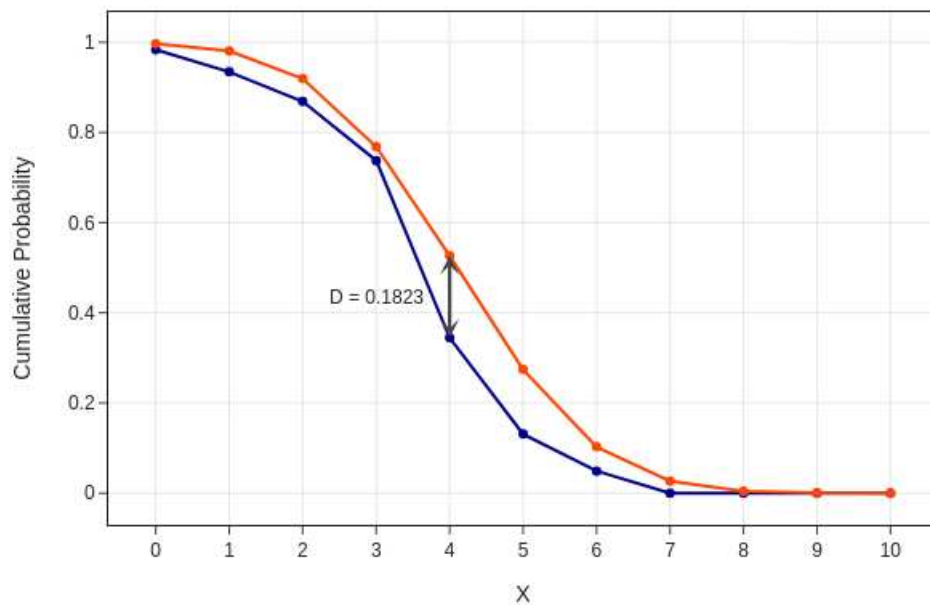


Figura 3.6: Exemplo do teste não paramétrico proposto por Kolmogorov-Smirnov, onde

$$D = \sup_x |F_1(x) - F_2(x)|.$$

Fonte: Dados originais da pesquisa.

Sahaf et al. (2018, 2019) propuseram um método probabilístico de redução (ou seleção) baseados na distância entre modelos de reservatórios de modo a agrupá-los e identificar possíveis MRs por meio de visualizações interativas, projetadas para tornar todo o processo mais compreensível. O objetivo desses trabalhos foi encontrar uma solução para o problema da complexidade do conjunto de dados, mapeando-os por meio de uma transformação de *kernel* para um novo espaço usando uma função de base radial. Isso possibilita que as amostras sejam linearmente separáveis, resolvendo o problema da complexidade das variáveis. Essa abordagem foi combinada com a aplicação de um MDS. Em seguida, foi usado o algoritmo de agrupamento *K-means*; a técnica foi chamada de *kernel K-means*. Em particular, no trabalho de Sahaf et al. (2018), essa técnica foi fundamentada em uma abordagem de agrupamento baseada na distância de pares que agrupa as realizações geológicas de acordo com suas semelhanças. Enquanto, em Sahaf et al. (2019), propuseram um método de cálculo de distância baseado em blocos, que emprega o conceito de informação mútua³. Em Sahaf et al. (2019), para capturar as incertezas de um conjunto de 100 modelos geológicos para a variável fator de recuperação de óleo, foram criados seis grupos por meio dessa técnica; os centroides, os quais são os MRs por grupo, estão destacados com uma estrela na Figura 3.7.

³Redução do nível de incerteza associado a uma variável em relação à informação fornecida por outra variável.

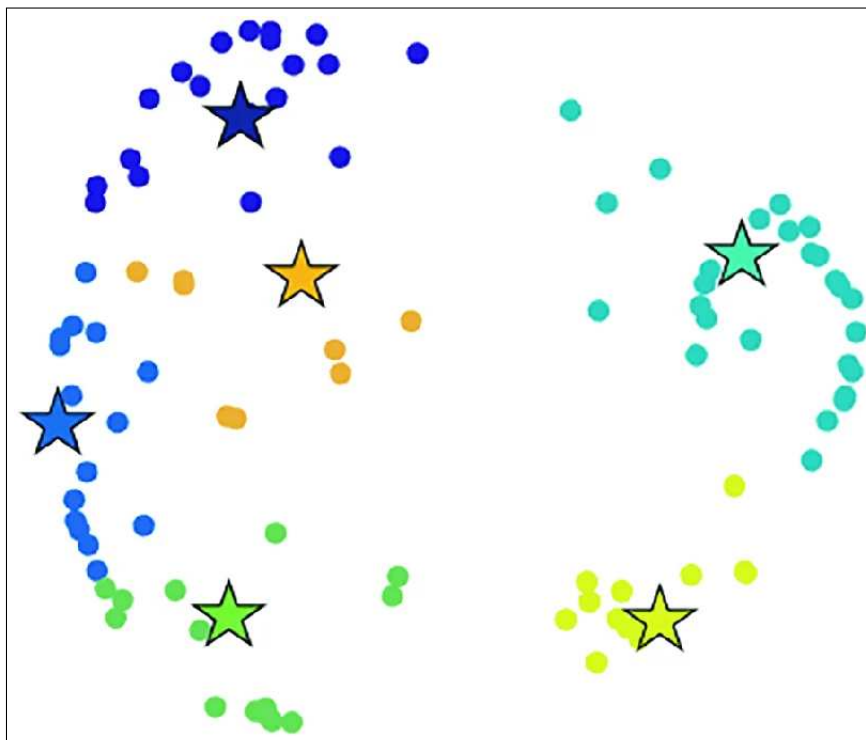


Figura 3.7: Resultado de agrupamento para 100 modelos geológicos, com seis agrupamentos para o fator de recuperação de óleo.

Fonte: (SAHAF et al., 2019).

Por fim, destaca-se uma semelhança por parte dos trabalhos elencados: aplicação de métodos de redução de dimensionalidade (ANUPAM; TEWARI, 2021; ZHENG et al., 2018; SAHAF et al., 2019, 2018) e/ou técnicas de agrupamento (ANUPAM; TEWARI, 2021; SILVA et al., 2019; ZHENG et al., 2018; SAHAF et al., 2019, 2018). A Tabela 3.2 apresenta uma comparação entre as publicações listadas na Tabela 3.1, bem como discutidas nesta seção. Esta tabela pode ajudar a compreender as diferenças entre elas, assim como possíveis lacunas a serem preenchidas neste trabalho. As colunas “*Heatmap* de Modelos – Variabilidade” e “CCDF – Sumarizada” da Tabela 3.2 não estão preenchidas, devido à ausência dessas técnicas nos trabalhos elencados. No entanto, a Tabela 7.2 apresentará a mesma comparação realizada na Tabela 3.2, incluindo a presente pesquisa, com o intuito de posicioná-la em relação à literatura atual. Nesse sentido, essas colunas “*Heatmap* de Modelos – Variabilidade” e “CCDF – Sumarizada” podem representar possíveis técnicas a serem desenvolvidas neste trabalho.

Tabela 3.2: Comparação entre os trabalhos relacionados.

Artigos	Projeção Multidimensional			Técnica de Agrupamento		Medida de Distância		Heatmap de Modelos		CCDF			Cenários com múltiplas		Mapa		Outros	
	PCA	MDS	t-SNE	K-means	X-means	Área	K-S	Representatividade	Variabilidade	Sobreposta	Justa-posta	Sumarizada	Soluções	Variáveis	2D	3D	Reordenação de Matriz	Interação
Anupam e Tewari, 2021	X		X	X														
Amaral, 2021										X	X							X
Mahjour et al., 2020							X			X								
Silva, Santos e Schiozer, 2019					X	X		X					X	X	X	X		X
Barbiero et al., 2019				X														
Sahaf et al., 2019		X		X												X		X
Sahaf et al., 2018		X		X												X		X
Zheng et al., 2018		X		X														
Meira et al., 2017										X	X							
Schiozer et al., 2004										X								

3.3 Considerações finais

O Capítulo 3 apresentou na Seção 3.1 a metodologia utilizada para realizar a revisão sistemática dos trabalhos relacionados, onde se destacou, além de um processo estruturado, outros meios de complementar a lista de publicações científicas. Em seguida, na Seção 3.2, foi apresentado um breve comentário a esses trabalhos, enfatizando as principais sinergias entre eles, a fim de que possam ser aproveitadas nos próximos capítulos deste trabalho; e uma tabela comparativa das técnicas utilizadas em cada um deles.

Capítulo 4

Metodologia

Como já mencionado na Seção 1.3, este projeto tem por objetivo geral propor e/ou aprimorar visualizações interativas de múltiplas curvas de risco (Seções 2.3.2 e 2.2.2) para ajudar os analistas de petróleo a efetuarem melhores escolhas de conjuntos de MRs de reservatórios de petróleo em cenários com múltiplas soluções e variáveis. Além disso, como objetivo específico, este trabalho procura-se focar na proposição de melhorias para as visualizações atualmente utilizadas pelo *software* RMFinder, bem como na proposição de visualizações de caráter mais geral, atualmente ausentes nesse sistema. Essas visualizações devem ser capazes de possibilitar comparações de mais alto nível de abstração, permitindo comparar cenários com diversas variáveis e soluções.

Nesse sentido, o presente Capítulo 4 aborda os métodos e materiais escolhidos para o desenvolvimento deste trabalho com intuito de alcançar esses objetivos propostos. Na Seção 4.1 consta uma descrição detalhada do conjunto de dados do RMFinder. Na Seção 4.2 são apresentadas as linguagens de programação e tecnologias para o desenvolvimento das visualizações e tratamento dos dados. Por fim, na Seção 4.3, é elencado o método de trabalho proposto para o desenvolvimento deste projeto.

4.1 Conjunto de dados

O conjunto de dados, a ser representado por visualizações interativas desenvolvidas neste trabalho, serão disponibilizados no repositório *Zenodo* com acesso público via o endereço <https://doi.org/10.5281/zenodo.6657617>. Nesse repositório, encontra-se atualmente o conjunto de dados brutos (Seção 2.1.2) gerados pelo RMFinder referentes a três

casos de estudo (Seção 2.2.4): OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017), UNISIM-I-D (GASPAR et al., 2015) e UNISIM-II-D (CORREIA et al., 2015). Esses dados estão estruturados em três grupos de arquivos, como segue:

- Dados de entrada (detalhes na Seção 4.1.1);
- Arquivo de parâmetro (detalhes na Seção 4.1.2; e
- Dados de saída (detalhes na Seção 4.1.3).

4.1.1 Dados de entrada

O conjunto de dados de entrada (exemplificado na Figura 4.1) gerado pelo RMFinder é um arquivo de valores separados por vírgula (*Comma Separated Values* – CSV) contendo os cenários $s \in T$ (ou seja, o conjunto de modelos τ simulados com uma estratégia P_r predefinida pelo analista de petróleo) para cada variável de saída f , dos quais serão escolhidos os conjuntos de MRs τ' , onde $\tau' \subseteq \tau$. Esse conjunto de dados está diretamente relacionado aos insumos para gerar as curvas de risco, elencadas na Seção 2.2.2. Esse arquivo (*rm_finder.csv*) está estruturado com os seguintes atributos:

- Número de identificação (ID) para cada modelo do conjunto (o ID deve ser único e começar em zero). Este ID será usado para identificar os conjuntos de MRs; e
- O valor das variáveis de saída f (uma coluna por variável) para cada cenário s .

ID	NPV	Np	ORF	Wp	Winj	OIP	WIP	WEIIN1	WEIIN2	WEIIN3	WEIIN4
0	1.65E+08	9817660	22.3906	2.90E+07	3.96E+07	4.38E+07	1.69E+08	-4.7E+07	-1.2E+07	-4.4E+07	-2.7E+07
1	5.32E+08	1.25E+07	23.6228	3.08E+07	4.43E+07	5.29E+07	1.65E+08	-4.2E+07	-3.1E+07	-4.3E+07	-5.6E+07
2	7.35E+08	1.38E+07	25.8977	3.62E+07	5.12E+07	5.35E+07	1.61E+08	-4.1E+07	-3.5E+07	-4.5E+07	-3.4E+07
3	80625175	8061910	17.4669	1.58E+07	2.45E+07	4.62E+07	1.57E+08	-6355816	-1.9E+07	-1.2E+07	-1.8E+07
4	2.22E+08	9269100	19.786	1.84E+07	2.83E+07	4.68E+07	1.56E+08	-3.4E+07	-2E+07	-3.8E+07	-1.3E+07
5	2.67E+08	1.02E+07	20.7491	2.37E+07	3.47E+07	4.94E+07	1.56E+08	-1E+07	-2.1E+07	-4E+07	-2.7E+07
6	2.46E+08	9838800	21.8796	2.22E+07	3.28E+07	4.50E+07	1.62E+08	-2.6E+07	-6138584	-4.5E+07	-2.2E+07
7	3.46E+08	1.08E+07	22.3118	2.56E+07	3.73E+07	4.86E+07	1.66E+08	-1.9E+07	-2.2E+07	-5.3E+07	-3.2E+07
8	5.01E+08	1.27E+07	24.6511	3.33E+07	4.70E+07	5.14E+07	1.64E+08	-1.5E+07	-1.9E+07	-3.2E+07	-4.6E+07
9	7.87E+08	1.36E+07	27.2373	2.81E+07	4.28E+07	5.00E+07	1.56E+08	-1.7E+07	-2.7E+07	-9725134	-3.6E+07
10	4.83E+08	1.18E+07	24.685	2.80E+07	4.08E+07	4.80E+07	1.59E+08	-2.3E+07	-1.9E+07	-4.8E+07	-1.9E+07
11	5.51E+08	1.21E+07	24.6961	2.74E+07	4.05E+07	4.91E+07	1.65E+08	-4.7E+07	-8058068	-1.1E+07	-5E+07
12	5.43E+08	1.21E+07	22.0048	2.70E+07	4.00E+07	5.49E+07	1.60E+08	-4.8E+07	-1.3E+07	-1.3E+07	-5.1E+07
13	4.56E+08	1.15E+07	22.3356	2.50E+07	3.73E+07	5.13E+07	1.59E+08	-3.6E+07	-2E+07	-2.5E+07	-3.1E+07
14	1.34E+08	9924990	21.1669	2.45E+07	3.52E+07	4.69E+07	1.60E+08	-1.7E+07	-7735580	-4.2E+07	-2.6E+07
15	3.1E+08	9637370	19.4241	1.68E+07	2.72E+07	4.96E+07	1.73E+08	-9058170	-7720500	-5526719	-2.2E+07
16	5.04E+08	1.21E+07	24.0272	3.24E+07	4.55E+07	5.05E+07	1.59E+08	-3.5E+07	-2.4E+07	-3.9E+07	-4.7E+07
17	2.09E+08	9443030	20.4936	1.98E+07	3.00E+07	4.61E+07	1.68E+08	-8835891	-8873856	-7776208	-3.4E+07
18	5.3E+08	1.25E+07	24.407	3.09E+07	4.44E+07	5.12E+07	1.59E+08	-4.1E+07	-2.2E+07	-5.5E+07	-1.2E+07
19	7.69E+08	1.40E+07	26.4743	3.47E+07	4.97E+07	5.28E+07	1.53E+08	-4.2E+07	-4.5E+07	-1.6E+07	-2.4E+07
20	5.78E+08	1.16E+07	24.2341	1.87E+07	3.12E+07	4.79E+07	1.67E+08	-2.6E+07	-2.2E+07	-9581725	-3.9E+07

Figura 4.1: Extrato de um exemplo do conjunto de dados de entrada. Vermelho: ID do modelo. Azul: valor das variáveis de saída por cenário.

Fonte: Dados originais da pesquisa.

4.1.2 Arquivo de parâmetro

O arquivo de parâmetro (*Parameters_<EXECUTION_ID>.csv*) é composto de um conjunto de configurações que definem a forma como o RMFinder é executado para cada execução E_i de um conjunto de execuções $E = \{E_1, \dots, E_n\}$. A Figura 4.2 apresenta um exemplo desse arquivo, no qual as palavras em negrito são os valores de cada parâmetro, definidos pelo usuário.

```

DEFINE INPUT_DATA_FILE unisim-i-d.csv
DEFINE NUMBER_RMS 9
DEFINE NUMBER_SOL_TO_BE_PRINTED 3
DEFINE USERNAME EXAMPLE
DEFINE EXECUTION_ID 1
DEFINE RANDOM TRUE
DEFINE EQUIPROBABLE FALSE
DEFINE FIXED_MODELS_BY_PROB 50
DEFINE FIXED_MODELS_BY_ID 11
DEFINE FO_WEIGHT 1,2,3,4
DEFINE NVAR 6
DEFINE INDIVIDUAL_CROSS_PLOT_WEIGHT 6,1,6,6,1,3,3,1,1,3
DEFINE INDIVIDUAL_RISC_CURVE_WEIGHT 6,6,6,1,3,3,1,1,1,1
DEFINE MANUAL_SOLUTIONS

```

Figura 4.2: Arquivo de parâmetros salvo pela *interface* gráfica do RMFinder, armazenando a configuração de execução definida pelo usuário.

Fonte: (SANTOS et al., 2021)

A descrição de cada parâmetro destacado na Figura 4.2 é apresentada na lista a seguir:

- **INPUT_DATA_FILE**: conjunto de dados com a instância a ser avaliada;
- **NUMBER_RMS**: quantidade de MRs a serem selecionados por solução;
- **NUMBER_SOL_TO_BE_PRINTED**: quantidade de melhores soluções a serem encontradas;
- **USERNAME**: nome do usuário, projeto, etc.;
- **EXECUTION_ID**: o ID da execução específica;
- **RANDOM**: se a execução é rápida (menos qualidade de solução) ou normal (com melhor qualidade);
- **EQUIPROBABLE**: se os MRs de uma solução devem ser considerados como equiprováveis;
- **FIXED_MODELS_BY_PROB**: modelos definidos pelo usuário (por probabilidades) como parte integrante de cada solução;
- **FIXED_MODELS_BY_ID**: modelos definidos pelo usuário (por IDs) como parte integrante de cada solução;
- **FO_WEIGHT**: lista de pesos de cada componente da função objetivo multicritério (Equação 2.3);
- **NVAR**: número de variáveis (indicadores globais e locais);
- **INDIVIDUAL_CROSS_PLOT_WEIGHT**: lista contendo os pesos de cada *cross-plot* (Equação 2.4);
- **INDIVIDUAL_RISK_CURVE_WEIGHT**: lista de pesos de cada curva de risco (Equação 2.6); e
- **MANUAL_SOLUTIONS**: solução definida pelo usuário para ser avaliada.

A parametrização via interface gráfica do RMFinder (Figura 4.3) tem as mesmas informações que o arquivo descrito na Figura 4.2. A única diferença é que ela tem campos específicos a serem detalhados a fim de orientar o usuário para o correto preenchimento.

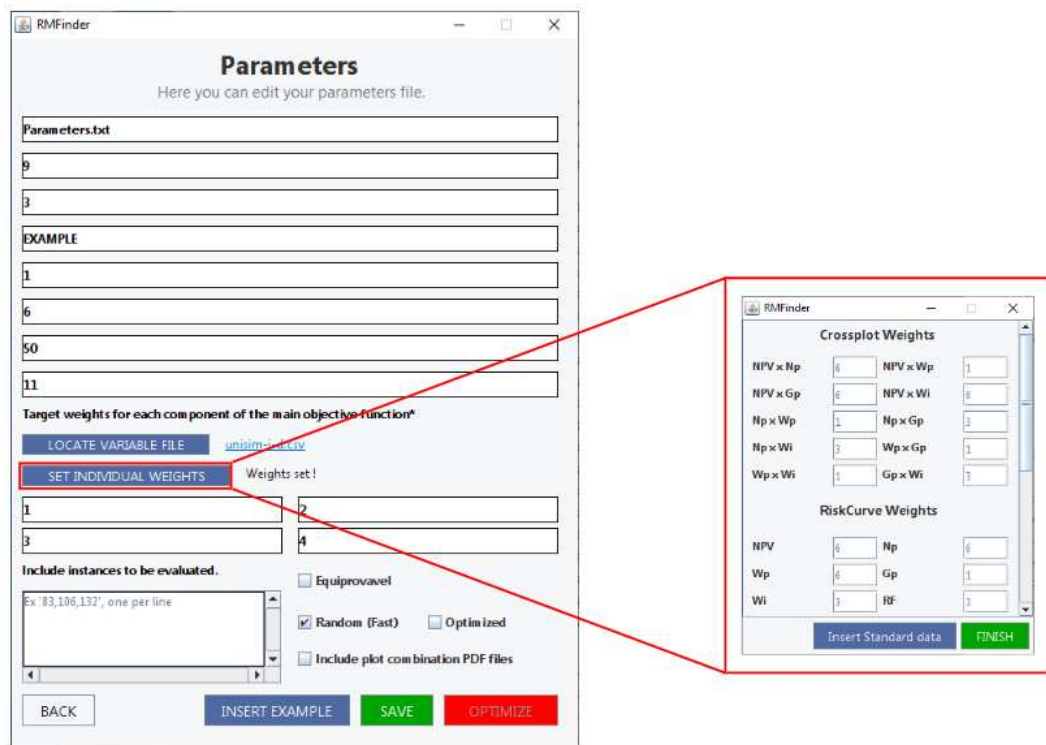


Figura 4.3: Exemplo da interface gráfico para definição de parâmetros do RMFinder.

Fonte: (SANTOS et al., 2021)

Os dados de parametrização do RMFinder serão úteis para possivelmente adicioná-los como informações interativas (p. ex. a técnica de conexão da Seção 2.1.5) nas visualizações de curvas de risco (p. ex., o usuário, ao analisar graficamente um conjunto de MRs, pode verificar quais foram os parâmetros utilizados pelo RMFinder para encontrar uma determinada solução).

4.1.3 Conjunto de dados de saída

Por fim, nos diretórios *results_<EXECUTION_ID>* estão os resultados de uma execução E_i . Em cada diretório, os arquivos *<SOLUTION_ID>_solucao.txt* contém os MRs escolhidos como sendo a solução S_i de uma execução E_i . Nesses arquivos constam a probabilidade atribuída pelo RMFinder para cada MR (Figura 4.4), e um dado arquivo de uma execução S_i representa uma curva de risco (CCDF) de um conjunto de MRs τ' . Em outras palavras, cada arquivo *<SOLUTION_ID>_solucao.txt* de uma execução E_i representa uma linha poligonal preta no gráfico com dez curvas de risco combinadas (sobrepostas) da Figura 2.19.

```
#modelID(3) with prob 0.08  
#modelID(38) with prob 0.08  
#modelID(6) with prob 0.08  
#modelID(47) with prob 0.1  
#modelID(7) with prob 0.1  
#modelID(24) with prob 0.1  
#modelID(16) with prob 0.14  
#modelID(33) with prob 0.1  
#modelID(11) with prob 0.12  
#modelID(2) with prob 0.1  
##ID  
3:38:6:47:7:24:16:33:11:2:
```

Figura 4.4: Arquivo de saída do RMFinder para um conjunto de dez MRs.

Fonte: Dados originais da pesquisa.

4.2 Ferramentas

O desenvolvimento das visualizações interativas de múltiplas curvas de risco será feito visando sua incorporação na plataforma *WebViz*. Essa restrição foi uma solicitação recebida da empresa financiadora deste projeto, que incentiva o desenvolvimento baseado em software livre e, particularmente, é uma das financiadoras também dessa plataforma. O *Webviz* é uma camada de configuração desenvolvida usando as bibliotecas *Dash* ¹ e *Plotly* ², que possibilita a criação de componentes e painéis reutilizáveis, os quais podem ser adicionados ou removidos ao criar um aplicativo usando um arquivo de configuração *YAML* ³. As visualizações e outros elementos desse aplicativo podem ser adicionados como *plugins*, altamente customizados com o uso de *Python* e *React*.

Os dados brutos do RMFinder (Seção 4.1) serão armazenados em um Sistema de Gerenciamento de Banco de Dados (SGBD) na tecnologia *PostgreSQL*⁴, com intuito de transformá-los em tabela de dados a serem utilizadas por uma estrutura visual (Seção 2.1.2). Portanto, para ser possível transformar esses dados, torna-se necessário criar um Modelo Entidade-Relacionamento (MER), assim como definir os processos de Extração, Transformação e Carga (ETL) desses dados. Para realizar a transferências desses dados entre a

¹O *Dash* é um framework *Python* para criar aplicativos *web* de aprendizado de máquina e ciência de dados.

²O *Plotly* é uma biblioteca de código aberto para criar gráficos interativos.

³*YAML* é uma linguagem de serialização de dados, frequentemente utilizada para escrever arquivos de configuração.

⁴O *PostgreSQL* é um sistema de banco de dados objeto-relacional de código aberto.

camada de *front-end web* e o SGDB, será desenvolvido um *back-end* em *Python* suportado por um *framework* Object-Relational Mappers (ORM) ⁵, nesse caso o *SQLAlchemy* ⁶.

4.3 Método de trabalho

A fim de realizar projetos de visualização com foco no usuário-alvo, é necessário realizar um estudo de modo a identificar suas características, motivações, necessidades, capacidades e preferências próprias (conforme a Seção 2.1.1). Essa etapa de caracterização dos usuários pode ajudar a determinar as tarefas que os eles executam (ou a sua jornada) e, portanto, deve ser o primeiro passo para criar uma técnica de visualização satisfatória.

Considerando que este trabalho tem sido feito para criar estratégias para obter um desenho de interface que seja centrado no usuário-alvo (conceitos estabelecidos na Seção 2.1.1), o método de desenvolvimento de visualização deve prever uma forte sinergia entre pesquisadores e usuários. A vantagem dessa interação é que os usuários podem dar *feedback* acerca das visualizações e o quanto elas ajudarão em suas tarefas. Nesse sentido, é necessário propor uma abordagem iterativa, conforme listada a seguir (KULYK et al., 2017):

1. Estudo dos usuários e suas tarefas (ou, simplesmente, a caracterização dos usuários detalhada na Seção 2.1.1);
2. Projetar um método de visualização;
3. Implementar o método de visualização (obedecendo às ferramentas estabelecidas da Seção 4.2);
4. Rever e testar o método de visualização (Seção 2.1.6). Se o resultado não for satisfatório, voltar ao passo 1 e repetir o ciclo.

Essa abordagem iterativa é proposta por Kulyk et al. (2017) como um ciclo de projeto, ilustrado na Figura 4.5. A etapa “Análise do usuário, análise de requisitos e análise de tarefas” é aplicada com uso das técnicas detalhadas na Seção 2.1.1. As etapas de “Desenho de Visualização” e “Implementação” são executadas com insumos baseados no primeiro passo e

⁵Uma técnica de programação na qual um descritor de metadados é usado para conectar objetos a um banco de dados relacional

⁶O *SQLAlchemy* é um conjunto de ferramentas *Python* e ORM, no qual flexibiliza o uso da linguagem *SQL* no desenvolvimento de aplicações.

com uso de referenciais apresentados nas Seções 2.1.2, 2.1.4 e 2.1.5. A última etapa do círculo contém tarefas para “Avaliar a Visualização”, em que são aplicados os métodos discutidos na Seção 2.1.6. A quantidade de ciclos a ser executada dependerá do andamento do cronograma do projeto.

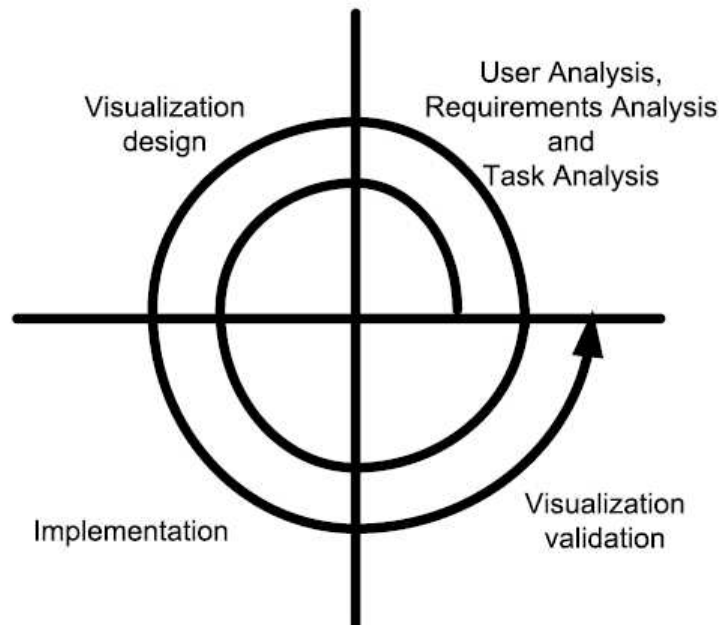


Figura 4.5: O ciclo de projeto proposto à área de InfoVis aplicada.
Fonte: (KULYK et al., 2017)

Por fim, com base na análise de usuários e modelagem de tarefas, precisa ser feito um processo de identificação da melhor técnica de visualização para uma determinada tarefa executada por um usuário e a interação necessária para eles. Na etapa “Desenho de Visualização”, dois aspectos são importantes: usuários (Seção 2.1.1) e dados (Seção 2.1.2), já que um grupo específico de usuários precisa de um tipo específico de visualização para uma determinada tarefa, com base em seu modelo mental.

Em virtude disso, o uso de protótipos⁷ será útil para absorver mudanças rápidas no projeto. Além disso, eles podem transmitir *feedback* se as ideias de projeto correspondem às ideias dos usuários sobre uma visualização específica, por exemplo.

O primeiro protótipo a ser construído, com intuito de direcionar a primeira interação com os usuários e torná-la mais objetiva, deve contemplar uma combinação de melhorias já identificadas quando se observa as visualizações da Seção 2.3.2 e os métodos discutidos na Seção 2.1; são elas: técnicas de interação da Seção 2.1.5, como filtragem, conexão, navegação,

⁷Os protótipos são produtos intermediários, não necessários, resultando em (partes de) produtos finais.

etc.; e possível agrupamento das curvas de risco (CCDF), inspirados nos trabalhos identificados no Capítulo 3, visando melhorar a percepção cognitiva dos gráficos de superposição, justaposição e abordagens híbridas (Figuras 2.19, 2.20 e 2.21, respectivamente).

4.4 Considerações finais

Nesse Capítulo 4 foram apresentados os métodos e materiais utilizados neste trabalho para atender o objetivo proposto na Seção 1.3. Foram apresentadas as características do conjunto de dados do RMFinder na Seção 4.1. Em seguida, dissertou-se na Seção 4.2 sobre os materiais (no caso ferramentas) a serem utilizadas no desenvolvimento do projeto. Por fim, a Seção 4.3 estabeleceu o método de trabalho deste projeto, dando enfoque ao ciclo iterativo de desenvolvimento.

Capítulo 5

Entrevista e protótipo estático

Como já mencionado na Seção 4.3, a primeira etapa da metodologia desse trabalho foi identificar as tarefas principais de análise de dados dos usuários por meio de entrevistas (uma das técnicas elencadas na Seções 2.1.1 e 2.1.6), observando suas principais necessidades de análise de dados e quais critérios utilizam na tomada de decisão envolvendo comparação de variáveis de saída, análise de curvas de risco envolvendo essas variáveis, MRs selecionados, e o conjunto de cenários originais.

As entrevistas conduzidas nessa etapa, por envolverem seres humanos por natureza, justificaram a elaboração de um projeto de pesquisa, submetido ao Comitê de Ética em Pesquisa (CEP) da UNICAMP para aprovação. Além disso, algumas questões foram levantadas quanto a como os participantes tomam suas decisões na seleção dos MRs para guiar as entrevistas. Também para essa etapa, desenvolveu-se protótipos estáticos (Seção 4.3) que serviram como artefatos potencializadores das discussões sobre o tema que ocorreu nessa etapa de entrevistas.

Com o objetivo de compreender mais claramente todas as etapas executadas pelos analistas de petróleo que usam a metodologia e *software* RMFinder, quais os objetivos durante o uso e quais as ações são executadas, foram construídos casos de uso a partir das análises dos resultados das entrevistas. A intenção foi estabelecer quais as ações e tarefas a que se pretende dar apoio e produzir o mapeamento da rotina de execução das atividades.

Esta etapa de entrevistas ocorreu uma única vez ao longo do ciclo iterativo de desenvolvimento proposto na metodologia desse trabalho (Seção 4.3).

Na Seção 5.1 é apresentada a preparação das entrevistas (documentos, local, data e projeto de pesquisa submetido ao CEP). Após isso, na Seção 5.2, constam os detalhes do

desenvolvimento do protótipo estático. Por fim, na Seção 5.3, são discutidas as análises dos resultados das entrevistas.

5.1 Preparação das entrevistas

Nessa etapa de entrevistas, antes de convidarmos os possíveis participantes, foi obrigatório submeter um projeto de pesquisa ao CEP/UNICAMP para aprovação. Após autorização do CEP¹, os participantes puderam ser convidados a participar dessa etapa do estudo.

Neste contexto, foi possível contar com o apoio do grupo de pesquisa UNISIM, ao qual o autor desta dissertação e seu orientador estão vinculados. Como parte desse apoio, solicitou-se ao ponto focal do UNISIM a sugestão de possíveis participantes do grupo que pudessem voluntariamente participar dessa primeira etapa de entrevistas. Os pré-requisitos desses voluntários deveriam ser:

- (i) serem engenheiros ou estudantes, em um amplo espectro de idades (aproximadamente entre 22 e 65 anos);
- (ii) terem experiência em engenharia de reservatórios de petróleo (de 2 a 20 anos);
- (iii) terem tido contato anterior com a metodologia e *software* RMFinder; e
- (iv) terem preferencialmente enfrentado problemas de seleção de MRs em cenários com diversos modelos, soluções e variáveis.

Em seguida, foram selecionados quatro voluntários com os pré-requisitos estabelecidos, entre eles o ponto focal do UNISIM. Aproximadamente, esses voluntários têm 37 anos e, em média, 12 anos de experiência como pesquisadores de engenharia do petróleo. Dois deles informaram que não estão atualmente trabalhando com o RMFinder, mas devem usá-lo em pesquisas futuras; dois deles utilizam a ferramenta desde a versão inicial; e dois a usam há pouco mais de três anos. Dois possuem doutorado e dois mestrado.

Sendo assim, as entrevistas foram agendadas para os dias 03 e 04 de julho de 2023, respeitando a disponibilidade de cada um dos participantes (consultados previamente sobre os dias e horários propostos). Cada entrevista foi planejada para ser realizada com duração de 2 horas a 2 horas e 30 minutos, sendo uma sessão agendada para o período da manhã, das

¹A aprovação deste projeto foi recebida em 12 de Junho de 2023 e foi protocolada pelo Certificado de Apresentação de Apreciação Ética (CAAE) número **68642523.9.0000.540**.

09:00 às 11:30, e outra sessão para o período da tarde, das 14:00 às 16:30. Essas seções foram divididas para ocorrer duas por dia, totalizando quatro entrevistas individuais, sendo uma para cada voluntário. Essas entrevistas ocorreram no formato presencial dentro das instalações do CEPETRO/UNICAMP².

Um Termo de Consentimento Livre e Esclarecido (TCLE) referente à colaboração dos participantes nessa etapa foi elaborado (Apêndice C), o qual foi assinado pelos participantes antes do início de cada entrevista, autorizando o registro de gravação de áudio, para uso restrito nesta pesquisa. Essa etapa de entrevistas foi composta por tópicos relacionados às atividades dos participantes, sendo, portanto, uma entrevista semiestruturada (Apêndice D), utilizando uma abordagem de avaliação qualitativa (Seção 2.1.6). Nessa etapa apresentou-se as visualizações preliminares aos participantes, que refletiam as primeiras ideias dos pesquisadores desse projeto para auxiliá-los em suas tarefas (ou seja, a demonstração do protótipo estático a ser detalhado na Seção 5.2). A intenção é que essas visualizações ajudassem a confirmar as hipóteses prévias do grupo de pesquisa sobre as tarefas de análise a serem apoiadas por visualizações, incentivassem os participantes a observarem o potencial do uso de gráficos interativos na tentativa de apoio às suas tarefas, e levantassem ideias complementares dos participantes nesse tema. Além disso, buscou-se seguir as orientações de entrevistas de Carpendale (2008) (Seção 2.1.6), tanto na elaboração das perguntas da entrevista semiestruturada (Apêndice D), quanto na sua condução (interlocução entre entrevistador e participante).

Por fim, essa etapa de entrevista, bem como seus preparativos, foi realizada em conjunto com outro pesquisador do SEIS, o qual tem trabalhado na visualização de variáveis de produção de campos e poços como auxílio na seleção de modelos representativos de reservatórios de petróleo (p. ex., os gráficos ilustrados na Figura 2.22) sob a liderança do mesmo orientador deste trabalho.

5.2 Preparação dos protótipos estáticos

A metodologia deste trabalho foi inspirada no ciclo de desenvolvimento iterativo apresentado por Kulyk et al. (2017), conforme detalhado na Seção 4.3. No entanto, acrescentou-se uma etapa preliminar, na qual formulou-se hipóteses (com o apoio da experiência do grupo de pesquisa

²CEPETRO – Centro de Estudos em Energia e Petróleo – é um centro de pesquisa da UNICAMP, situado na Rua Cora Coralina, 350, Cidade Universitária “Zeferino Vaz”, Distrito de Barão Geraldo, Campinas–SP CEP:13083–896

SEIS) das características dos usuários do RMFinder, e das principais tarefas que eles desejam realizar para selecionar melhor os MRs ou comparar conjuntos de MRs.

Neste contexto, desenvolveu-se um protótipo estático de visualizações com o objetivo de fornecer percepções a esses usuários nessa primeira interação (ou entrevistas), supondo que suas principais tarefas poderiam ser descritas sucintamente como:

- (i) análise da representatividade de uma solução; e
- (ii) comparação de um conjunto de soluções para encontrar a mais representativa.

Entender a chamada *representatividade* é um desafio porque ela tem componentes explícitos e tácitos. Os componentes explícitos são aqueles já definidos nas funções de otimização do RMFinder (Seção 2.3.1). Os componentes tácitos devem ser obtidos provocando reflexões dos usuários – uma tarefa realizada ao discutir o protótipo com eles na etapa de entrevistas.

Esse trabalho de prototipagem de visualizações como suporte para a seleção de modelos representativos de reservatórios de petróleo foi apresentado na 17th *International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing* (CGVCVIP) realizada em Porto, Portugal em Junho de 2023 (PÓVOA; DA SILVA; SILVA, 2023). Esse protótipo contemplou quatro representações visuais estáticas; são elas: projeção multidimensional de curvas de risco, *heatmap* reordenável de variáveis por soluções, *heatmap* reordenável de modelos utilizados por solução e curvas de risco sumarizadas por envoltório, apresentadas nas Seções 5.2.1, 5.2.2, 5.2.3 e 5.2.4, respectivamente.

5.2.1 Projeção multidimensional de curvas de risco

A Figura 5.1 utiliza uma projeção multidimensional das curvas de risco das soluções no conjunto $(\tau, S_1, \dots, S_n)$, referente à medida de distância “área entre curvas de risco” (Seção 2.3.1 e Figura 2.17), considerando simultaneamente um conjunto de variáveis de saída f . Com o apoio de um agrupamento, assinalado por sete cores diferentes na Figura 5.1, essa representação visual visa possibilitar ao usuário identificar a similaridade entre pares de soluções (S_i, S_j) , e também a similaridade entre cada solução e o conjunto de todos os modelos τ . A similaridade de uma solução S_i com relação a τ é, na verdade, a representatividade de S_i . Na Figura 5.1, cada ponto representa uma solução, e o ponto com um tamanho maior identifica o conjunto de todos os modelos (τ).

Nessa representação visual, usou-se o algoritmo t-SNE (MAATEN; HINTON, 2008) para calcular a projeção multidimensional, reduzindo a dimensionalidade de um conjunto de dados, referente às execuções do RMFinder do caso de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017), de 251 elementos (sendo o conjunto de todos os modelos e mais 250 soluções), considerando apenas a variável de saída Np , para duas dimensões. Executou-se um agrupamento, após a projeção, que definiu sete grupos, com o uso do algoritmo *X-Means* (PELLEG; MOORE, 2000).

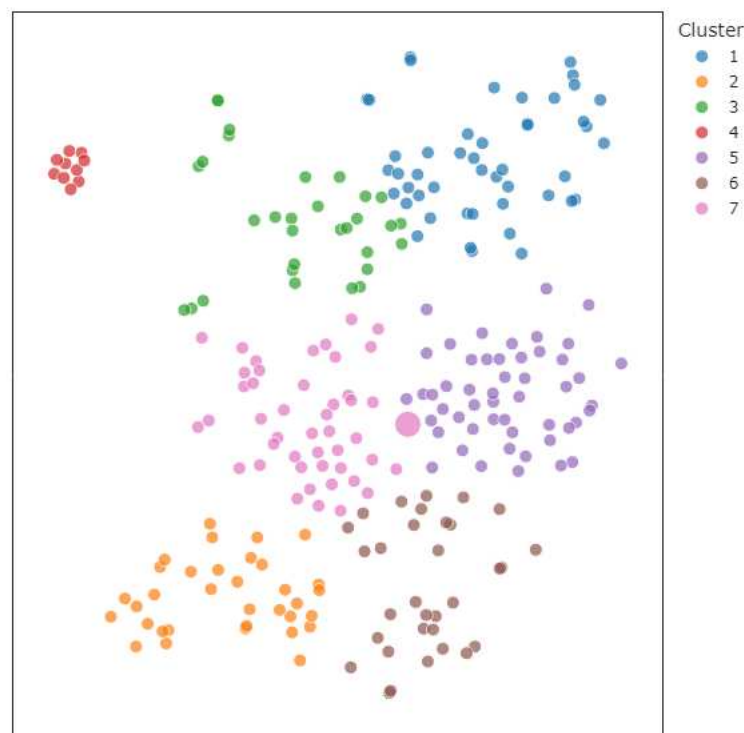


Figura 5.1: Projeção multidimensional das curvas de risco da variável de saída Np , usando dados do caso de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017).

Fonte: Dados originais da pesquisa.

5.2.2 *Heatmap* reordenável de variáveis por soluções

Na Figura 5.2 é ilustrado um *heatmap* de soluções por variáveis, como o de Silva et al. (2020), cujas células indicam a “representatividade baseada em curvas de risco” de uma variável de saída f para uma solução S_i . Essa representatividade é definida a partir de uma medida de similaridade entre a curva de risco do conjunto de todos os modelos τ e a curva de risco de uma solução S_i (ou seja, quanto mais próximas essas curvas, melhor a representatividade). Conforme essa medida, a escala de cores do *heatmap* segue um padrão divergente que se inicia

em verde (ausência de erro), passando pelo branco (valor neutro) e seguindo até roxo (erro máximo no conjunto de dados). O valor neutro usado é arbitrário e neste trabalho está fixado conforme utilizado por Silva et al. (2020).

Na Figura 5.2, a medida de similaridade utilizada foi a “área entre curvas de risco” (Seção 2.3.1 e Figura 2.17). Além disso, para ajudar a revelar possíveis informações ocultas, utilizou-se o algoritmo *Optimal Leaf Ordering*, uma das técnicas citadas em Bar-Joseph, Gifford e Jaakkola (2001) e Baroni e Silva (2022), para reordenar as linhas e colunas desse *heatmap* (Seção 2.1.4). O intuito dessa representação visual é apoiar o usuário em duas tarefas; são elas:

- (i) permitir ao usuário analisar a representatividade de uma dada solução com relação a várias variáveis de saída; e
- (ii) auxiliar o usuário a comparar soluções com números distintos de MRs e observar quais variáveis de saída provocam mais ou menos erro (ou seja, quais variáveis de saída precisam de mais MRs para serem representadas corretamente).

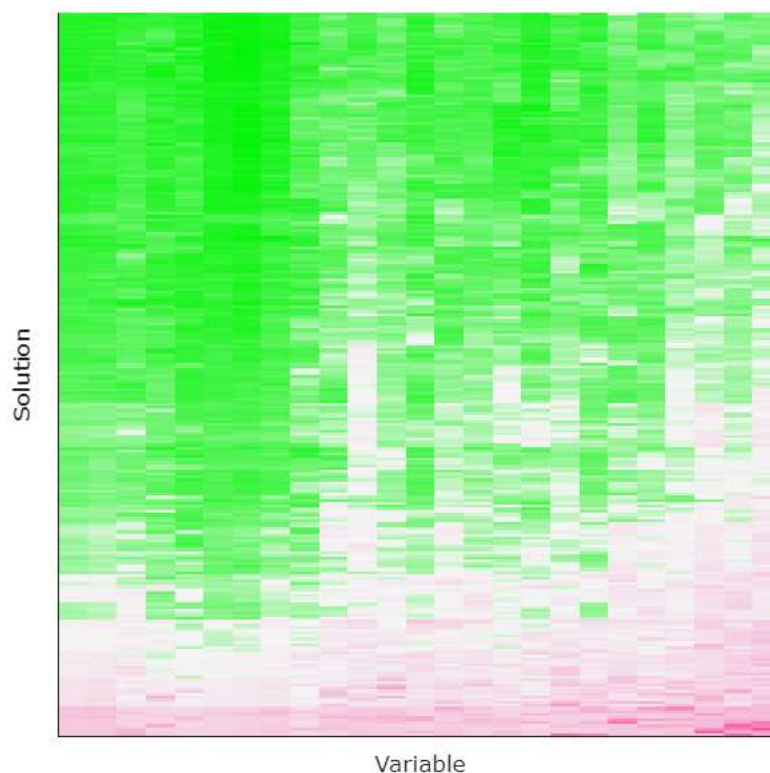


Figura 5.2: *Heatmap* reordenável de soluções por variáveis, usando dados do caso de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017).

Fonte: Dados originais da pesquisa.

5.2.3 *Heatmap* reordenável de modelos utilizados por solução

A Figura 5.3 demonstra um *heatmap* binário dos modelos usados por cada solução para que o usuário possa encontrar padrões relevantes em conjuntos de soluções com relação à sua variabilidade, p.ex. subconjuntos comuns de MRs. Embora a variabilidade não seja uma tarefa intrinsecamente relacionada às curvas de risco, julgou-se relevante essa proposta de representação visual para o usuário observar como o RMFinder selecionou os MRs com diferentes parâmetros de execuções, gerando inúmeras soluções de um mesmo conjunto de dados. Na Figura 5.3, uma célula na cor preta representa o modelo utilizado na solução, enquanto a cor cinza em uma célula sinaliza a ausência do modelo na solução.

O heatmap é reordenado pelo algoritmo *Optimal Leaf Ordering* (BAR-JOSEPH; GIFFORD; JAAKKOLA, 2001), visando agrupar soluções similares (ou seja, que usem aproximadamente os mesmos modelos) e agrupar modelos que foram usados nas mesmas soluções. Com isso, blocos densos de células pretas indicam que várias soluções utilizaram os mesmos modelos; além disso, colunas com escassez de células pretas indicam que a maioria das soluções não utilizou os modelos referentes a essas colunas.

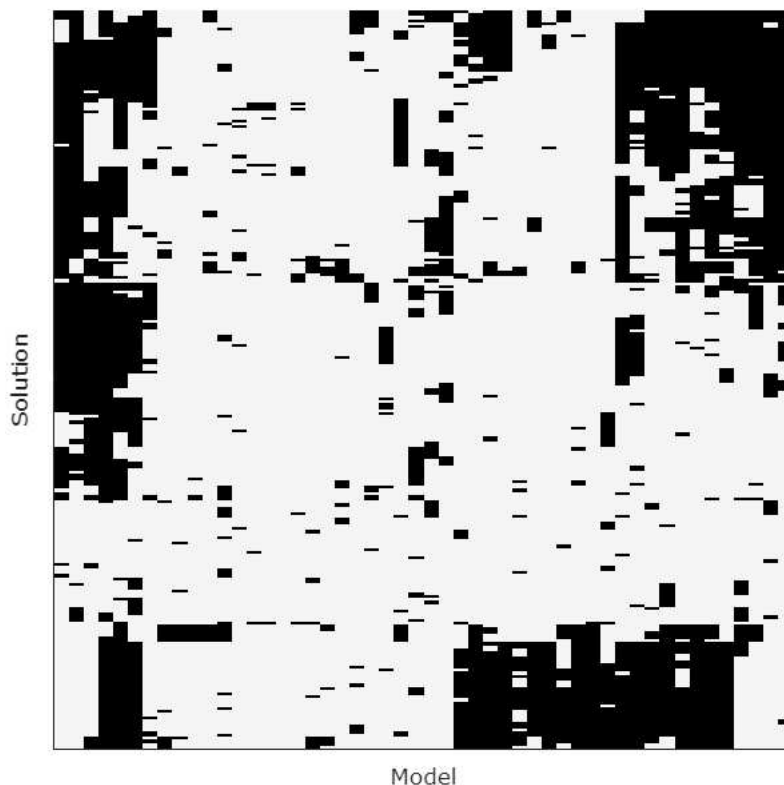


Figura 5.3: *Heatmap* reordenável de soluções por modelos, usando dados do caso de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017).

Fonte: Dados originais da pesquisa.

5.2.4 Curvas de risco sumarizadas por envoltório

Na Figura 5.4, observa-se uma sequência de gráficos de curvas de risco (um para cada variável de saída f) do conjunto de todos os modelos τ com envoltórios (faixas de confiança) que resumem os conjuntos de curvas de risco de várias soluções ($S_1, \dots, S_n$), para ajudar os usuários na análise da representatividade das soluções baseada em sua curva de risco e os seus respectivos detalhes. Os pontos roxos representam o conjunto de todos os modelos, e o envoltório cinza (similar a uma “sombra”) sumariza um conjunto de curvas de risco (ou de soluções), sendo uma alternativa aos gráficos de curvas de risco justapostos e sobrepostas (Figura 2.21). Assim, nesse gráfico, quanto maior o envoltório cinza, pior é ao menos uma das curvas de risco sumarizadas por ele.

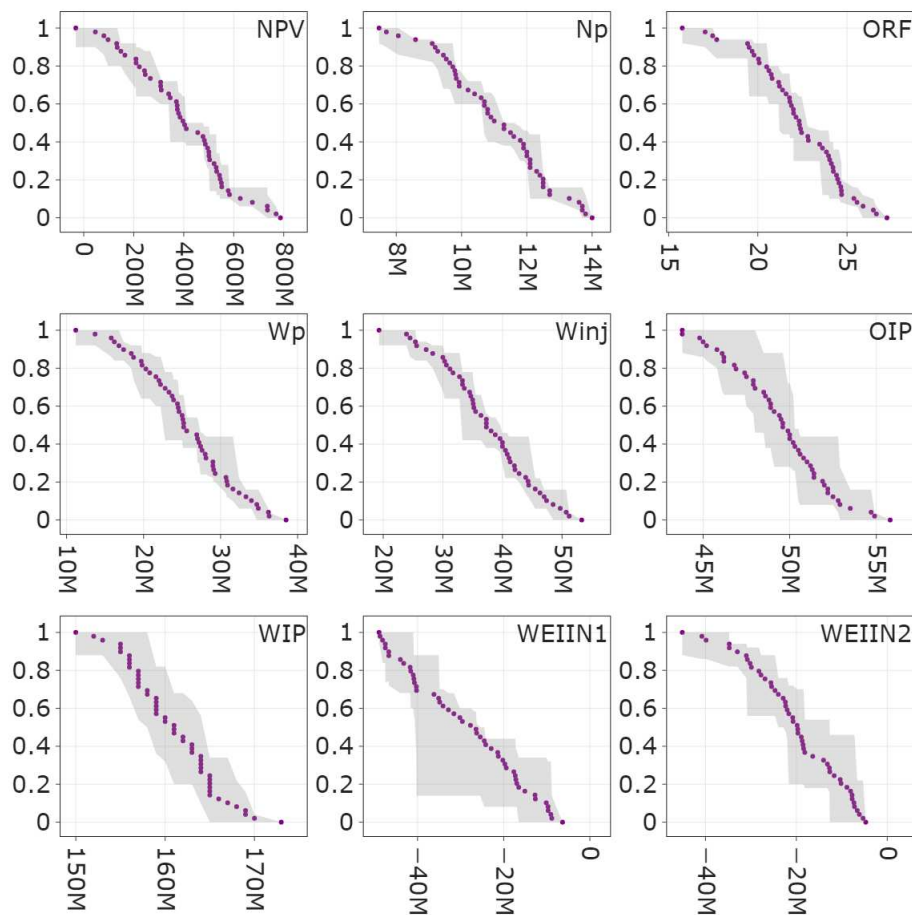


Figura 5.4: Curvas de risco com envoltório, usando dados do caso de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017).

Fonte: Dados originais da pesquisa.

5.3 Análise dos resultados das entrevistas

A análise dos resultados das entrevistas foi dividida em duas etapas: interpretar e sumarizar as questões presentes no Apêndice D, totalizando dez perguntas para guiar a entrevista semiestruturada; e avaliar o protótipo estático, composto por quatro representações visuais (Seção 5.2). Essas etapas estão descritas detalhadamente nas Seções 5.3.1 e 5.3.2, respectivamente.

5.3.1 Avaliação da entrevista semiestruturada

Para tornar os resultados da entrevista semiestruturada mais claros, após a análise do material das entrevistas (gravação e anotação), as dez perguntas do Apêndice D foram divididas em quatro categorias; são elas:

- Como lidar com as diversas variáveis de saída;
- Como identificar bons MRs;
- Como identificar soluções adequadas para uma execução; e
- Pontos a serem aprimorados no relatório gráfico apresentado ao final de cada execução do RMFinder (Seção 2.3.2).

Em relação à categoria “Como lidar com diversas variáveis de saída”, foi mencionado que é preciso estabelecer um protocolo de uso para permitir uma análise adequada e padronizada do relatório gráfico do RMFinder (Seção 2.3.2). A seleção de pesos das funções objetivo do RMFinder para aumentar ou diminuir a importância de uma boa distribuição dos MRs (conforme descrito nas Seções 2.3.1 e 4.1.2) é um processo de tentativa e erro como parte de uma abordagem de otimização iterativa, com o objetivo de atingir os resultados desejados. Esses pesos são definidos sem uma regra (ou heurística) estabelecida. Nesse sentido, os participantes selecionam as principais variáveis de saída de acordo com a relevância da sua pesquisa (p. ex., NPV , ORF , Np e Wp) para serem analisadas nos relatórios gráficos do RMFinder.

Na categoria “Como identificar bons MRs”, os participantes relataram que, ao analisar a distribuição dos MRs na nuvem de pontos dos *crossplots* (Seção 2.3.2), em seus quadrantes, verificam a ausência de discrepâncias. Além disso, procuram por soluções com curva de risco

próxima à curva de risco de todos os modelos (ou seja, soluções adequadas de acordo com a representatividade da curva de risco) para determinar MRs nos percentis *P10*, *P50* e *P90* (ou cenários otimista, mais provável e pessimista, respectivamente). Sendo assim, quando um “bom” MR é encontrado, ele é fixado na configuração dos parâmetros para uma próxima execução do RMFinder (Seção 4.1.2), enfatizando a necessidade de executar o RMFinder mais vezes para tentar cobrir uma maior variedade de cenários e comparar os resultados dessas diversas execuções, de acordo com os participantes.

Os participantes mencionaram, na categoria “Como identificar soluções adequadas para uma execução”, que executam o RMFinder até encontrarem uma solução adequada em relação à representatividade, baseada nos *cross-plots* e nas curvas de risco, podendo ser mais de dez execuções. O número de soluções solicitadas para cada execução pode variar de cinco a dez. A quantidade de MRs pode variar de acordo com os requisitos de cada pesquisador, mas selecionar nove MRs é o procedimento mais usual, uma vez que, em geral, os atributos de entrada do problema de otimização do RMFinder são divididos (ou discretizados) em três níveis (conceito de atributo-nível, mencionado na Seção 2.2) e cada atributo seria representado por três MRs, totalizando nove modelos.

Por fim, na categoria “Pontos a serem aprimorados no relatório gráfico apresentado ao final de cada execução do RMFinder”, os participantes relataram que o relatório gráfico do RMFinder não apresenta nenhum recurso interativo, dificultando a visualização de um determinado MR em uma curva de risco nas outras curvas de risco (uma para cada variável de saída); ou a possibilidade de aplicar filtros por execução, solução e variável de saída, por exemplo. Além disso, os participantes mencionaram que, para comparar resultados de execuções distintas, é necessário abrir os relatórios gráficos do RMFinder em janelas diferentes e alterná-los (via atalho do teclado ALT + TAB, p. ex.), ou imprimi-los para fins de comparação.

5.3.2 Avaliação do protótipo estático

As visualizações presentes no protótipo estático (Seção 5.2) foram apresentadas aos participantes de acordo com o *visual information seeking mantra* (SHNEIDERMAN, 1996), começando por uma visão geral, seguindo para possibilidades de uso de *zoom* e filtros, e então apresentando detalhes sob demanda, de forma a simular a navegação entre eles.

Nesse sentido, foram apresentadas as seguintes representações visuais de visão geral: projeção multidimensional de curvas de risco, *heatmap* reordenável de variáveis por soluções

e *heatmap* reordenável de modelos utilizados por solução, contextualizando as possibilidades de recursos interativos (descritas na Seção 2.1.5) a serem implementados; e, por fim, as curvas de risco sumarizadas por envoltório e as curvas de risco justapostas e sobrepostas, detalhadas por meio da seleção das marcas dos gráficos de visão geral. Sendo assim, as percepções dos entrevistados para cada representação visual do protótipo estático são descritas a seguir:

- (i) **Projeção multidimensional de curvas de risco:** primeiramente, uma vez que se trata de uma proposta completamente nova, foi necessário explicar o conceito de projeção multidimensional como sendo uma técnica para simplificar o conjunto de dados, na qual é feita uma transformação para reduzir as dimensões originais (geralmente para apenas duas), preservando a semelhança entre os elementos. Para exemplificar esse conceito foi citada a técnica PCA (Seção 3.2), amplamente difundida em diversas áreas de pesquisa. Em seguida, apresentou-se a Figura 5.1, permitindo ao usuário identificar a similaridade entre os pares de soluções, bem como a semelhança entre cada solução e o conjunto de todos os modelos, usando uma medida de similaridade (a ser selecionada pelo usuário), como: a “área entre curvas de risco”, ou o teste não-paramétrico proposto por Kolmogorov-Smirnov (exemplificado aos participantes por meio da Figura 3.6). As opiniões dos participantes foram “legal” e “muito bom”; no entanto, sugeriram ter um texto em destaque ao posicionar o mouse na solução (ou marca) da representação visual, contendo informações sobre a configuração dos parâmetros de execução do RMFinder (Seção 4.1.2);
- (ii) **Heatmap reordenável de variáveis por soluções:** dois participantes conheciam o *heatmap* ilustrado na Figura 5.3, por conta do trabalho de Silva et al. (2020), e afirmaram ser “interessante” ter essa representação visual integrada às outras do protótipo (descritas na Seção 5.2). Os outros dois participantes disseram ser “muito interessante” essa representação visual para analisar a representatividade de uma solução em relação a várias variáveis de saída e demonstraram especial interesse em experimentar o recurso interativo de reordenação, alertando para a importância de acrescentar à Figura 5.3 os rótulos dos eixos *X* e *Y*, bem como um texto em destaque ao posicionar o mouse na célula do *heatmap* com as informações da execução, solução e variável de saída. Também foi mencionado aos usuários que a “representatividade baseada em curvas de risco” de uma variável de saída para uma solução (i.e., cada célula do *heatmap*) seria calculada a partir das medidas de similaridade: “área entre

curvas de risco” (utilizada no *heatmap* de Silva et al. (2020)) e teste não-paramétrico proposto por Kolmogorov-Smirnov, comparando a curva de risco do conjunto de todos os modelos e a curva de risco de uma solução. Assim como no item (i), o usuário selecionaria a medida de similaridade a ser usada na célula do *heatmap*;

- (iii) **Heatmap reordenável de modelos utilizados por solução:** os participantes mencionaram ser “legal” o mapeamento visual da Figura 5.3 para encontrar subconjuntos comuns de MRs, ajudando a identificar a variabilidade em conjuntos de soluções, por exemplo. Tal qual no item (ii), reafirmaram o interesse pelo recurso de reordenação do *heatmap* e sugeriram adicionar os rótulos dos eixos X e Y, bem como um texto em destaque ao posicionar o mouse na célula do *heatmap* com as informações da execução, solução e modelo. Somado a isso, destacaram a necessidade de assinalar com uma cor diferente as células referentes aos modelos fixados na configuração dos parâmetros de execução do RMFinder (Seção 4.1.2);
- (iv) **Curvas de risco sumarizadas por envoltório:** os participantes comentaram ser uma alternativa “muito interessante” o gráfico apresentado na Figura 5.4 em relação ao gráfico de curvas de risco justapostas e sobrepostas (Figura 2.21) do relatório gráfico do RMFinder, evitando, de acordo com eles, uma “poluição visual”. Segundo um dos participantes, seria “relevante” acrescentar linhas na representação visual para identificar a média e o desvio padrão das curvas de risco (ou soluções) sumarizadas no envoltório; e
- (v) **Curvas de risco justapostas e sobrepostas:** como mencionado no item (iii), os participantes enfatizaram a importância de assinalar com uma cor diferente os modelos fixados na configuração dos parâmetros de execução do RMFinder (Seção 4.1.2) no gráfico de curvas de risco justapostas e sobrepostas (Figura 2.21). Além disso, sugeriu-se a inclusão de recursos interativos à representação visual, tais como: filtros, *zoom* e selecionar um MR em uma curva de risco e destacá-lo nas outras curvas de risco.

5.4 Considerações finais

O Capítulo 5 apresentou o estudo dos usuários e suas tarefas (ou seja, a caracterização de usuários, elencada na Seção 2.1.1) realizada por meio de entrevistas e apoiado por um protótipo estático. Na Seção 5.1 destacaram-se os preparativos para essas entrevistas iniciais, tais como: documentos, local, data e projeto de pesquisa submetido ao CEP. Em seguida, na Seção 5.2, a elaboração do protótipo estático contendo quatro representações visuais. A última seção Seção 5.3.2 abordou a análise dos resultados dessas entrevistas, dando enfoque à entrevista semiestruturada, e às percepções e ideias do usuário em relação ao protótipo estático.

Capítulo 6

Desenvolvimento do sistema de visualização

Após a análise dos resultados das entrevistas (Seção 5.3), o sistema de visualização interativo foi desenvolvido visando auxiliar o usuário nas tarefas de análise da representatividade de uma solução e de comparação de um conjunto de soluções para encontrar a mais representativa, adicionando as percepções e ideias coletadas na etapa de entrevista (Capítulo 5).

Previamente, para realizar o desenvolvimento das visualizações interativas, foi necessário transformar os dados brutos do RMFinder (Seções 4.1.1, 4.1.2 e 4.1.3) em tabelas de dados (Seção 2.1.2). Esses dados transformados foram armazenados em um SGBD, respeitando o MER detalhado no Apêndice G. A transformação dos dados foi desenvolvida com uso da linguagem *Python* e as bibliotecas *Numpy*¹ e *Pandas*², extraíndo as informações dos conjuntos de dados do RMFinder; transformando-as nas estruturas de dados (ou classes) do MER, com o uso do ORM *SQLAlchemy*; e carregando-as no SGDB na tecnologia *PostgreSQL* (Seção 4.2). Após o tratamento e armazenamento desses dados, foi desenvolvido um sistema de visualização interativo (Seção 2.1.5), conforme ilustrado na Figura 6.1,

¹O *Numpy* é uma biblioteca para computação científica com *Python*.

²O *Pandas* é uma ferramenta de análise e manipulação de dados de código aberto desenvolvida com base na linguagem de programação *Python*.

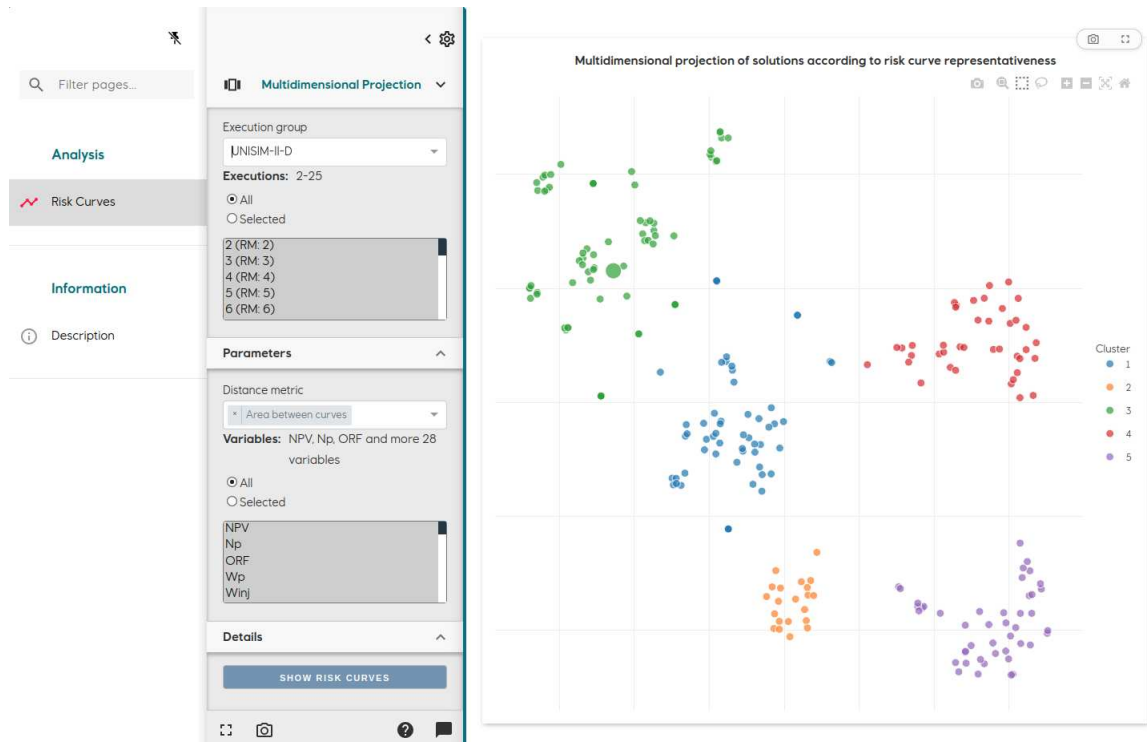


Figura 6.1: Visão geral do sistema de visualização interativo.

Fonte: Dados originais da pesquisa.

Esse sistema contém quatro visualizações. Três delas estão assinaladas como opções a serem selecionadas no retângulo vermelho “1” na Figura 6.2; são elas: projeção multidimensional de soluções de acordo com a representatividade da curva de risco (*Multidimensional Projection*); *heatmap* de modelos usados em cada solução (*Solution by Model*); e *heatmap* de soluções por variáveis de acordo com a representatividade da curva de risco (*Solution by Variable*). Além delas, há uma visualização com o detalhamento das curvas de risco, acionado a partir do botão *Show Risk Curve*, destacado no retângulo vermelho “2” na Figura 6.2.

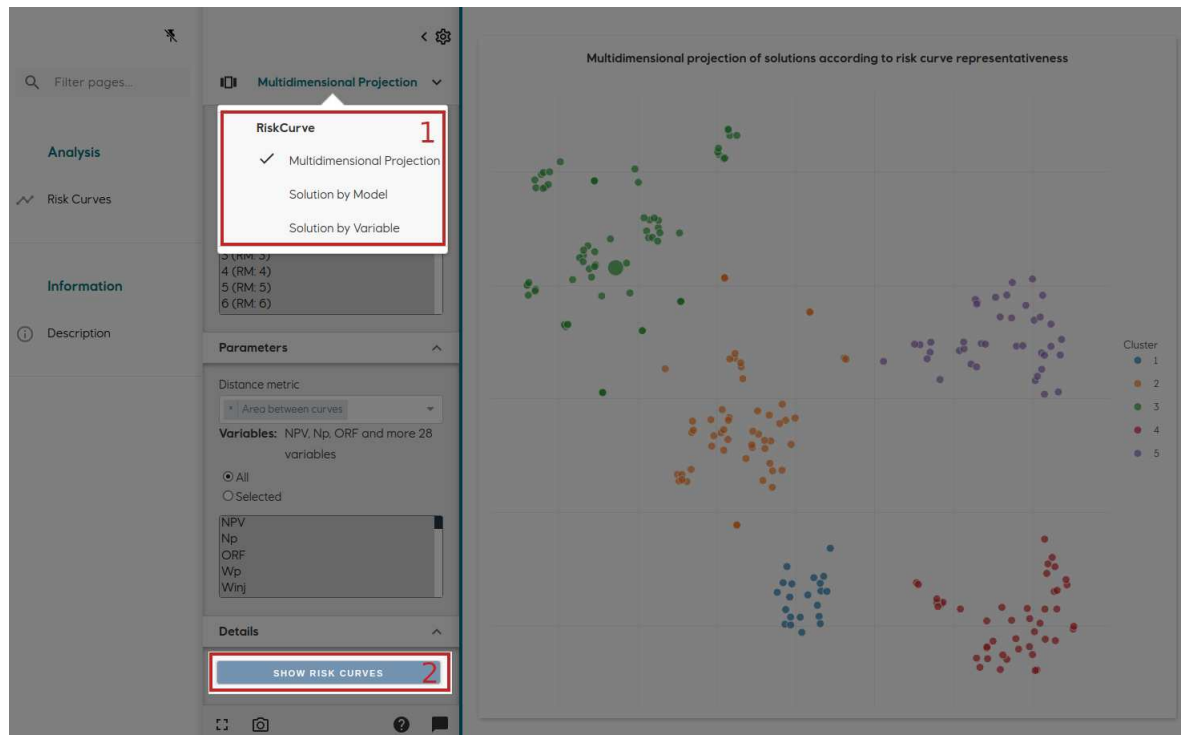


Figura 6.2: Representações visuais disponíveis no sistema de visualização interativo.

Fonte: Dados originais da pesquisa.

Na Seção 6.1 constam os detalhes do desenvolvimento do sistema de visualização interativo. Após isso, na Seção 6.2, será detalhado um caso de uso, com um conjunto de dados do RMFinder, para explorar situações diferentes nas quais as visualizações interativas propostas podem auxiliar os usuários em suas tarefas.

6.1 Visualizações interativas propostas

A implementação do sistema de visualização interativo é descrita nesta presente Seção 6.1. Em cada uma das Seções 6.1.1, 6.1.2 e 6.1.3, há um detalhamento do desenvolvimento de gráficos de visão geral, são eles: projeção multidimensional de curvas de risco; *heatmap* reordenável de variáveis por soluções; e *heatmap* reordenável de modelos utilizados por solução, respectivamente. Na Seção 6.1.4, são apresentadas as seleções de marcas nos gráficos de visão geral de modo a detalhar sob demanda as curvas de riscos. Nas Seções 6.1.5, 6.1.6 e 6.1.7, são abordadas as implementações das representações visuais sob demanda, tais como: curvas de risco sumarizadas por envoltório, curvas de risco justapostas e sobrepostas e curvas de risco do conjunto de todos os modelos, respectivamente. Por fim, na Seção 6.1.8, consta os recursos interativos adicionais presentes no sistema de visualização.

6.1.1 Projeção multidimensional de curvas de risco

O conjunto de dados do RMFinder (Seções 4.1.1, 4.1.2 e 4.1.3) compreende um conjunto τ de modelos. Cada modelo tem valores para um conjunto de variáveis de saída $f \in \{f_1, \dots, f_m\}$, onde $m \in \mathbb{N}$, $1 \leq m \leq 50$, corresponde à quantidade dessas variáveis de saída. O usuário do RMFinder define a quantidade de soluções $n \in \mathbb{N}^*$ para seu problema de otimização, gerando um conjunto de soluções $S = \{S_1, \dots, S_n\}$, sendo $S_i \subseteq \tau$. Cada solução (ou conjunto de MRs) S_i , assim como o conjunto τ de modelos, tem associada uma curva de risco (ou CCDF) Δ^f para cada variável de saída f_i , a qual pode ser representada matematicamente pela Equação 2.2, onde $\Delta^{f_i} : [0, 1] \rightarrow [0, 1]$. Vale considerar que os valores das variáveis de saída f_i são padronizados pela técnica *min-max*, obedecendo a Equação 6.1.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (6.1)$$

Somado a isso, cada variável de saída f_i possui um conjunto de curvas de risco, sendo uma delas associada a τ e as demais associadas às soluções S_1 a S_n . Assim, seja ξ_i o conjunto de curvas de risco associado à variável f_i , de tal forma que $\xi_i = \{\Delta_\tau^{f_i}, \Delta_{S_1}^{f_i}, \dots, \Delta_{S_n}^{f_i}\}$. Considere ξ o conjunto de curvas de risco associado a todas as variáveis de saída, ou seja, $\xi = \{\xi_1, \dots, \xi_m\}$.

Para cada ξ_i é possível determinar uma matriz de dissimilaridade $D_i^\delta = (d_{p,q}^{i,\delta})$, descrita na Equação 6.2, tal que $1 \leq p \leq k$, $1 \leq q \leq k$, e $k = n + 1$ representa a quantidade de soluções n somada a 1 (sendo que este valor “1” representa o conjunto τ de modelos). A matriz de dissimilaridade D_i^δ compara todas as curvas de risco Δ^{f_i} entre si, par a par, de uma variável de saída f_i , utilizando uma medida de distância δ (a ser definida posteriormente nesta seção).

$$D_i^\delta = \begin{bmatrix} 0 & d_{1,2}^{i,\delta} & d_{1,3}^{i,\delta} & \dots & d_{1,k}^{i,\delta} \\ d_{2,1}^{i,\delta} & 0 & d_{2,3}^{i,\delta} & \dots & d_{2,k}^{i,\delta} \\ d_{3,1}^{i,\delta} & d_{3,2}^{i,\delta} & 0 & \dots & d_{3,k}^{i,\delta} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{k,1}^{i,\delta} & d_{k,2}^{i,\delta} & d_{k,3}^{i,\delta} & \dots & 0 \end{bmatrix} \quad (6.2)$$

Conforme essa definição, há m matrizes de dissimilaridade (D_1^δ a D_m^δ). Contudo, para utilizar uma técnica de projeção multidimensional (p. ex., as abordagens assinaladas na Seção 2.1.4), é

necessário haver apenas uma matriz de dissimilaridade $D^{k \times k}$. Para gerar essa matriz, realizou-se o cálculo da média entre as m matrizes de dissimilaridade, como representado pela Equação 6.3:

$$\overline{D}^\delta = \frac{1}{m} \sum_{i=1}^m D_i^\delta \quad (6.3)$$

Para o cálculo da matriz de dissimilaridade, foram propostas duas medidas de distância: (i) a “área entre as curvas de risco” (Figura 2.17); e (ii) o teste não-paramétrico proposto por Kolmogorov-Smirnov (Figura 3.6), representadas matematicamente pelas Equações 6.4 e 6.5, respectivamente, nas quais δ foi substituído por “area” ou “kol_smir” para diferenciar as medidas.

$$d_{p,q}^{i,area} = \int_0^1 |\Delta_p^{f_i}(x) - \Delta_q^{f_i}(x)| dx \quad (6.4)$$

$$d_{p,q}^{i,kol_smir} = \sup_x |\Delta_p^{f_i}(x) - \Delta_q^{f_i}(x)| \quad (6.5)$$

Cada função Δ^{f_i} foi interpolada com uso do método linear³ para realizar o cálculo numérico das medidas de distância das Equações 6.4 e 6.5. Neste contexto, utilizou-se a regra dos trapézios⁴ com precisão $n = 1000$ para integração numérica da Equação 6.4. Esses métodos numéricos foram desenvolvidos com o uso da linguagem *Python* e as bibliotecas *Numpy* e *SciPy*⁵.

Em seguida, após a matriz de dissimilaridade ser pré-computada, é realizado o cálculo da projeção multidimensional, usando o algoritmo t-SNE (MAATEN; HINTON, 2008), o qual foi implementado via biblioteca *Scikit-learn*⁶, de acordo com os parâmetros listados no Apêndice H. Além disso, o usuário pode parametrizar o cálculo da projeção multidimensional, selecionando as variáveis de saída f_i e as métricas de distância das Equações 6.4 e 6.5 a serem consideradas, conforme ilustrado no retângulo vermelho número “2” na Figura 6.3. Se o usuário selecionar ambas as métricas de distância, é calculada a média entre as matrizes de distância de ambas as métricas, sendo $\overline{D} = \frac{1}{2} (\overline{D}^{area} + \overline{D}^{kol_smir})$.

³A interpolação consiste em aproximar $f(x)$ no intervalo $[a, b]$ para qualquer x por meio de outra função $g(x)$.

⁴Consiste em substituir a função integrando $f(x)$ por um polinômio interpolador de grau 1 (reta). O polinômio interpolará os extremos do intervalo de integração.

⁵Algoritmos para computação científica em *Python*.

⁶*Scikit-learn* é uma biblioteca de *machine learning* para a linguagem de programação *Python*.

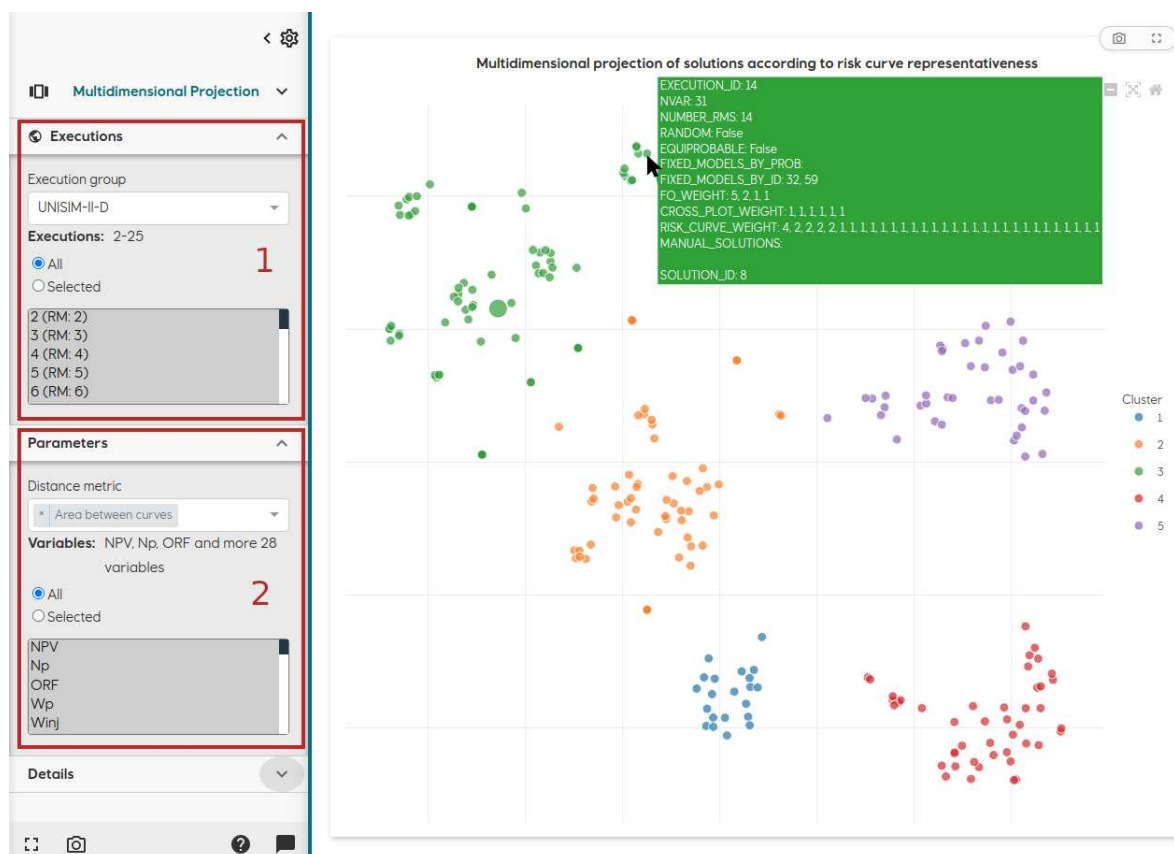


Figura 6.3: Recursos interativos do gráfico de projeção multidimensional de soluções, usando dados do caso de estudo UNISIM-II-D (CORREIA et al., 2015). Os filtros estão destacados pelo retângulo vermelho número “1”, os parâmetros assinalados pelo retângulo vermelho número “2”, e o texto em destaque no retângulo verde com as informações complementares da solução selecionada.

Fonte: Dados originais da pesquisa.

O usuário também pode aplicar filtros (retângulo vermelho número “1” na Figura 6.3) para visualizar a projeção multidimensional de soluções de um grupo de execuções (ou caso de estudo, conforme detalhado na Seção 2.2.4), e das execuções do RMFinder relacionadas a esse grupo. Outro recurso interativo consiste em permitir o detalhamento das execuções (Seção 4.1.2) ao posicionar o ponteiro do *mouse* nas marcas do gráfico (ou seja, nas soluções), como ilustrado na Figura 6.3; neste caso, o texto mostrado na janela de detalhes sob demanda corresponde aos parâmetros informados para a execução do RMFinder que foram responsáveis por gerar a solução em questão.

Além disso, para apoiar o usuário na identificação de possíveis agrupamentos, combinou-se a abordagem de projeção multidimensional com um algoritmo de *clustering* (Seção 2.1.4) para determinar grupos de soluções, estabelecendo cores nas marcas do gráfico associado a um grupo específico, conforme elencado na legenda da Figura 6.3. Nesse caso, utilizou-se o

algoritmo de *clustering X-means* (PELLEG; MOORE, 2000) para encontrar a melhor divisão de *clusters* k , onde $\{k \in \mathbb{N} \mid 2 \leq k \leq 10\}$. Essa abordagem foi implementada na linguagem *Python* com uso da biblioteca *PyClustering*⁷, respeitando os parâmetros elencados no Apêndice I.

A projeção multidimensional via algoritmo t-SNE Maaten e Hinton (2008) é calculada a cada interação do usuário no sistema de visualização (p.ex., ao alterar qualquer valor dos filtros e/ou parâmetros nos retângulos vermelhos números “1” e “2” na Figura 6.3, respectivamente). Em seguida, com base nos dados projetados em duas dimensões, resultado do algoritmo t-SNE, divide-se esse novo conjunto de dados em grupos por meio do algoritmo de *clustering X-means* (PELLEG; MOORE, 2000). Segundo Maaten e Hinton (2008), o algoritmo t-SNE tem uma função de custo que não é convexa, ou seja, com diferentes inicializações (vide os parâmetros utilizados nessa técnica, elencados no Apêndice H), obtém-se resultados diferentes; portanto, a cada interação do usuário no sistema de visualização, a dispersão dos pontos (ou soluções) na Figura 6.3, assim como a divisão em grupos desses pontos, podem ser diferentes. Embora o algoritmo de *clustering X-means* não contenha componentes aleatórios que possam influenciar o seu resultado, a divisão em grupos pode variar de acordo com os dados de entrada utilizados nessa abordagem. Ou seja, a distribuição dos dados em um espaço de duas dimensões, projetada por meio do algoritmo t-SNE, pode variar, influenciando o resultado da clusterização.

Por fim, como exemplo de interpretação possível da Figura 6.3, observa-se a formação de cinco grupos (*clusters*) de curvas de risco. Dentre eles, o grupo com a cor verde, no canto superior esquerdo da projeção, contém a curva de risco do grupo de modelos originais, destacada como marca de maior tamanho. As soluções ao redor dessa marca pertencem ao mesmo grupo, o que é um indicativo de que as curvas desse grupo são as que mais se aproximaram da curva dos modelos originais. Também é possível verificar que os outros grupos formados apresentam uma certa coesão e podem ser investigados em outros gráficos para analisar melhor a variabilidade dessas soluções e sua representatividade.

6.1.2 *Heatmap* reordenável de variáveis por soluções

O *heatmap* de variáveis por soluções é uma representação visual compacta para comparar pares de curvas de risco. Dadas duas curvas de risco de uma mesma variável f_q – uma para

⁷A biblioteca *PyClustering* é uma coleção de análise de *cluster*, ferramentas para visualização e análise de resultados, etc.

modelos representativos de uma solução S_p e outra para todo o conjunto de modelos τ – a soma das áreas entre elas (Figura 2.17) ou a distância máxima entre elas (Figura 3.6) pode ser considerada como um valor de erro (ou de “representatividade baseada em curvas de risco”).

Sendo assim, determina-se uma matriz numérica $H^{n \times m} = (h_{p,q})$, $1 \leq p \leq n$, $1 \leq q \leq m$, onde n corresponde à quantidade de soluções e m corresponde à quantidade de variáveis de saída. Cada célula $h_{p,q} = d_{S_p,\tau}^{q,\delta}$ dessa matriz representa o valor do erro relacionado à solução S_p e à variável f_q , com relação a uma das medidas de distância δ das Equações 6.4 e 6.5; nesse caso, são comparadas as curvas de risco $\Delta_{S_p}^{f_q}$ e $\Delta_{\tau}^{f_q}$, sendo a variável f_q padronizada pela Equação 6.1. Em seguida, os valores calculados $h_{p,q}$ também são padronizados por meio da Equação 6.1, gerando a matriz $\hat{H}^{n \times m}$.

Tal qual realizado no cálculo da projeção multidimensional (Seção 6.1.1), as curvas de risco $\Delta_{S_p}^{f_q}$ e $\Delta_{\tau}^{f_q}$ foram interpoladas com uso do método linear para realizar o cálculo numérico das medidas de distância das Equações 6.4 e 6.5, e a integração numérica da Equação 6.4 utilizou a regra dos trapézios com precisão $n = 1000$. Essa matriz numérica foi implementada com uso da linguagem *Python* e as bibliotecas *Numpy*, *SciPy* e *Pandas*.

Na Figura 6.4, a matriz $\hat{H}^{n \times m}$ é ilustrada como um *heatmap*. O eixo Y apresenta as soluções S_p , enquanto o eixo X contém as variáveis de saída f_q . A cor de cada célula representa o valor do erro ao qual a célula se refere. O mapa de cores tem uma escala verde-branca-rosa, em que o verde representa a ausência de erro, branco um valor neutro arbitrário, e o rosa, o maior erro no conjunto de dados (SILVA et al., 2020). Além disso, o usuário pode escolher a medida de distância a ser considerada no cálculo da matriz, selecionando a “área entre as curvas de risco” ou o teste não paramétrico Kolmogorov-Smirnov (ambos disponíveis no retângulo vermelho número “3” na Figura 6.4). Vale notar que valores baixos são desejados para ambas as medidas.

É importante destacar que a legenda de cores foi adotada de forma idêntica ao feito por Silva et al. (2020), cuja escolha das cores visava usar cores diferentes das já utilizadas em outros gráficos dos usuários-alvo e que já possuíam outros significados para eles (referentes a concentrações de óleo, gás e água). O valor neutro na cor branca no mapa de cores foi definido empiricamente como 0, 2.

Se o usuário posicionar o cursor do *mouse* em uma célula do *heatmap*, como ilustrado na Figura 6.4, é possível visualizar o detalhamento das execuções (Seção 4.1.2) e os valores categóricos dos eixos X e Y (solução e variável de saída, respectivamente).



Figura 6.4: Recursos interativos do *heatmap* reordenável de variáveis por soluções, usando dados do caso de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017). Os filtros estão destacados pelos retângulos vermelhos “1” e “2”, as medidas de distância assinaladas pelo retângulo vermelho “3”, e o texto em destaque no retângulo preto contém as informações complementares da célula selecionada.

Fonte: Dados originais da pesquisa.

Os retângulos vermelhos “1” e “2” na Figura 6.4 correspondem a filtros a serem aplicados pelo usuário de acordo com suas ações e tarefas. O usuário pode selecionar as soluções de um grupo de execuções e, ao mesmo tempo, selecionar as execuções do RMFinder desse grupo (retângulo vermelho “1” na Figura 6.4); e/ou filtrar as variáveis de saída consideradas no conjunto de execuções selecionado (retângulo vermelho “2” na Figura 6.4). Somado a isso, o usuário tem à disposição outros recursos interativos no retângulo vermelho “3” na Figura 6.4; são eles: mostrar (ou ocultar) os rótulos dos eixos *X* e *Y*, redimensionando o tamanho da fonte de acordo com a resolução do navegador e a quantidade de rótulos a ser exibida; mostrar (ou ocultar) a legenda do mapa de cores; e escolher entre reordenar a matriz ou mantê-la na sua forma original (ou seja, com colunas ordenadas em ordem alfanumérica crescente dos nomes das variáveis, e com linhas ordenadas em ordem alfanumérica crescente dos rótulos que identificam as soluções).

Como as variáveis de saída e as soluções são variáveis categóricas, as linhas e colunas da matriz podem ser permutadas se necessário; portanto, reordenações dessa matriz podem ser calculadas e apresentadas aos usuários para revelar possíveis padrões nos dados. Sendo assim, utilizou-se o algoritmo *Optimal Leaf Ordering* (BAR-JOSEPH; GIFFORD; JAAKKOLA, 2001; BARONI; SILVA, 2022) para reordenar essa matriz, implementado na linguagem *Python* com uso da biblioteca *SciPy*. Ele efetua um agrupamento hierárquico (*hierarchical clustering*) e, em seguida, procura permutar o dendrograma de forma que elementos vizinhos nas folhas do dendrograma sejam o mais semelhantes possível entre si.

Em particular, a Figura 6.4 apresenta a medida de distância Kolmogorov-Smirnov, que deve ter o menor valor possível. A matriz está reordenada, e nessa reordenação é possível ver que aproximadamente um terço das soluções se apresenta com valores baixos (células verdes) para quase todas as variáveis de saída. Também é possível notar certa dificuldade de cerca de metade das soluções em prover boa representatividade para as variáveis *WEIPR2* e *WEIPR3* (índice econômico dos poços produtores números dois e três, respectivamente).

6.1.3 *Heatmap* reordenável de modelos utilizados por solução

O *heatmap* de modelos utilizados por solução é um mapeamento visual compacto para encontrar padrões relevantes em um conjunto de soluções $S = \{S_1, \dots, S_n\}$, sendo $S_p \subseteq \tau$, em relação à sua variabilidade de subconjuntos comuns de MRs. Cada solução S_p tem um conjunto de MRs $\tau'_p = \{r_1, \dots, r_n\}$ (vide arquivo de saída do RMFinder, ilustrado na Figura 4.4), onde $\{\tau'_p \subseteq \tau \wedge |\tau'_p| \ll |\tau|\}$ e τ é o conjunto de todos os modelos.

Sendo assim, determina-se uma matriz binária $B^{n \times m} = (b_{p,q})$, $1 \leq p \leq n$, $1 \leq q \leq m$, onde n corresponde à quantidade de soluções e m corresponde à quantidade de modelos. Cada célula $b_{p,q}$ dessa matriz binária com domínio $\beta = \{0, 1\}$ é definida pela Equação 6.6, de tal forma que a célula com o valor 1 é um modelo $\tau_q \in \tau'_p$ (i. e., modelo selecionado pelo RMFinder como MR na solução S_p) e a célula com o valor 0 é um modelo $\tau_q \notin \tau'_p$ (i. e., modelo não selecionado pelo RMFinder como MR na solução S_p). Essa matriz binária foi implementada com uso da linguagem *Python* e as bibliotecas *Numpy* e *Pandas*.

$$b_{p,q} = \begin{cases} 1 & \tau_q \in \tau'_p \\ 0 & \tau_q \notin \tau'_p. \end{cases} \quad (6.6)$$

Na Figura 6.5 a matriz binária $B^{n \times m}$ é ilustrada como um *heatmap*. O eixo Y apresenta as soluções S_p , enquanto o eixo X contém os modelos τ_q . A célula azul-escura representa o modelo utilizado na solução como MR, enquanto a cor azul-clara indica a ausência do modelo na solução. A cor vermelha representa o modelo fixado na configuração dos parâmetros de execução do RMFinder (Seção 4.1.2).

O cursor do mouse é posicionado em uma célula do *heatmap*, conforme ilustrado na Figura 6.5, permitindo a visualização de detalhes das execuções (Seção 4.1.2) e os valores categóricos dos eixos X e Y (solução e modelo, respectivamente).

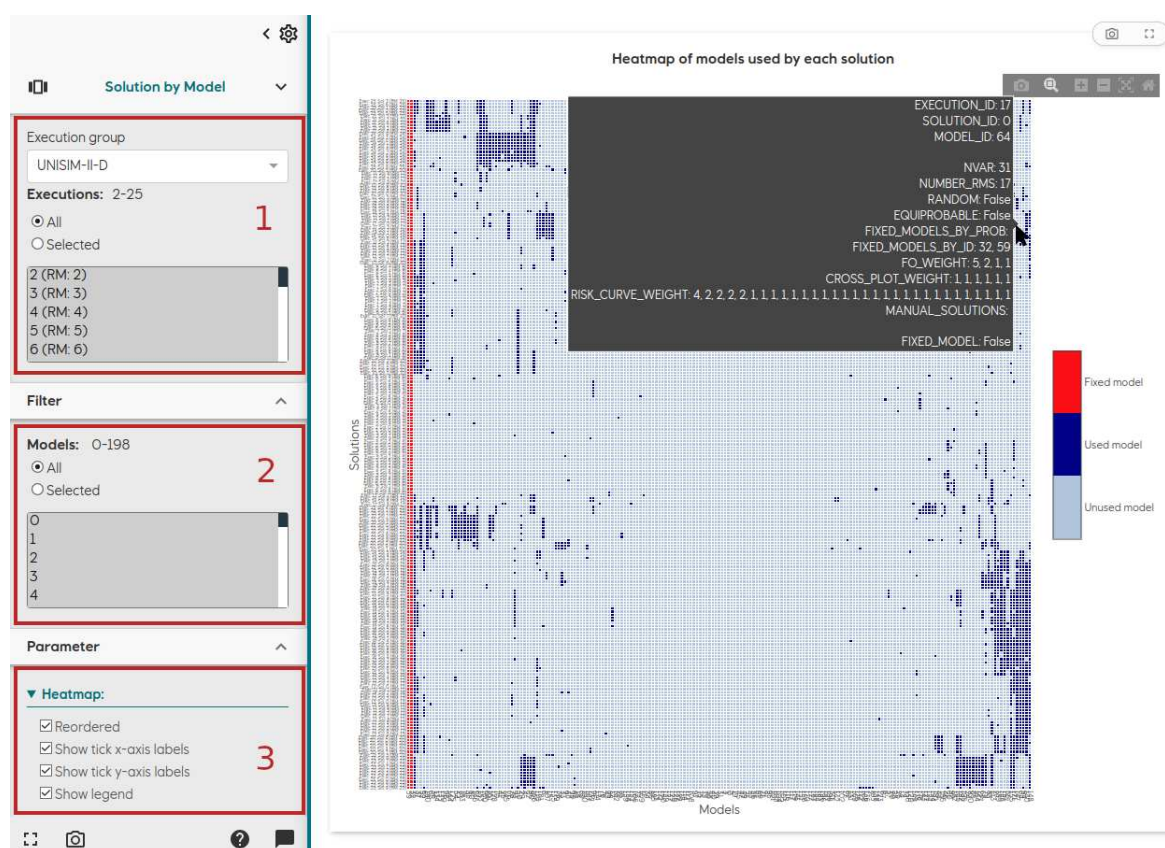


Figura 6.5: Recursos interativos do *heatmap* reordenável de modelos utilizados por solução, usando dados do caso de estudo UNISIM-II-D (CORREIA et al., 2015). Os filtros estão destacados pelos retângulos vermelhos “1” e “2”, as medidas de distância assinaladas pelo retângulo vermelho “3”, e o texto em destaque no retângulo preto contém as informações complementares da célula selecionada.

Fonte: Dados originais da pesquisa.

Tal qual no *heatmap* de variáveis por soluções (Seção 6.1.2), os retângulos vermelhos “1” e “2” na Figura 6.5 correspondem a filtros a serem aplicados pelo usuário de acordo com suas ações e tarefas. O usuário pode escolher soluções para um grupo de execução do RMFinder, e as execuções do RMFinder relacionadas a esse grupo (retângulo vermelho “1” na Figura 6.5);

e/ou filtrar os modelos considerados no conjunto de execuções selecionado (retângulo vermelho “2” na Figura 6.5). Além disso, o usuário tem disponíveis outros recursos interativos no retângulo vermelho “2” na Figura 6.5, tais como: mostrar (ou ocultar) os rótulos dos eixos X e Y , redimensionando a fonte de acordo com a resolução do navegador e a quantidade de rótulos a serem exibidos; mostrar (ou ocultar) a legenda; e escolher entre reordenar a matriz ou mantê-la na sua forma original (ou seja, respeitando a ordenação crescente dos valores dos eixos X e Y).

De forma idêntica ao feito no *heatmap* da Seção 6.1.2, este também pode ser reordenado pelo usuário para tentar evidenciar padrões relevantes, uma vez que solução e modelo são variáveis categóricas. Também se utilizou nesse caso o algoritmo *Optimal Leaf Ordering* (BAR-JOSEPH; GIFFORD; JAAKKOLA, 2001; BARONI; SILVA, 2022) para reordenar essa matriz (opção de “Reordered” no retângulo “3” na Figura 6.5).

Na Figura 6.5 é possível visualizar a matriz reordenada das 230 soluções (que variam de 2 a 25 MRs) e dos 198 modelos do conjunto de dados do caso de estudo UNISIM-II-D (CORREIA et al., 2015) por meio de um *heatmap*. Há dois modelos fixos (de número 32 e 59, listados no texto em destaque no rótulo “FIXED_MODELS_BY” na Figura 6.5), posicionados nas duas primeiras colunas e assinalados com a cor vermelha. Após a reordenação da matriz, é possível visualizar, nas primeiras e últimas colunas, modelos presentes em diversas soluções (cor azul-escura). No entanto, as colunas no meio do *heatmap* (cor azul-clara) indicam que aproximadamente um terço dos modelos desse conjunto de dados não foram utilizados em nenhuma solução. Dessa forma, é possível afirmar uma tendência do problema de otimização do RMFinder selecionar dois terços dos modelos como representativos para as diferentes soluções desse conjunto de dados; no entanto, os modelos não utilizados (compreendendo uma fração de um terço), sob a perspectiva do RMFinder, estão sempre bem representados por outros MRs, apesar de não terem sido escolhidos como representativos.

6.1.4 Detalhamento das curvas de risco

Conforme ilustrado na Figura 6.6, o usuário pode selecionar marcas no gráfico de projeção multidimensional de curvas de risco, no *heatmap reordenável* de variáveis por soluções e no *heatmap reordenável* de modelos utilizados por solução (Figuras 6.6a, 6.6b e 6.6c, respectivamente), e, em seguida, clicar no botão “Show Risk Curve” (retângulo vermelho “2” da Figura 6.2) para visualizar o detalhamento das curvas de risco associadas a essas marcas

(recurso interativo denominado elaboração, elencado na Seção 2.1.5). Se nenhuma marca for selecionada nestes gráficos, ao clicar no botão “Show Risk Curve”, considera-se apenas os filtros aplicados pelo usuário para realizar a seleção do detalhamento das curvas de risco. A lógica por trás do detalhamento das curvas de risco (ou do sistema de visualização implementado) foi inspirada no *visual information seeking mantra* de Shneiderman (1996) (visão geral, *zoom* e filtros, e então detalhes sob demanda), elencado na Seção 2.1.5.

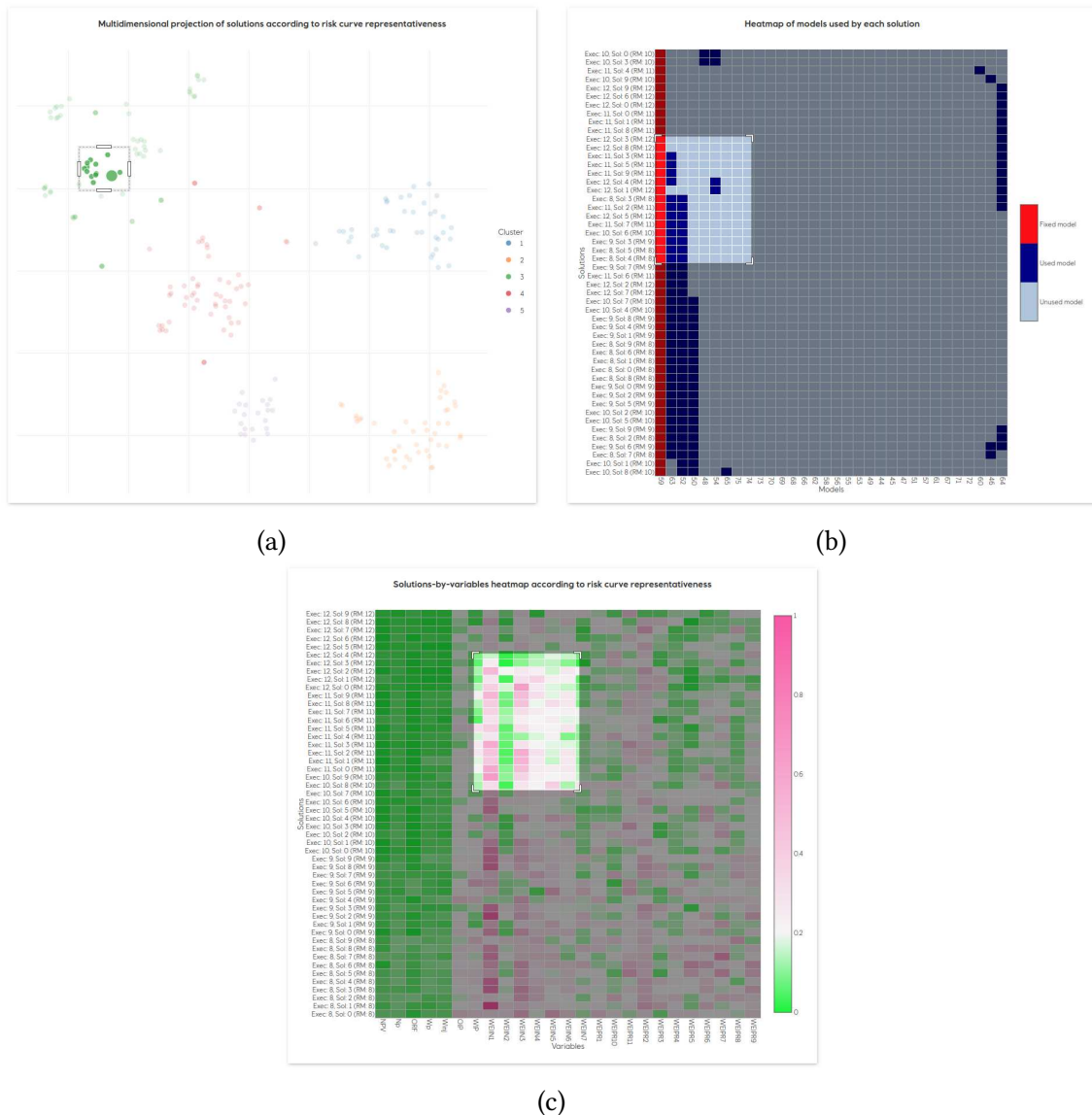


Figura 6.6: Seleção das marcas dos gráficos para serem detalhadas as curvas de risco associadas a elas: (a) projeção multidimensional de curvas de risco; (b) *heatmap* reordenável de variáveis por soluções; e (c) *heatmap* reordenável de modelos utilizados por solução.

Fonte: Dados originais da pesquisa.

No detalhamento das curvas de risco, o usuário tem três visualizações interativas, as quais serão descritas a seguir; são elas:

1. Curvas de risco sumarizadas por envoltório (Seção 6.1.5);
2. Curvas de risco justapostas e sobrepostas (Seção 6.1.6); e
3. Curvas de risco do conjunto de todos os modelos (Seção 6.1.7).

6.1.5 Curvas de risco sumarizadas por envoltório

O gráfico de curvas de risco sumarizadas por envoltório é uma visualização para fornecer uma visão resumida da semelhança das soluções $S = \{S_1, \dots, S_n\}$, sendo $S_j \subseteq \tau$, em relação às suas curvas de risco (ou CCDF) Δ^{f_i} para cada variável de saída f_i .

Cada curva de risco pode ser representada matematicamente pela Equação 2.2 aplicada a uma solução S_j ; ou seja, $\Delta_{S_j}^{f_i} = acc^{f_i}(S_j, \tau)$, onde $\Delta_{S_j}^{f_i} : \mathbb{R} \rightarrow [0, 1]$. Com isso, o gráfico de uma curva de risco de uma solução S_j com relação à variável f_i mapeia visualmente no eixo X a variável f_i , e no eixo Y os valores correspondentes de $\Delta_{S_j}^{f_i}$. Ou seja, se x for um valor da variável f_i , haveria no gráfico da curva de risco de f_i um ponto $(x, \Delta_{S_j}^{f_i}(x))$.

Por sua vez, o envoltório, como o nome sugere, envolve visualmente todas as curvas de risco, definindo uma curva que representa o limite superior e outra que representa o limite inferior do conjunto de curvas de risco em questão. Desse modo, o envoltório referente às curvas de risco relativas à variável f_i é definido pelo menor e pelo maior valor de $\Delta_{S_j}^{f_i}(x)$, para todas as soluções $S_j \in S$, sendo que x pode variar entre o menor e o maior valor possível de f_i para todas as soluções em S . Em outras palavras, os limites superior e inferior do envoltório referente à variável f_i e às soluções em S são dados pelas Equações 6.7 e 6.8, respectivamente.

$$u_{f_i}^S(x) = \max \left\{ \Delta_{S_1}^{f_i}(x), \dots, \Delta_{S_n}^{f_i}(x) \right\} \quad (6.7)$$

$$l_{f_i}^S(x) = \min \left\{ \Delta_{S_1}^{f_i}(x), \dots, \Delta_{S_n}^{f_i}(x) \right\} \quad (6.8)$$

Cada curva de risco $\Delta_{S_j}^{f_i}$ foi interpolada com uso do método linear para realizar o cálculo dos limites superior e inferior, Equações 6.7 e 6.8, respectivamente. Esse cálculo foi implementado com uso da linguagem *Python* e as bibliotecas *Numpy*, *SciPy* e *Pandas*.

Na Figura 6.7 são ilustradas as curvas de risco sumarizadas por envoltório. Os pontos assinalados em roxo correspondem à curva de risco de todos os modelos do conjunto τ (rotulada como “*Original model*” na legenda). Essa curva é envolvida por uma região cinza

(rotulada como “*Summarization*” na legenda), a qual representa, nesse exemplo, o envoltório dos conjuntos das curvas de risco de 97 soluções (retângulo número “3”), previamente selecionadas, nesse caso, do gráfico de projeção multidimensional de curvas de risco (Seção 6.1.1).

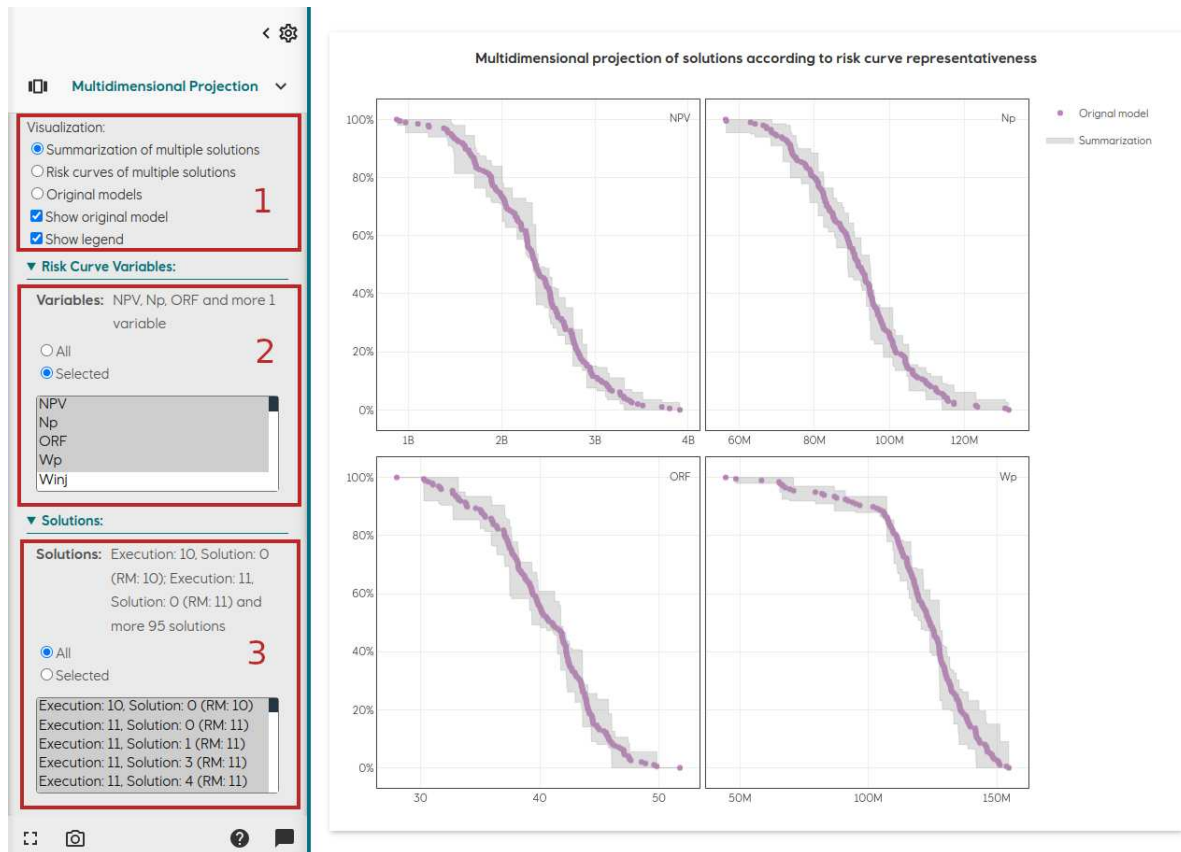


Figura 6.7: Recursos interativos do gráfico curvas de risco sumarizadas por envoltório, usando dados do caso de estudo UNISIM-II-D (CORREIA et al., 2015). Os parâmetros assinalados pelo retângulo vermelho número “1” e os filtros estão destacados pelos retângulos vermelho números “2” e “3”.

Fonte: Dados originais da pesquisa.

O retângulo vermelho “2” na Figura 6.7 compreende a quantidade de curvas de risco (uma para cada variável de saída) a ser exibida na visualização. O usuário pode filtrar as soluções a serem consideradas na sumarização de suas curvas de risco (retângulo vermelho “3” na Figura 6.7). Somado a isso, o usuário tem à disposição outros recursos interativos no retângulo vermelho “1” na Figura 6.7; são eles: mostrar (ou ocultar) a curva de risco de todos os modelos (pontos roxos); mostrar (ou ocultar) a legenda; e selecionar o tipo de visualização a ser exibida, nesse caso, “*Summarization of multiple solutions*”.

Na Figura 6.7 observa-se o gráfico de curvas de risco sumarizadas de quatro variáveis de saída (*NPV*, *Np*, *ORF* e *Wp*) sendo exibidas. Essas variáveis de saída correspondem às

variáveis utilizadas no conceito de distorção de representatividade do conjunto de MRs do problema de otimização do RMFinder (i.e., a função objetivo *cross-plot*, detalhada na Seção 2.3.1). Nesse exemplo, a quantidade de MRs que compõem cada solução retornada pelo RMFinder varia entre 10 e 25 MRs. Visualmente, os envoltórios dessas quatro variáveis de saída representam um contorno contínuo (considerando os limites superior e inferior) ao longo de todo o seu domínio em relação à curva de risco de todos os modelos. Além disso, os limites visualizados nos envoltórios dessas quatro variáveis de saída apontam para uma pequena variação, ou seja, as curvas de risco associadas às soluções selecionadas estão próximas da curva de risco de todos os modelos. Dessa forma, uma possível interpretação, considerando as quatro variáveis de saída analisadas, seria hipotetizar a existência de soluções com 10 MRs com uma “representatividade baseada em curvas de risco” tão adequada quanto a soluções com 25 MRs, onde quanto menor o número de MRs selecionados, melhor para reduzir a quantidade de recursos computacionais a serem utilizados em passos de simulação seguintes ao da escolha de MRs.

Os recursos interativos do sistema de visualização (filtros, abstração e elaboração, elencados na Seção 2.1.5) ajudaram a guiar essa descoberta, após serem aplicados na visão geral (Figura 6.8) do caso de estudo UNISIM-II-D (CORREIA et al., 2015), compreendendo 231 soluções, e 31 variáveis de saída.

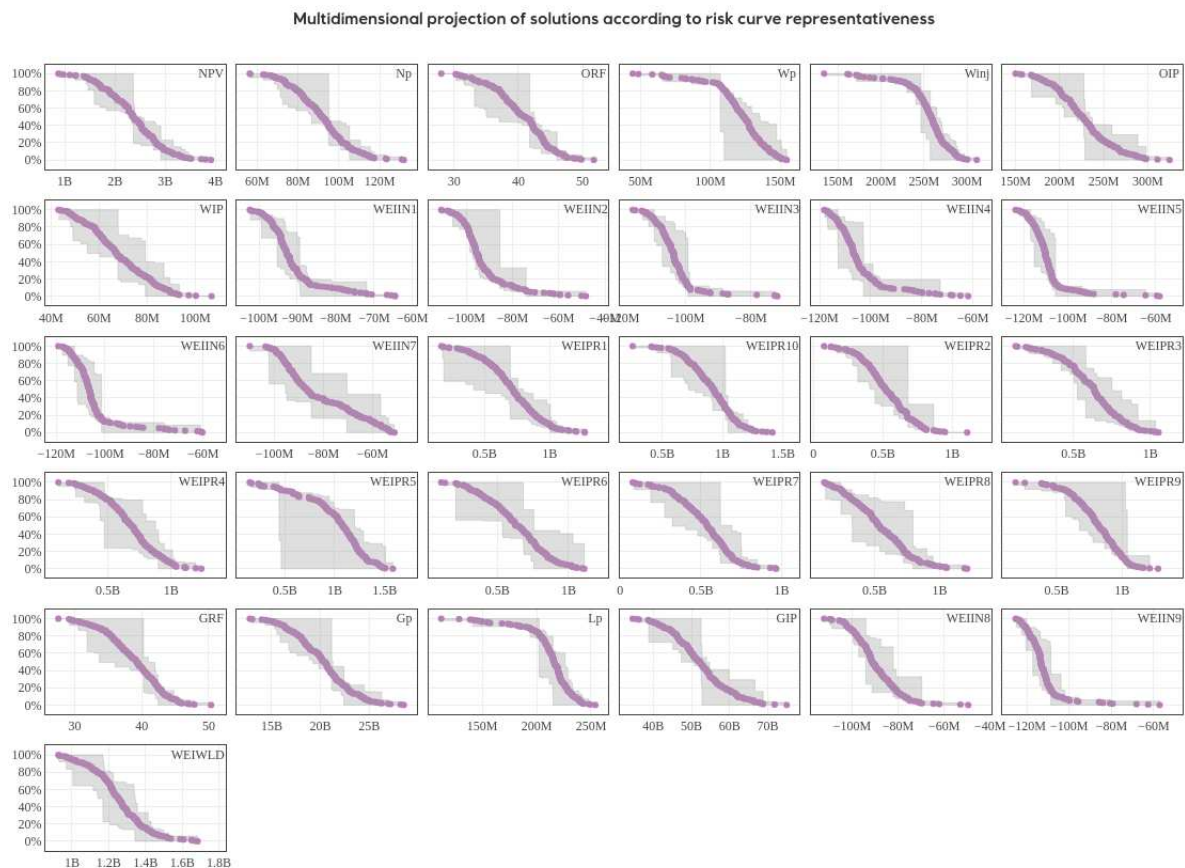


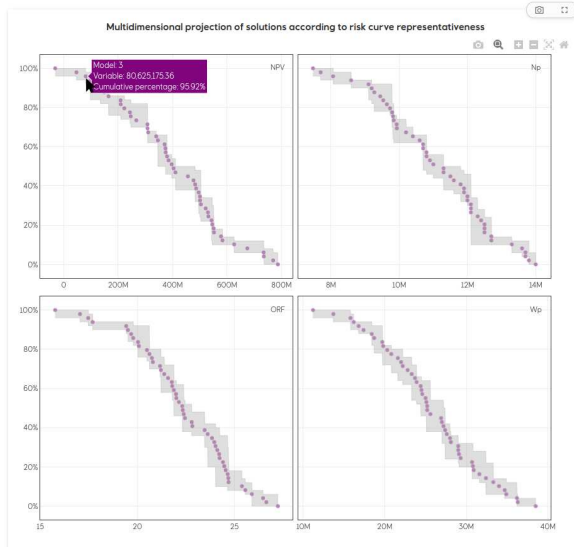
Figura 6.8: Visão geral das curvas de risco de 231 soluções, e 31 variáveis de saída sumarizadas por envoltório, usando dados do caso de estudo UNISIM-II-D (CORREIA et al., 2015).

Fonte: Dados originais da pesquisa.

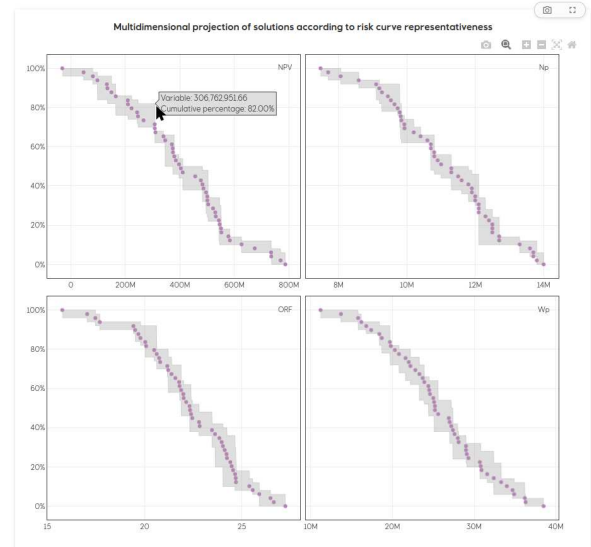
Para auxiliar o usuário a identificar os valores das marcas do gráfico, adicionou-se outros recursos interativos (Figura 6.9) ao posicionar ou clicar o ponteiro do *mouse* nelas; são eles:

- (i) texto em destaque com informações do modelo (valor da variável de saída e percentual acumulado) ao posicionar o ponteiro do *mouse* na marca (pontos roxos) da curva de risco de todos os modelos (Figura 6.9a);
- (ii) texto em destaque com os valores do limite superior (ou inferior) do envoltório das curvas de risco associadas às soluções (sombra cinza) ao posicionar o ponteiro do *mouse* na marca desse gráfico (Figura 6.9b);
- (iii) seleção de modelos (pontos roxos) na curva de risco de uma variável de saída, ao clicar com o botão esquerdo do *mouse* (utilizando tecla SHIFT para multi-seleção), e destacá-los nas curvas de risco das outras variáveis de saída (vide recurso interativo de *brushing* e *linking* na Seção 2.1.5) (Figura 6.9c); e

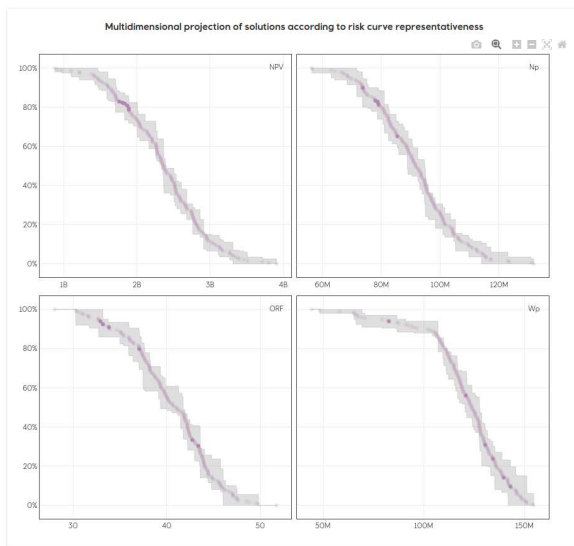
- (iv) curvas de risco sumarizadas por envoltório (sombra cinza) com a oclusão da curva de risco de todos os modelos (pontos roxos), desabilitada por meio da opção “Show original model” (retângulo vermelho “1” na Figura 6.7) (Figura 6.9d).



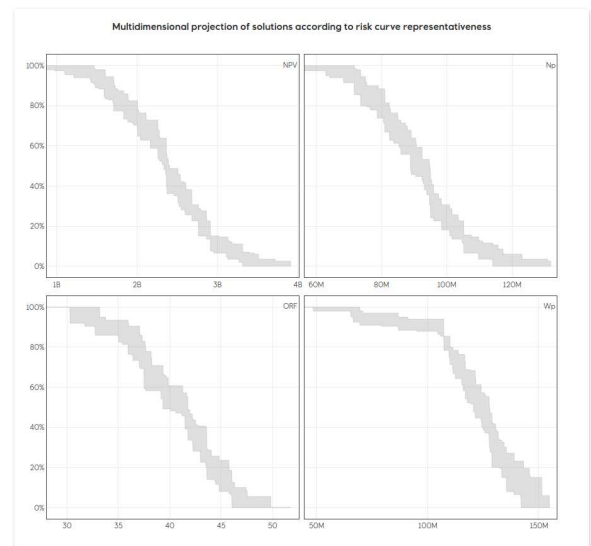
(a)



(b)



(c)



(d)

Figura 6.9: Recursos interativos do gráfico curvas de risco sumarizadas por envoltório: (a) o texto em destaque no retângulo roxo contém as informações do modelo na marca selecionada; (b) o texto em destaque no retângulo cinza contém as informações do limite do envoltório na marca selecionada; (c) seleção de modelos em uma curva de risco de uma variável de saída para destacá-los nas outras variáveis de saída; e (d) visualização contendo apenas a curva de risco sumarizadas por envoltório (sombra cinza).

Fonte: Dados originais da pesquisa.

6.1.6 Curvas de risco justapostas e sobrepostas

O gráfico curvas de risco justapostas e sobrepostas é uma matriz de gráficos de curva de risco, onde cada célula compreende uma variável de saída f_i , sobrepondo suas respectivas curvas risco (ou CCDF) Δ^{f_i} das soluções $S = \{S_1, \dots, S_n\}$, sendo $S_j \subseteq \tau$. Cada curva de risco aplicada a uma solução S_j tem um conjunto de cenários mais representativos (ou conjunto de MRs) $\tau' = \{r_1, \dots, r_n\}$, de tal forma que cada MR r_k tem uma probabilidade cumulativa $p_c(r_k)$ para cada solução S_j (arquivo de saída do RMFinder, ilustrado na Figura 4.4). Sendo assim, a distribuição empírica $F^{f_i}(r)$ é uma função degrau que dá um salto de $p_c(r_k)$ na ocorrência de cada MR $r \in \tau'$ para uma variável de saída f_i , na qual corresponde à definição de três pontos: superior, MR e inferior, representados matematicamente pelas Equações 6.9, 6.10 e 6.11, respectivamente.

$$F_{sup}^{f_i}(r) = \sum_{k=1}^N p_c(r_k)_{\{f_i(r) \leq f_i(r_k)\}} \quad (6.9)$$

$$F_{rm}^{f_i}(r) = \frac{p_c(r)}{2} + \sum_{k=1}^N p_c(r_k)_{\{f_i(r) < f_i(r_k)\}} \quad (6.10)$$

$$F_{inf}^{f_i}(r) = \sum_{k=1}^N p_c(r_k)_{\{f_i(r) < f_i(r_k)\}} \quad (6.11)$$

Nas Equações 6.9, 6.10 e 6.11, $f_i(r)$ é o valor da variável de saída f_i do MR $r_k \in \tau'$ e $N = |\tau'|$, ou seja, a quantidade de MRs de cada solução S_j definida pelo usuário para seu problema de otimização (arquivo de entrada do RMFinder, detalhado na Seção 4.1.2). Portanto, é possível definir a distribuição empírica $F(r)$ pela Equação 6.12 $\forall r \in \tau'$, considerando uma variável de saída f_i .

$$F^{f_i}(r) = \langle F_{sup}^{f_i}(r), F_{rm}^{f_i}(r), F_{inf}^{f_i}(r) \rangle : \mathbb{N} \rightarrow [0, 1] \times [0, 1] \times [0, 1] \quad (6.12)$$

Para elucidar o funcionamento da distribuição empírica $F(r)$, na Figura 6.11, é apresentado um exemplo de uma curva de risco da variável de saída NPV , considerando as probabilidades cumulativas $p_c(r_k)$ de uma das soluções dos dados do caso de estudo UNISIM-I-D (GASPAR et al., 2015) com quantidade de MRs $N = 3$, ilustrado na Figura 6.10.


```
#modelID(209) with prob 0.3177570093457944
#modelID(83) with prob 0.4158878504672897
#modelID(144) with prob 0.26635514018691586
##ID
209:83:144:
```

Figura 6.10: Arquivo de saída do RMFinder para uma solução com três MRs dos dados do caso de estudo UNISIM-I-D (GASPAR et al., 2015).

Fonte: Dados originais da pesquisa.

A Tabela 6.1 apresenta detalhadamente o cálculo das Equações 6.9, 6.10 e 6.11 para cada MR da solução destacada na Figura 6.10, considerando a variável de saída NPV e os dados do caso de estudo UNISIM-I-D (GASPAR et al., 2015).

Tabela 6.1: Cálculo da distribuição empírica $F(r)$ de uma solução com três MRs da variável de saída NPV dos dados do caso de estudo UNISIM-I-D (GASPAR et al., 2015).

k	MR (ou r)	$f_{NPV}(r_k)$	$p_c(r_k)$	$F_{sup}^{NPV}(r_k)$	$F_{rm}^{NPV}(r_k)$	$F_{inf}^{NPV}(r_k)$
1	209	924.534.688,40	0,3177	1,0000	0,8411	0,6823
2	83	1.634.816.578,00	0,4159	0,6823	0,4743	0,2664
3	144	2.236.494.379,00	0,2664	0,2664	0,1332	0,0000

Na Figura 6.11 os pontos roxos representam a curva de risco de todos os modelos do conjunto τ ; os pontos pretos representam os MRs, obedecendo à coordenada $(f_{NPV}(r_k), F_{rm}^{NPV}(r_k))$; o tamanho das barras verticais pretas indica a probabilidade individual $p_c(r_k)$ de cada MR; e os pontos superior e inferior (linha preta horizontal na extremidade de cada barra vertical) compreendem as coordenadas $(f_{NPV}(r_k), F_{sup}^{NPV}(r_k))$ e $(f_{NPV}(r_k), F_{inf}^{NPV}(r_k))$, respectivamente. Esses cálculos foram implementados com uso da linguagem *Python* e as bibliotecas *Numpy* e *Pandas*.

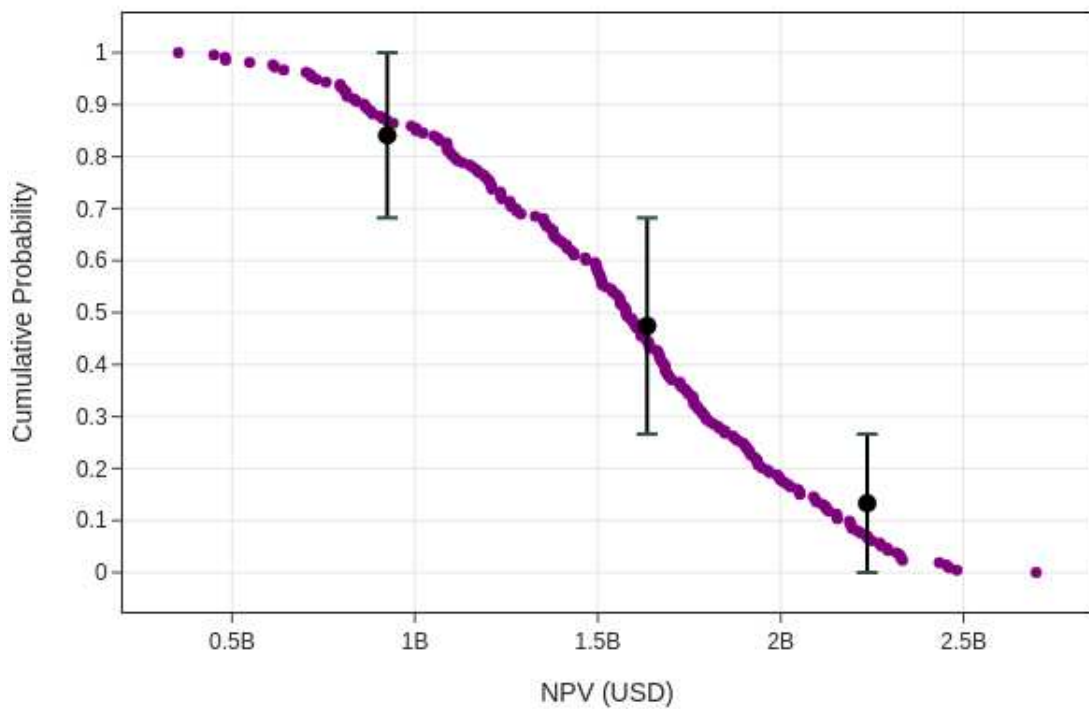


Figura 6.11: Exemplo do cálculo da distribuição empírica $F(r)$ de uma solução com três MRs da variável de saída NPV , usando dados do caso de estudo UNISIM-I-D (GASPAR et al., 2015).

Fonte: Dados originais da pesquisa.

Na Figura 6.12 é ilustrado o gráfico de curvas de risco justapostas e sobrepostas. Os pontos assinalados em roxo correspondem à curva de risco de todos os modelos do conjunto τ (rotulada como “Original Model” na legenda). Os gráficos de linhas sobrepostas, rotulados como “Solution” na legenda, representam as curvas de risco de cada solução selecionada (nesse caso, do gráfico *heatmap* reordenável de modelos utilizados por solução – Seção 6.1.3), onde o eixo X compreende os valores da variável de saída f_i , enquanto o eixo Y corresponde à probabilidade cumulativa p_c (ou a distribuição empírica, detalhada na Equação 6.12).

Ainda na Figura 6.12, os pontos assinalados na cor vermelha (rotulados como “Fixed model” na legenda) correspondem aos modelos definidos como parte integrante de cada solução, i. e. estes modelos foram solicitados a serem fixados na solução pelo usuário na configuração dos parâmetros de execução do RMFinder (Seção 4.1.2). Os pontos destacados na cor azul (rotulados como “Used model” na legenda) compreendem os modelos selecionados pelo usuário no gráfico *heatmap* reordenável de variáveis por soluções (Figura 6.6c) e pertencem ao conjunto de MRs de uma ou mais soluções, destacadas no retângulo vermelho número “3”. Esses pontos na cor azul somente são assinalados no gráfico de curvas de risco justapostas e sobrepostas quando oriundos de uma seleção aplicada pelo

usuário no gráfico *heatmap* reordenável de modelos utilizados por solução (Figura 6.6c). Os pontos na cor preta representam os demais MRs de cada solução.

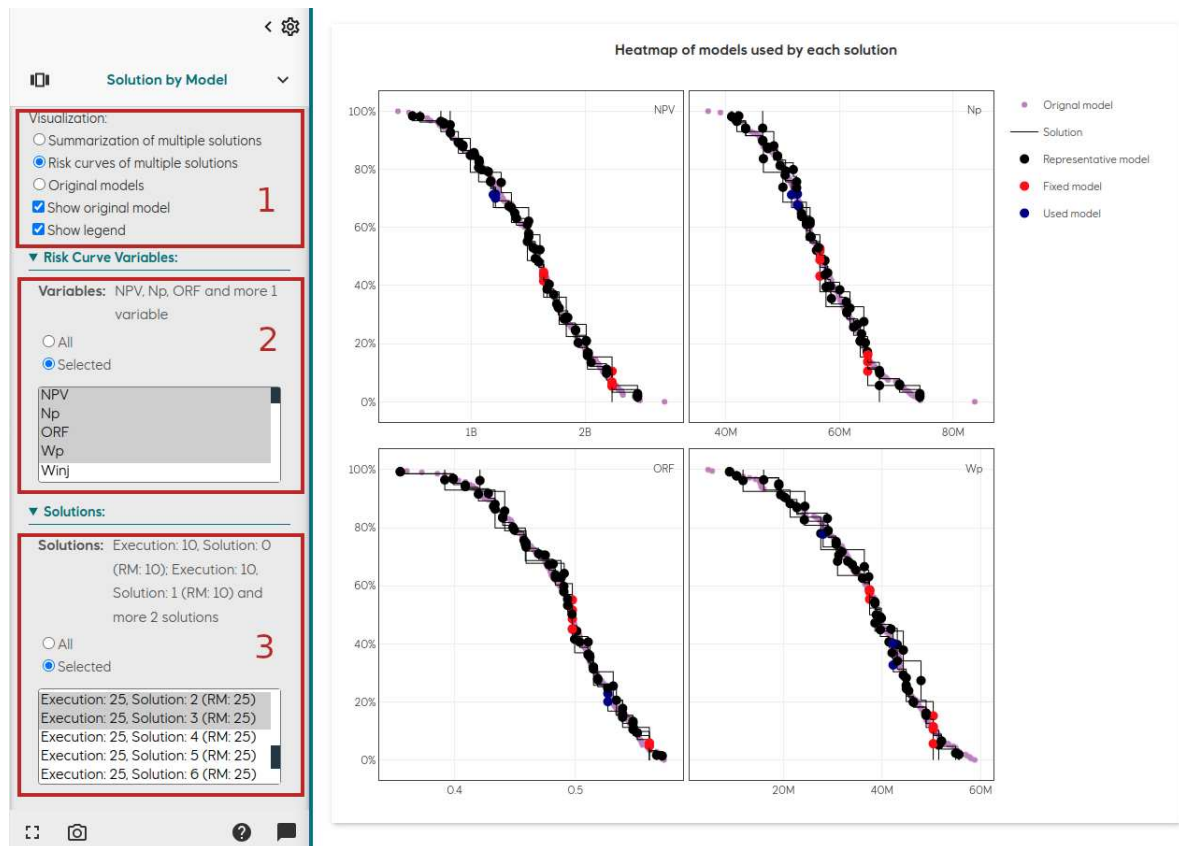


Figura 6.12: Recursos interativos do gráfico curvas de risco justapostas e sobrepostas, usando dados do caso de estudo UNISIM-I-D (GASPAR et al., 2015). Os parâmetros assinalados pelo retângulo vermelho número “1” e os filtros estão destacados pelos retângulos vermelho números “2” e “3”.

Fonte: Dados originais da pesquisa.

Tal qual descrito no gráfico curvas de risco sumarizadas por envoltório, o usuário pode selecionar a quantidade de gráficos de curvas de risco (uma para cada variável de saída) a ser exibida de forma justaposta na visualização e/ou filtrar as soluções (ou curvas de risco) a serem sobrepostas, por meio dos recursos interativos disponíveis nos retângulos vermelho número “2” e “3” na Figura 6.12, respectivamente. O usuário também tem à disposição outros recursos interativos no retângulo vermelho “1” na Figura 6.12; são eles: mostrar (ou ocultar) a curva de risco de todos os modelos (pontos roxos); mostrar (ou ocultar) a legenda; e selecionar o tipo de visualização a ser exibida, nesse caso, “*Risk curves of multiple solutions*”.

Como exemplo, na Figura 6.12 observa-se a justaposição de quatro variáveis de saída (*NPV*, *Np*, *ORF* e *Wp*). Foi aplicado um filtro para inspecionar duas soluções sobrepostas com 25 MRs cada. As curvas de risco dessas soluções sobrepõem ao longo de todo o seu domínio à

curva de risco de todos os modelos (com exceção dos modelos nas extremidades das curvas de risco das variáveis de saída NPV , Np , e Wp). Um aspecto visual, para ajudar a entender essa sobreposição, é a oclusão da curva de risco de todos os modelos (ou os pontos roxos). Sendo assim, considerando as quatro de variáveis de saída analisadas, a “representatividade baseada em curvas de risco” dessas duas soluções pode ser considerada adequada (i. e., quanto mais próximas da curva de risco de todos os modelos, melhor a representatividade). Por outro lado, as duas soluções filtradas são relativamente próximas entre si; no entanto, apresentam algumas diferenças, p. ex.: diferença do percentual acumulado (eixo Y) dos modelos assinalados na cor vermelha e azul, principalmente nas curvas de risco das variáveis de saída Np , ORF e Wp ; e as linhas pretas de cada solução não se sobrepõem ao longo de todo o domínio, apresentando diferenças acentuadas nos tracejados das curvas de risco das variáveis de saída Np e Wp . Os recursos interativos mencionados (filtros, abstração e elaboração, destacados na Seção 2.1.5) ajudaram a guiar essa análise, ao serem aplicados na visão geral das curvas de risco justapostas e sobrepostas (Figura 6.13) do caso de estudo UNISIM-I-D (GASPAR et al., 2015), inicialmente compreendendo 231 soluções, e 29 variáveis de saída. Além disso, os recursos interativos implementados nas curvas de risco justapostas e sobrepostas podem guiar descobertas ocultas nos gráficos estáticos produzidos pelo RMFinder (Figuras 2.19, 2.20 e 2.21).

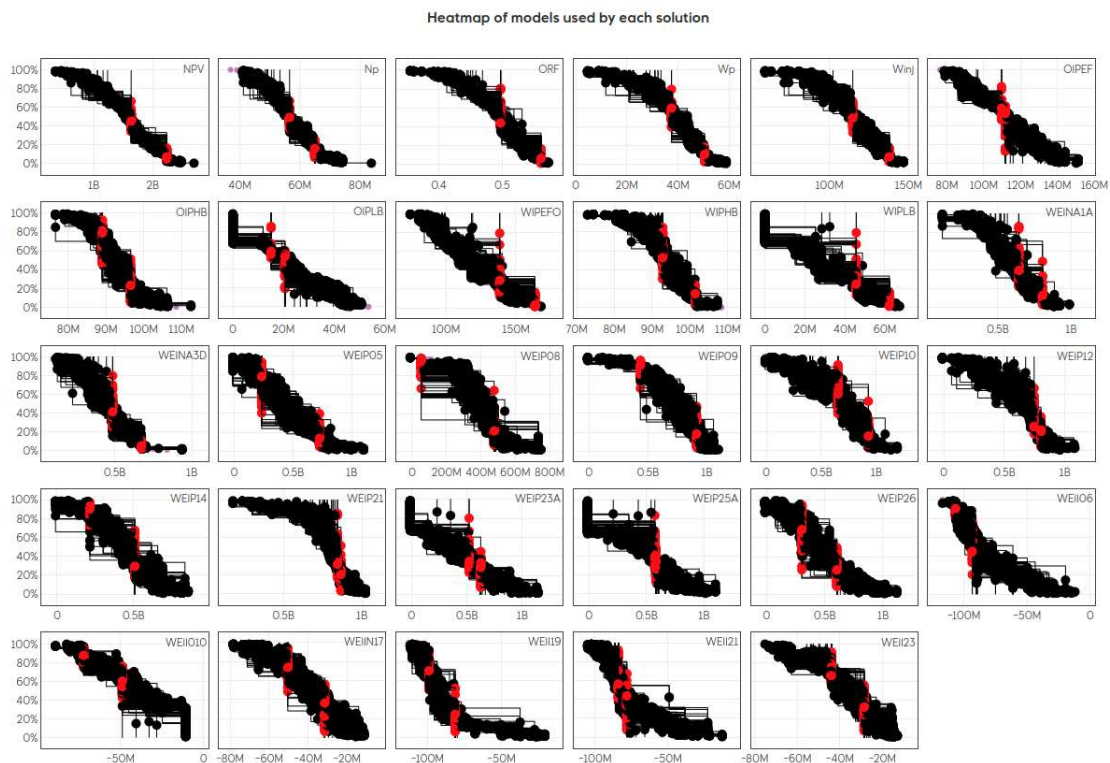


Figura 6.13: Visão geral das curvas de risco de 231 soluções, e 29 variáveis de saída justapostas e sobrepostas, usando dados do caso de estudo UNISIM-I-D (GASPAR et al., 2015).
Fonte: Dados originais da pesquisa.

Conforme mencionado na Seção 5.2, as curvas de risco sumarizadas por envoltório podem ser uma alternativa às curvas de risco justapostas e sobrepostas. Na Figura 6.14 é ilustrada a visão geral das curvas de risco sumarizadas por envoltório, sendo uma alternativa à abordagem híbrida de justaposição e sobreposição da Figura 6.13, usando dados do caso de estudo UNISIM-I-D (GASPAR et al., 2015) em ambos os gráficos. Embora na Figura 6.13 seja possível visualizar as faixas de confiança dos conjuntos de curvas de risco de várias soluções, os envoltórios da Figura 6.14 podem ajudar a identificar a sumarização das curvas de risco ao longo de todo o seu domínio sem oclusão do conjunto de todos os modelos. No entanto, o gráfico de curvas de risco justapostas e sobrepostas pode ser utilizado como um detalhe sob demanda (SHNEIDERMAN, 1996) para investigar as características de cada solução, p. ex.: comparar o tracejado das curvas de risco entre soluções; identificar os MRs utilizados em cada solução e como estão dispostos nas variáveis de saída; analisar os MRs destacados como fixos; e comparar as curvas de risco das soluções com a curva de todos os modelos.

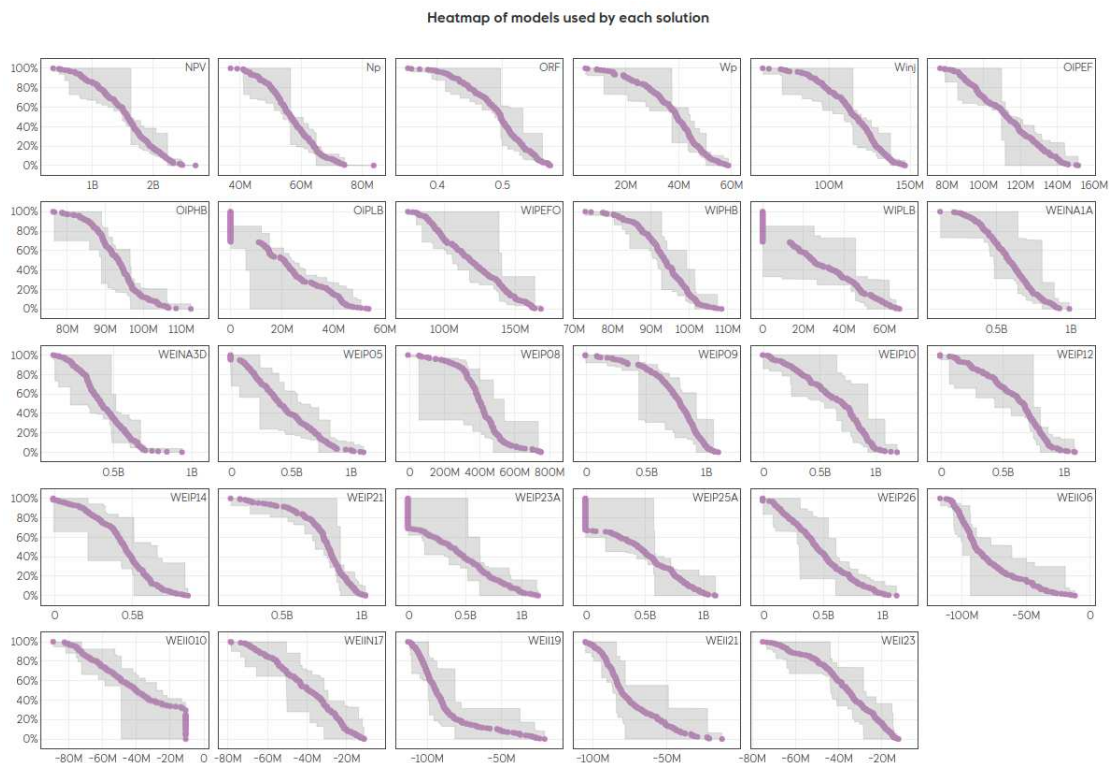


Figura 6.14: Visão geral das curvas de risco sumarizadas por envoltório como alternativa à abordagem híbrida de justaposição e sobreposição, usando dados do caso de estudo UNISIM-I-D (GASPAR et al., 2015).
Fonte: Dados originais da pesquisa.

Por fim, para auxiliar o usuário a identificar os valores das marcas no gráfico de curvas de risco justapostas e sobrepostas, adicionou-se outros recursos interativos (Figura 6.15) ao posicionar ou clicar o ponteiro do mouse nelas; são eles:

- (i) texto em destaque na Figura 6.15a com os valores do modelo (variável de saída e percentual acumulado) ao posicionar o ponteiro do *mouse* na marca da curva de risco de todos os modelos;
- (ii) texto em destaque na Figura 6.15b com os valores do modelo (variável de saída e percentual acumulado) ao posicionar o ponteiro do mouse no ponto (ou MR) da curva de risco de uma solução. A cor da caixa de texto corresponde à cor do MR onde o *mouse* é posicionado, respeitando os pontos assinalados na legenda da Figura 6.12;
- (iii) texto em destaque na Figura 6.15c com o detalhamento da execução (Seção 4.1.2) ao posicionar o ponteiro do *mouse* no gráfico de linhas na cor preta, no qual representa a curva de risco de uma solução;

- (iv) seleção de modelos (incluindo os MRs de uma solução) nas curvas de risco de uma variável de saída, ao clicar com o botão esquerdo do mouse (utilizando tecla SHIFT para multi-seleção), e destacá-los nas curvas de risco das outras variáveis de saída (recurso interativo de *brushing* e *linking*, destacado na Seção 2.1.5), conforme ilustrado na Figura 6.15d; e
- (v) seleção das curvas de risco justapostas e sobrepostas (Figura 6.15e) com a oclusão da curva de risco de todos os modelos (pontos roxos), desabilitada por meio da opção “Show original model” (retângulo vermelho “1” na Figura 6.12).

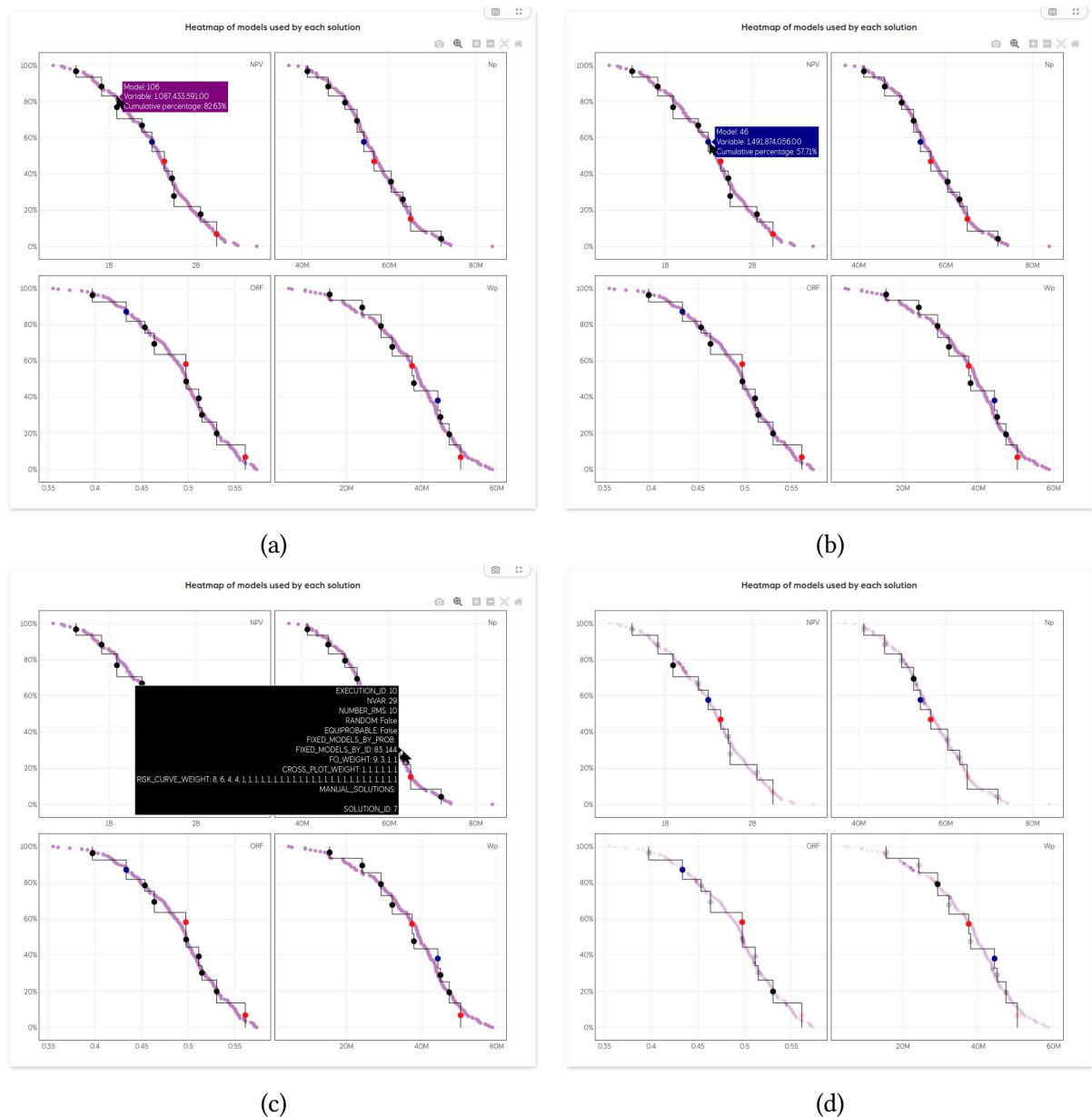
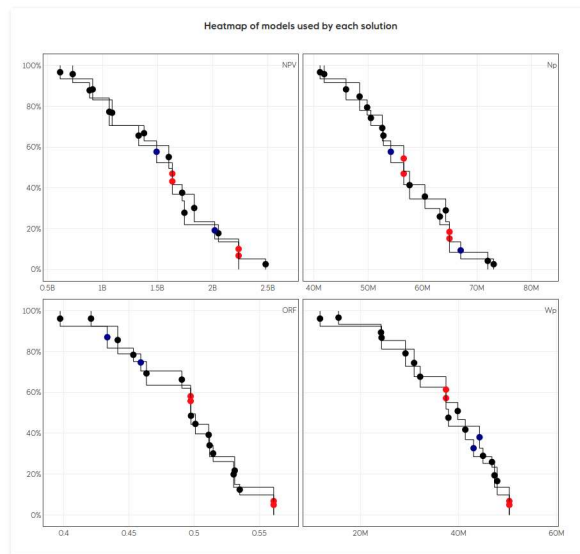


Figura 6.15: Recursos interativos do gráfico curvas de risco justapostas e sobrepostas: (a) o texto em destaque no retângulo roxo contém as informações do modelo na marca selecionada; (b) o texto em destaque no retângulo azul contém as informações do MR na marca selecionada; (c) o texto em destaque no retângulo preto contém as informações complementares da solução (linha preta) selecionada; e (d) seleção de modelos e MRs em uma curva de risco de uma variável de saída para destacá-los nas curvas de risco das demais variáveis de saída.

Fonte: Dados originais da pesquisa.



(e)

Figura 6.15: Recursos interativos do gráfico curvas de risco justapostas e sobrepostas: (e) visualização contendo apenas curva de risco.

Fonte: Dados originais da pesquisa.

6.1.7 Curvas de risco do conjunto de todos os modelos

Nos gráficos curvas de risco sumarizadas por envoltório e curvas de risco justapostas e sobrepostas, o usuário pode optar por visualizar somente a curva de risco de todos os modelos τ de cada variável de saída f_i , escondendo a região cinza do envoltório e as curvas de risco das soluções, por meio da opção “Original models” do retângulo vermelho “1” na Figura 6.16. Esse recurso interativo pode ajudar o usuário a analisar somente as curvas de risco de todos os modelos, sem a presença de outro mapeamento visual, para cada variável de saída (uma para cada curva de risco) a ser selecionada no retângulo vermelho “2” na Figura 6.16. Além disso, essa visualização pode ser considerada uma opção adicional implementada neste trabalho por conta dela não estar presente nos gráficos estáticos gerados pelo RMFinder (Seção 2.3.2).

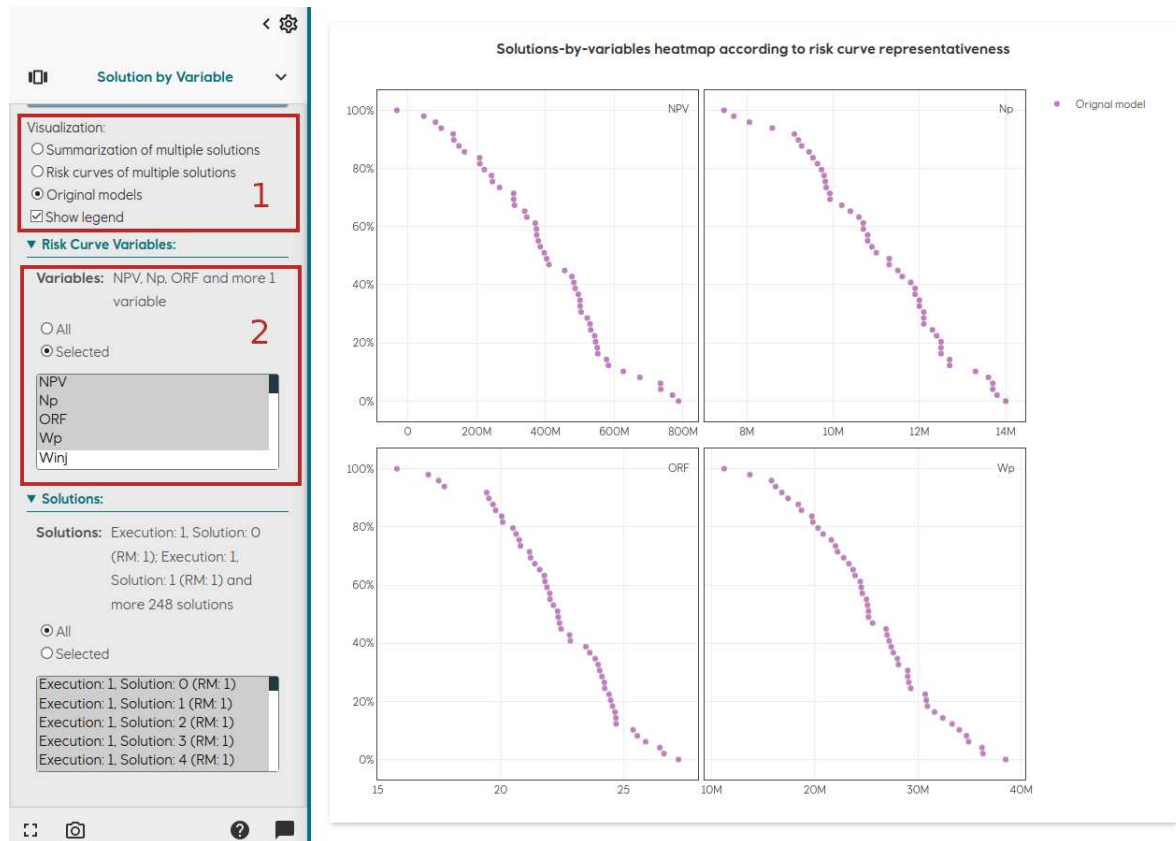


Figura 6.16: Recursos interativos do gráfico curvas de risco do conjunto de todos os modelos, usando dados do caso de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017), com parâmetros assinalados no retângulo vermelho número “1” e filtro destacado no retângulo vermelho número “2”.

Fonte: Dados originais da pesquisa.

Bem como nos gráficos curvas de risco sumarizadas por envoltório e curvas de risco justapostas e sobrepostas, adicionou-se outros recursos interativos (Figura 6.17) ao posicionar ou clicar o ponteiro do mouse nas marcas do gráfico curvas de risco do conjunto de todos os modelos; são eles:

- (i) texto em destaque na Figura 6.17a com os valores do modelo (variável de saída e percentual acumulado) ao posicionar o ponteiro do mouse na marca da curva de risco de todos os modelos; e
- (ii) seleção de modelos (uso da tecla SHIFT para multi-seleção) na curva de risco de uma variável de saída, destacando-os nas curvas de risco das outras variáveis de saída (recurso interativo de *brushing* e *linking*, assinalado na Seção 2.1.5) ao clicar com o ponteiro do mouse na marca do gráfico, conforme ilustrado na Figura 6.17b.

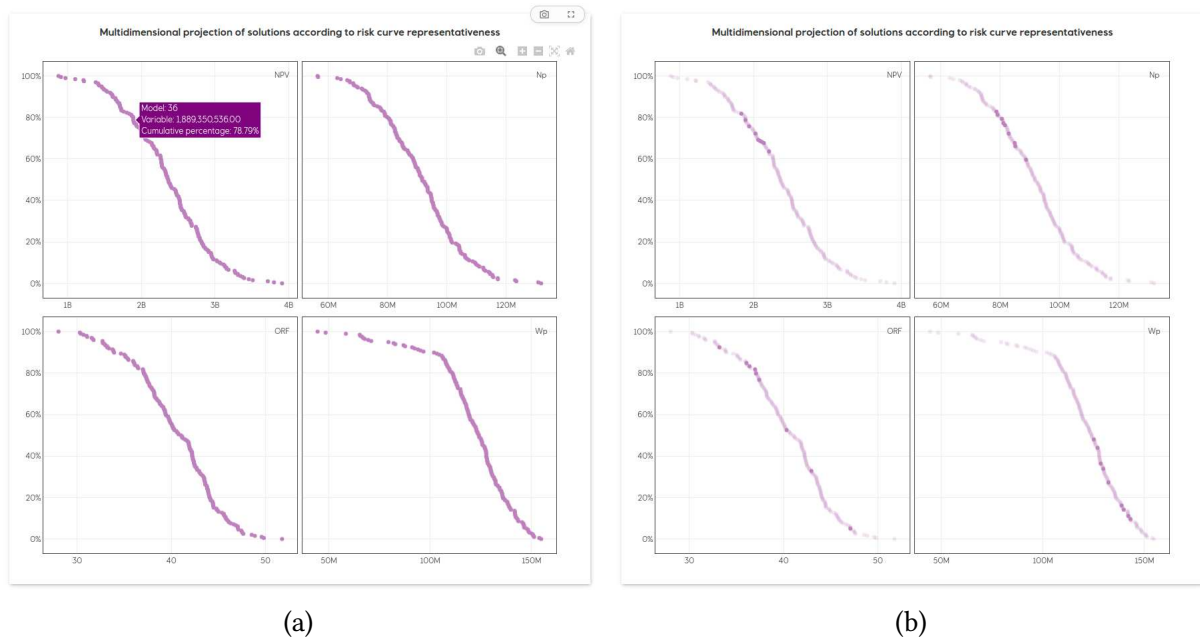


Figura 6.17: Recursos interativos do gráfico curvas de risco do conjunto de todos os modelos: (a) o texto em destaque no retângulo roxo contém as informações do modelo na marca selecionada; e (b) seleção de modelos em uma curva de risco de uma variável de saída para destacá-los nas outras variáveis de saída.

Fonte: Dados originais da pesquisa.

6.1.8 Outros recursos interativos

Ao passar o mouse em qualquer representação visual do sistema de visualização interativo, as barras de botões dos retângulos “1” e “2” na Figura 6.18 são exibidas no canto superior direito de cada gráfico. Somado a isso, está à disposição do usuário uma outra barra fixa de botões (retângulo “3” na Figura 6.18).

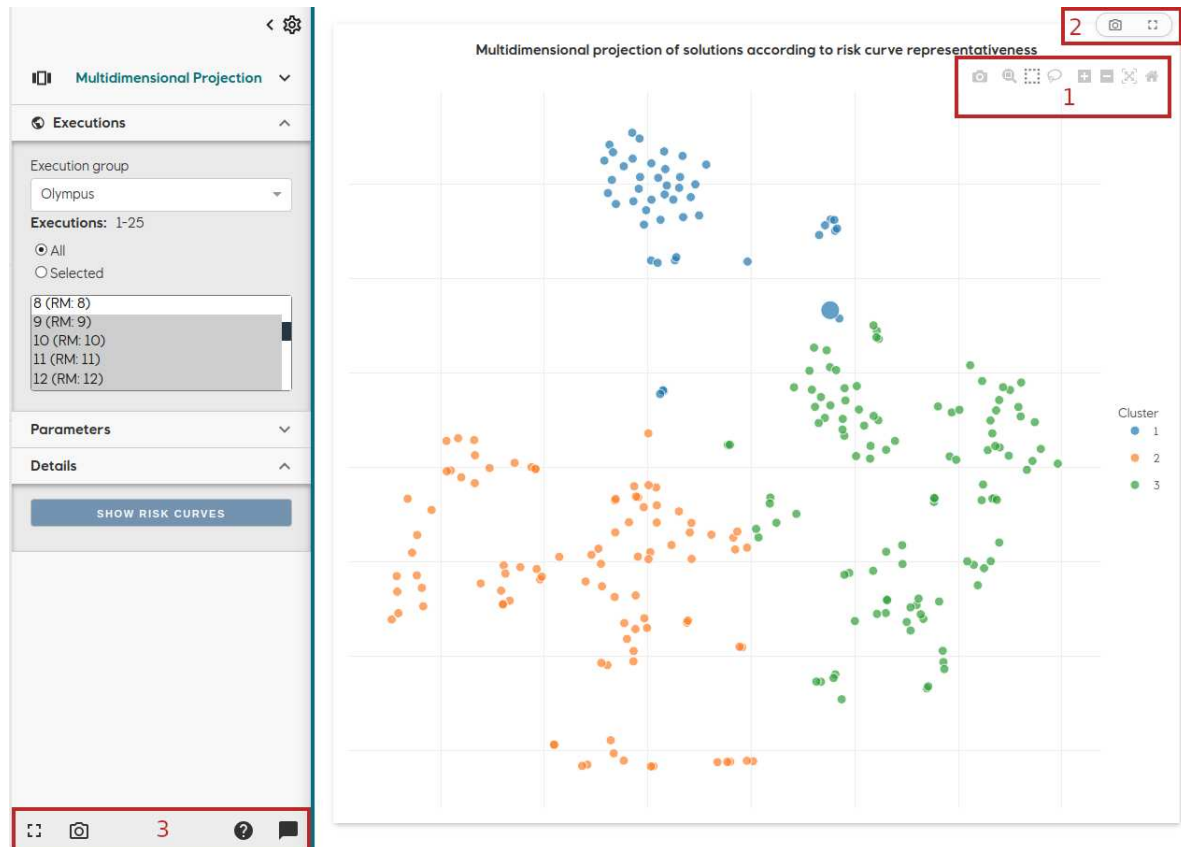


Figura 6.18: Barras de botões exibidas no sistema de visualização interativo.

Fonte: Dados originais da pesquisa.

Os botões do retângulo “1” na Figura 6.18, da esquerda à direita, realizam as seguintes ações, respectivamente:

1. baixar uma figura no formato *Portable Network Graphics* (PNG) de tamanho 700 por 450 *pixels* via navegador do usuário contendo uma versão estática do estado atual do gráfico;
2. permitir ao usuário aumentar e diminuir o *zoom* do gráfico usando a barra de rolagem do *mouse* e/ou a rolagem com dois dedos;
3. criar uma seleção retangular ou quadrada no gráfico (clicar, manter pressionado e arrastar o *mouse* para criar essa seleção);
4. criar uma seleção personalizada livremente (a borda da região selecionada seguirá o cursor do mouse à medida que ele for arrastado no gráfico e a forma geométrica será automaticamente fechada com uma linha reta do local atual do cursor até o ponto inicial);
5. aumentar o *zoom* do gráfico e redimensioná-lo usando como referência o centro dos pontos selecionados (ou todos os pontos, na ausência de uma seleção);

6. análogo ao item 5, diminuir o *zoom* do gráfico;
7. alterar o leiaute do gráfico para mostrar todos os pontos (i. e., redimensionar e ajustar automaticamente o gráfico); e
8. retornar às configurações de leiaute original atribuídas a cada gráfico.

Somando a isso, os botões do retângulo “2” na Figura 6.18, da esquerda à direita, realizam as seguintes ações, respectivamente:

1. capturar a tela da seção de um gráfico⁸ e, automaticamente, realizar o *download* dessa tela como uma imagem local em formato PNG no sistema de arquivos do usuário, respeitando o tamanho em *pixels* da seção exibida na tela do usuário. A seção do gráfico corresponde à versão estática do estado atual da representação visual somado ao seu respectivo título; e
2. acionar a opção de tela cheia, onde a seção do gráfico ocupa toda a área disponível no navegador do usuário, ocultando as barras dos recursos interativos. Após acionar a tela cheia, outro botão é exibido no canto superior direito (substituindo a opção de tela cheia) para retornar a seção do gráfico ao seu tamanho anterior e as barras dos recursos interativos.

Por fim, os botões do retângulo “3” na Figura 6.18, da esquerda à direita, realizam as seguintes ações, respectivamente:

1. os dois primeiros botões realizam as funcionalidades de: acionar a opção de tela cheia e capturar a tela da seção de um gráfico, descritas nos itens 2 e 1 do retângulo “2” na Figura 6.18, respectivamente;
2. iniciar um passeio pelos recursos interativos e representações visuais do sistema de visualização. Ao clicar nesse botão, uma barra de navegação é exibida (retângulo vermelho na Figura 6.19) contendo um botão “*Next*” no canto direito para avançar um passo no passeio, um botão “*Back*” no canto esquerdo para retornar um passo no passeio e, no meio, uma barra de progresso indicando o passo atual na navegação do

⁸Tecnicamente, o termo “seção de um gráfico” corresponde a uma *tag* HTML `<div>`, na qual cria um bloco genérico para agrupar conteúdos. A função dessa *tag* é separar os conteúdos do código HTML em divisões de espaço (ou conjunto), sem qualquer peso semântico.

passeio. Os recursos interativos ou representações visuais, relacionados ao passo do passeio, são destacados no sistema de visualização acompanhado de uma descrição, conforme exemplificado na Figura 6.19 com o gráfico *heatmap* reordenável de variáveis por soluções; e

3. reportar problemas no sistema de visualização interativo. Os problemas podem ser usados para manter o controle de erros, aprimoramentos ou outras solicitações.

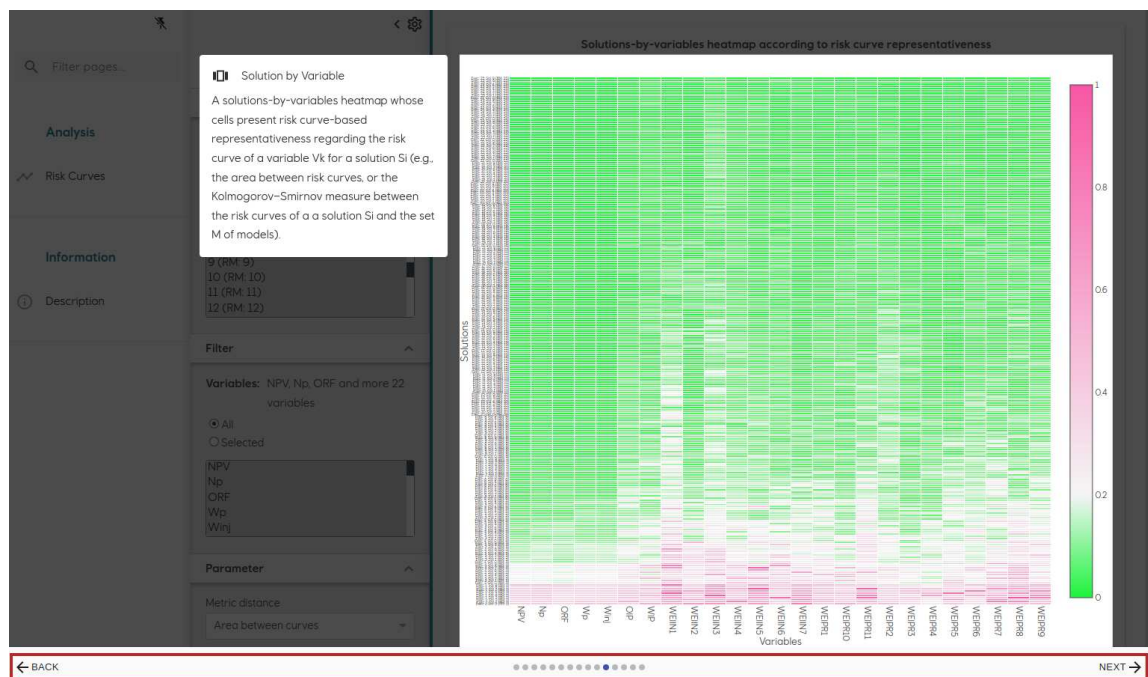


Figura 6.19: Passeio pelos recursos interativos e representações visuais do sistema de visualização.

Fonte: Dados originais da pesquisa.

6.2 Exemplos de caso de uso

De acordo com Ward, Grinstein e Keim (2015), ao invés de usar participantes voluntários para avaliar o sistema de visualização, alguns pesquisadores de InfoVis tentam demonstrar exemplos reais (ou, às vezes, artificiais) de como seu método pode ser aplicado na solução de um problema específico ou na execução de uma tarefa específica. Ainda segundo os mesmos autores, é crucial garantir que os casos de uso sejam suficientemente realistas para que alguém com uma tarefa específica tenha a convicção de que, pelo menos, um dos casos de uso é bastante similar à sua própria tarefa e que o sistema de informação será capaz de fornecer suporte eficiente.

Nesse sentido, com o intuito de explorar o sistema de visualização interativo de forma integrada, utilizou-se como caso de uso os dados do caso de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017) para selecionar uma solução com base na representatividade em curvas de risco. Essa solução deve conter um conjunto de até nove MRs, de acordo com uma premissa estabelecida para a etapa 8 da metodologia de Schiozer et al. (2019) (Seção 2.2.1) de um projeto específico conduzido por um pesquisador. Essa premissa de nove MRs pode ser motivada por um requisito de pesquisa (p.ex., os atributos incertos dos cenários estarem discretizados em três níveis — conceito de atributo-nível descrito na Seção 2.2 — com até três MRs representando cada nível, totalizando até nove MRs) e/ou uma restrição de recurso computacional em passos de simulação seguintes ao da escolha de MRs.

O conjunto de dados do caso de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017) contém o resultado de 25 execuções do RMFinder (variando de 1 a 25 MRs, sendo a primeira execução composta por 1 MR, a segunda por 2 MRs e assim por diante, até a última execução de número 25, contendo 25 MRs); cada execução tem 10 soluções, totalizando 250 soluções no conjunto de dados; 25 variáveis de saída; 50 cenários (ou modelos); e 251 curvas de risco (sendo uma delas referente ao conjunto de todos os modelos, e as demais referentes a 250 soluções).

Primeiramente, para atender a premissa de selecionar uma solução com um conjunto de até nove MRs dos dados do caso de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017), foram aplicados dois filtros no sistema de visualização interativo; são eles: grupo de execução igual a “OLYMPUS” e as execuções dentro do intervalo de um a nove (i. e., soluções de um a nove MRs), conforme ilustrado nos retângulos “1” e “2” na Figura 6.20, respectivamente. Esses filtros de caráter mais abrangente estão presentes em todas as visualizações a serem analisadas a seguir nesta seção, retornando 90 soluções.



Figura 6.20: Filtros para selecionar soluções com conjuntos de até nove MRs dos dados do caso de estudo OLYMPUS.

Fonte: Dados originais da pesquisa.

Em seguida, realizou-se uma análise nos gráficos de visão geral contidos no sistema de visualização interativo, obedecendo a seguinte ordem: *heatmap* reordenável de variáveis por soluções, *heatmap* reordenável de modelos utilizados por solução e projeção multidimensional de curvas de risco.

Na Figura 6.21 é ilustrado o gráfico *heatmap* de variáveis por soluções da medida de distância “área entre curvas de risco”, que deve ter o menor valor possível, com a matriz reordenada. A medida de distância e a reordenação de matriz foram parametrizadas por meio das opções disponíveis no sistema de visualização (retângulo vermelho “3” na Figura 6.4). Nesse gráfico, é possível notar que aproximadamente um terço das soluções (retângulo vermelho na Figura 6.21) estão com valores baixos (células verdes) para quase todas as variáveis de saída, com exceção da variável de saída *WEIN1* (índice econômico do poço injetor número um). Contudo, aproximadamente metade das soluções assinaladas dentro do retângulo vermelho na Figura 6.21 estão com valores baixos (células verdes) na variável de saída *WEIN1*. Nesse sentido, essa visualização permitiu identificar a existência de soluções com boa “representatividade baseada em curvas de risco” em relação a todas as variáveis de saída, bem como constatar que a variável de saída *WEIN1* tem um erro maior ao comparar soluções que variam de um a nove MRs.

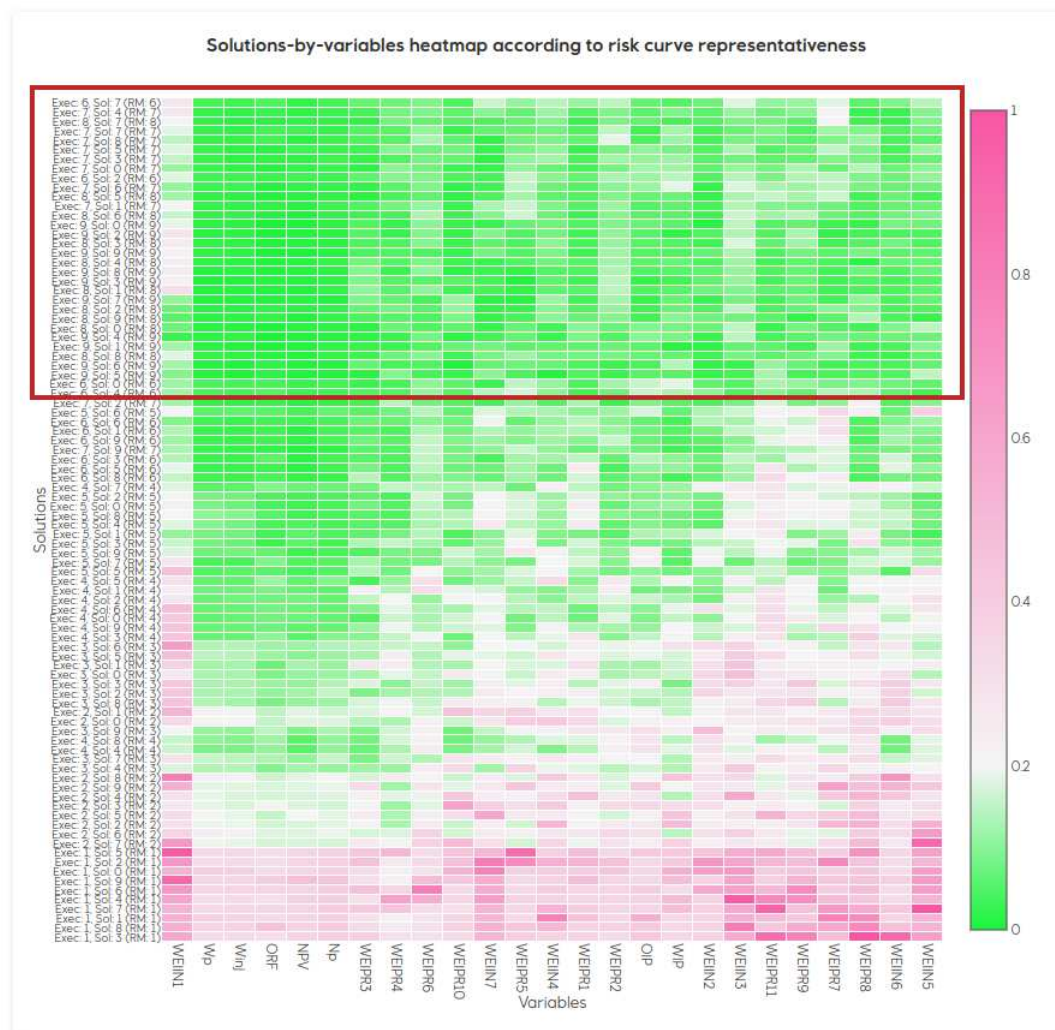


Figura 6.21: Caso de uso do *heatmap* reordenável de variáveis por soluções, com base nos dados do caso de estudo OLYMPUS, filtrados pelas execuções de um a nove.

Fonte: Dados originais da pesquisa.

Em seguida, foram analisadas as 90 soluções filtradas (Figura 6.20) em relação à sua variabilidade, como subconjuntos comuns de MRs entre elas, usando o *heatmap* de modelos utilizados por solução (como ilustrado na Figura 6.22). A matriz na Figura 6.22 está reordenada (recurso interativo disponível nessa representação visual no sistema de visualização, conforme indicado no retângulo vermelho “3” na Figura 6.5) e, após essa reordenação, é possível visualizar que cinco modelos (47, 2, 16, 3 e 11, obedecendo essa ordenação na Figura 6.22, da esquerda à direita do eixo X) estão presentes no conjunto de MRs de aproximadamente dois terços das soluções selecionadas.

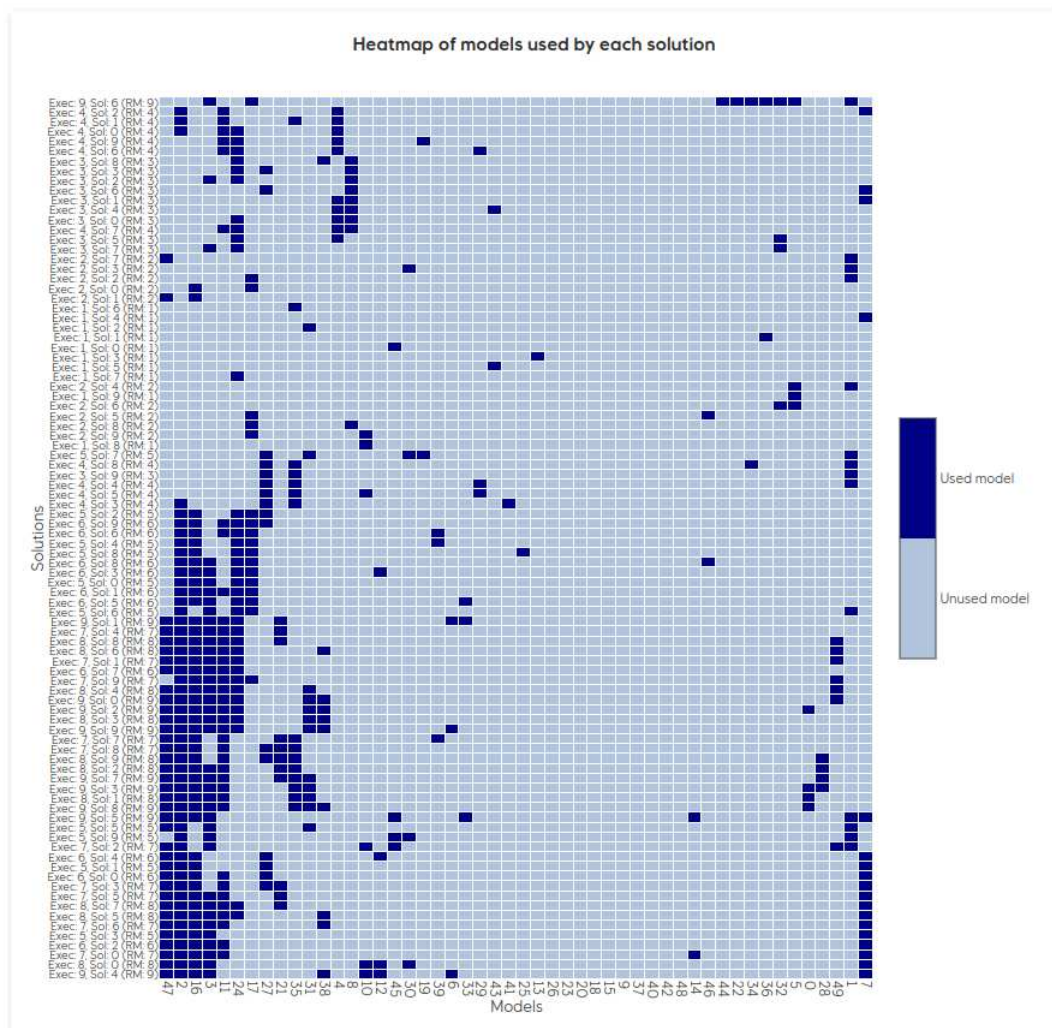


Figura 6.22: Caso de uso do *heatmap* de modelos utilizados por solução, com base nos dados do caso de estudo OLYMPUS, filtrados pelas execuções de um a nove.

Fonte: Dados originais da pesquisa.

Além disso, após selecionar os modelos não utilizados em nenhuma solução (ilustrado na Figura 6.23a), observa-se que somente 10 dos 50 modelos (i. e., 20% dos modelos) não foram utilizados em nenhuma solução (Figura 6.23b). Sendo assim, é possível observar uma variedade de modelos selecionados para o problema de otimização do RMFinder ao examinar as 90 soluções filtradas (Figura 6.20) nos dados do caso de estudo OLYMPUS; no entanto, os 10 modelos não selecionados podem estar bem representados por aqueles selecionados pelo RMFinder.

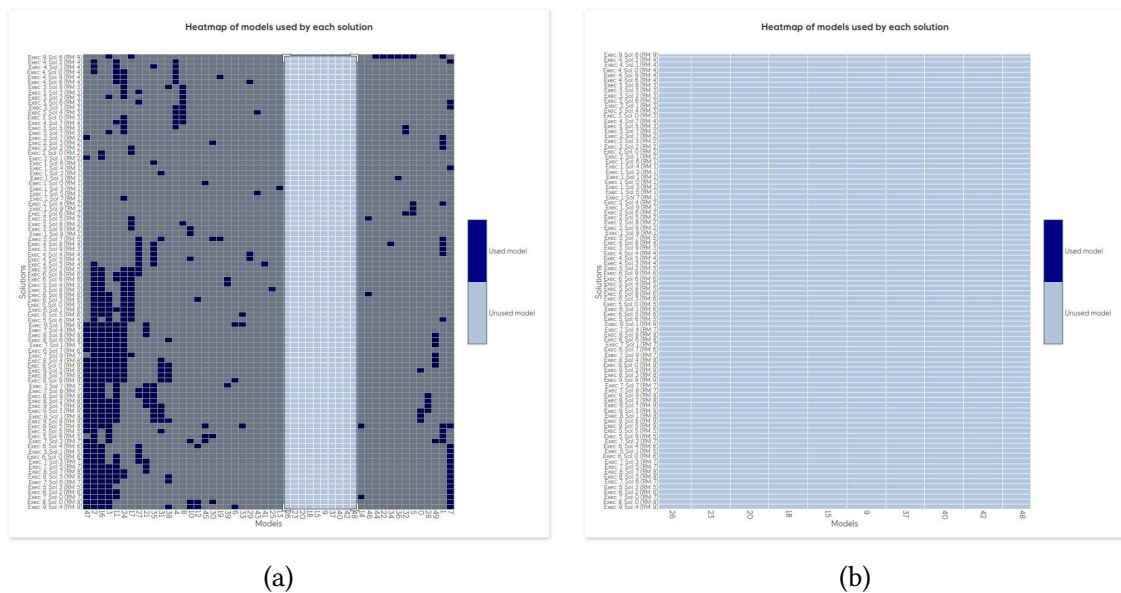


Figura 6.23: Seleção das células do *heatmap* considerando somente os modelos não utilizados em nenhuma solução, usando dados do caso de estudo OLYMPUS: (a) seleção retangular nas células do *heatmap*, ao clicar, manter pressionado e arrastar o mouse; e (b) células selecionadas no *heatmap*, após soltar o clique do mouse.

Fonte: Dados originais da pesquisa.

Em resumo, após as análises descritas até aqui nesta seção, pode-se afirmar que as 90 soluções (com até 9 MRs), filtradas (Figura 6.20) nos dados do caso de estudo OLYMPUS, apresentam as seguintes características:

- (i) aproximadamente um terço das soluções apresenta uma boa “representatividade baseada em curvas de risco” em relação a todas as variáveis de saída;
- (ii) o problema de otimização do RMFinder demonstrou variabilidade ao selecionar 80% dos modelos como representativos; e
- (iii) cinco modelos (números 47, 2, 16, 3 e 11) estiveram presentes em aproximadamente dois terços das soluções, demonstrando ser um subconjunto comum de MRs.

Para prosseguir com esse processo de análise, utilizou-se o gráfico de projeção multidimensional de curvas de risco para identificar a similaridade entre soluções e conjunto de todos os modelos. Na Figura 6.24, cada ponto corresponde a uma das 90 soluções filtradas (Figura 6.23b) nos dados do caso de estudo OLYMPUS, e o ponto com maior tamanho representa o conjunto de todos os modelos. Observa-se ainda a formação de dois grupos (*clusters*) de curvas de risco. No grupo com a cor azul, no canto inferior esquerdo da projeção,

encontra-se a curva de risco do conjunto de todos os modelos. Essa projeção multidimensional foi parametrizada com a medida de distância “área entre curvas de risco”, considerando simultaneamente todas as 25 variáveis de saída (configurações realizadas por meio dos parâmetros assinalados no retângulo vermelho “2” na Figura 6.3).

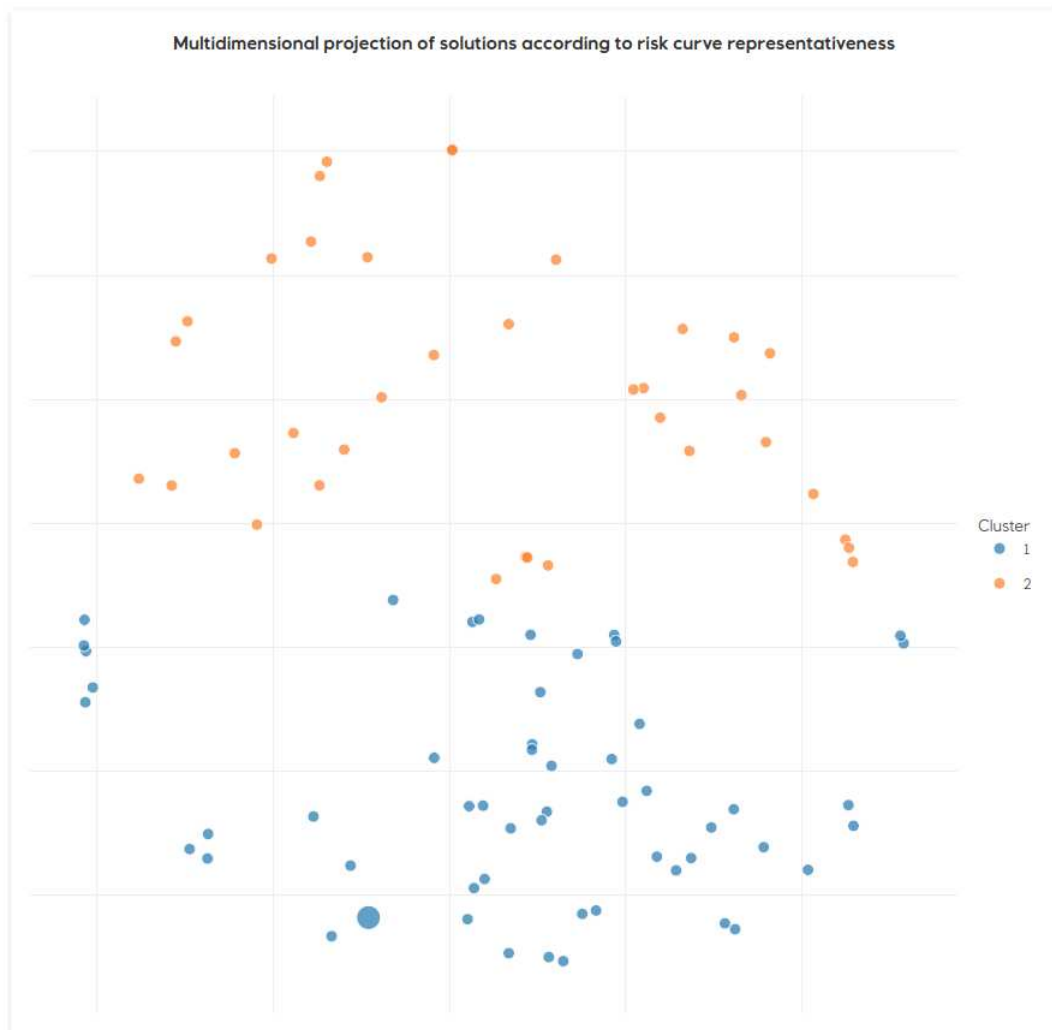


Figura 6.24: Caso de uso do gráfico de projeção multidimensional de curvas de risco, com base nos dados do caso de estudo OLYMPUS, filtrados pelas execuções de um a nove.

Fonte: Dados originais da pesquisa.

Com o objetivo de inspecionar potenciais soluções com boa “representatividade baseada em curvas de risco”, realizou-se uma seleção no gráfico de projeção multidimensional, compreendendo as três soluções mais próximas ao conjunto de todos os modelos, conforme destacado na Figura 6.25. Após realizar essa seleção no gráfico de projeção multidimensional, clicou-se no botão *Show Risk Curve* (ilustrado no retângulo vermelho “2” na Figura 6.2) para

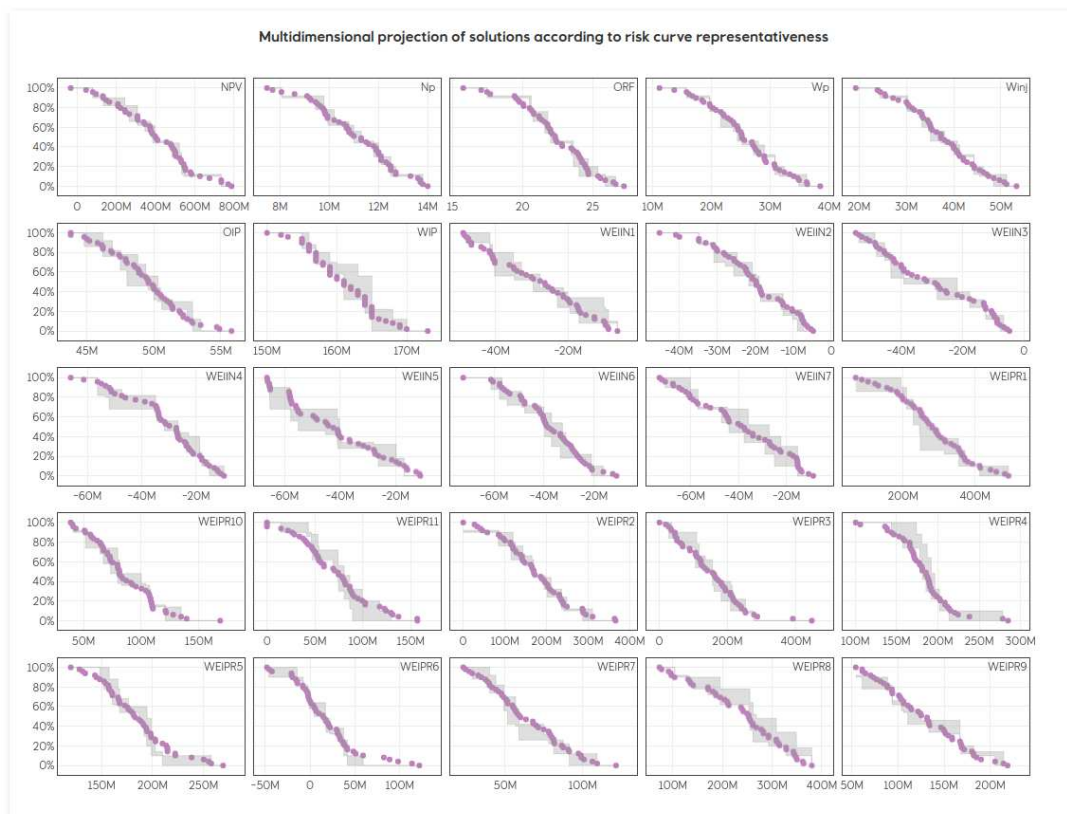
acionar o detalhamento das curvas de risco, primeiramente, com o gráfico curvas de risco sumarizadas por envoltório.



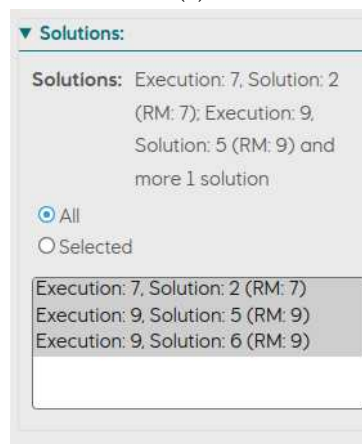
Figura 6.25: Seleção de soluções no gráfico projeção multidimensional de curvas de risco.
Fonte: Dados originais da pesquisa.

Na Figura 6.26a é apresentado o gráfico curvas de risco sumarizadas por envoltório, contemplando as três soluções selecionadas na Figura 6.25. Como ilustrado na Figura 6.26b, duas delas pertencem à execução 9, tendo um conjunto de nove MRs cada, identificadas por 2 e 5. Enquanto a outra solução, com sete MRs, pertencente à execução 7, tem identificação 6. A matriz de gráficos de curva de risco é composta por 25 células, sendo uma para cada variável de saída. Visualmente, os envoltórios (sombras cinza) das variáveis de saída na primeira linha da matriz de gráficos (*NPV*, *Np*, *ORF*, *Wp* e *Winj*) apresentam um contorno constante, envolvendo a curva de risco do conjunto de todos os modelos ao longo de todo o seu domínio. Além disso, a diferença entre os limites superiores e inferiores (i. e., a faixa de

confiança) dessas variáveis de saída pode ser considerada pequena; portanto, pode-se afirmar que as três soluções selecionadas na Figura 6.25 provém uma “boa representatividade baseada em curvas de risco” para essas variáveis de saída. Contudo, essa afirmação seria inválida para a variável de saída *WEIPR4* (índice econômico do poço produtor número quatro) localizada na última coluna da linha quatro, por exemplo. Nota-se que o envoltório dessa variável de saída não reveste totalmente a curva de risco do conjunto de todos os modelos ao longo de todo o seu domínio. Em outras palavras, as curvas de risco das três soluções selecionadas na Figura 6.25 consideram que este poço produtor terá mais produção do que os modelos originais previam.



(a)



(b)

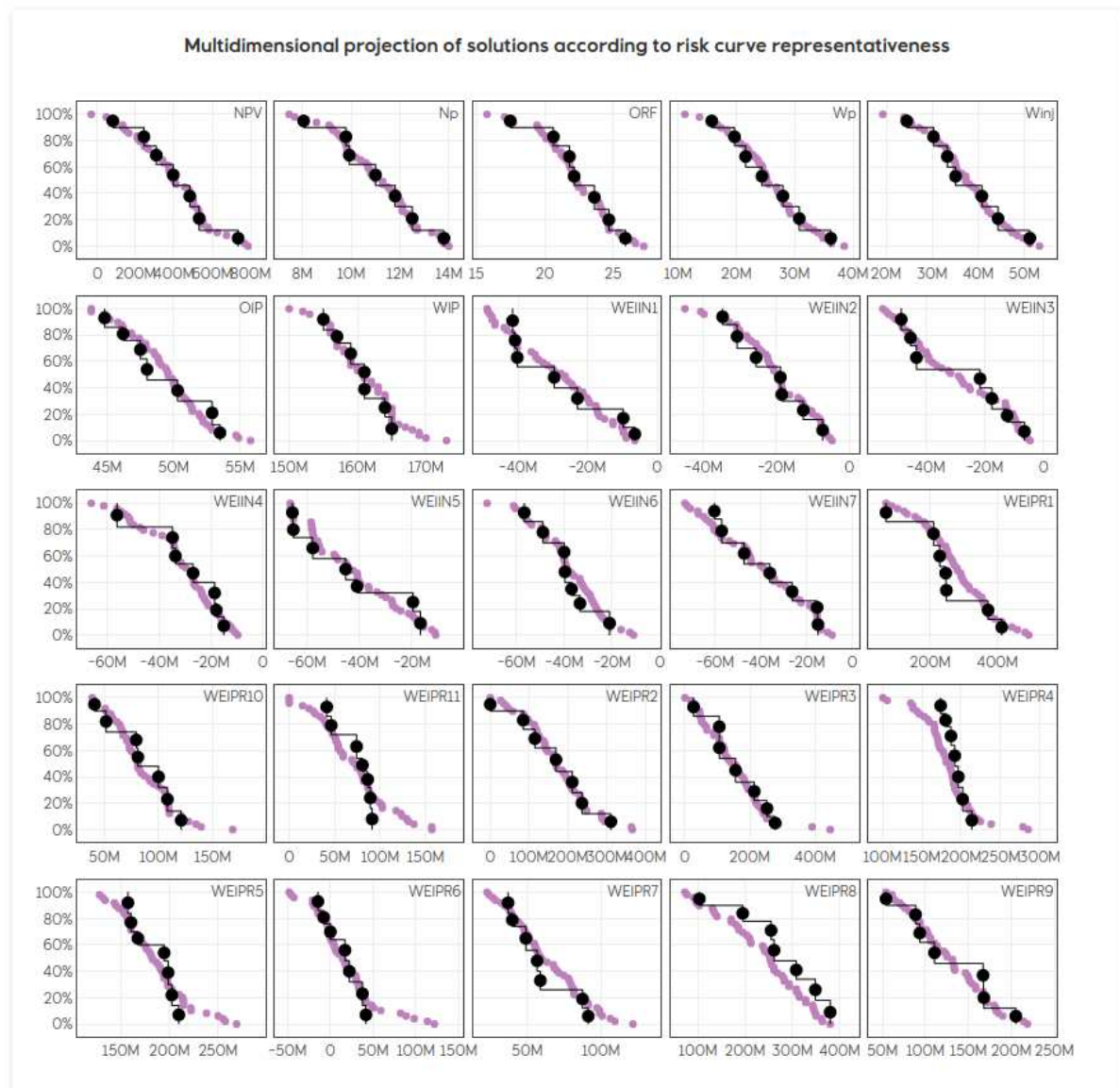
Figura 6.26: Caso de uso do gráfico curvas de risco sumarizadas por envoltório das três soluções selecionadas na Figura 6.25, usando dados do caso de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017): (a) matriz dos gráficos de curva de risco com 25 células, sendo uma para cada variável de saída; e (b) as três soluções selecionadas na Figura 6.25, ilustrado por meio do filtro de solução (retângulo “3” nas Figuras 6.7 e 6.12).

Fonte: Dados originais da pesquisa.

A próxima etapa dessa análise foi a verificação do gráfico curvas de risco justapostas e sobrepostas de cada uma das três soluções selecionadas na Figura 6.25, assinaladas na Figura 6.26b como solução 2 da execução 7, solução 5 da execução 9 e solução 6 da execução

9. As curvas de risco das 25 variáveis de saída de cada uma dessas soluções são apresentadas nas Figuras 6.27a, 6.27b e 6.27c, respectivamente, por meio de uma matriz de gráficos.

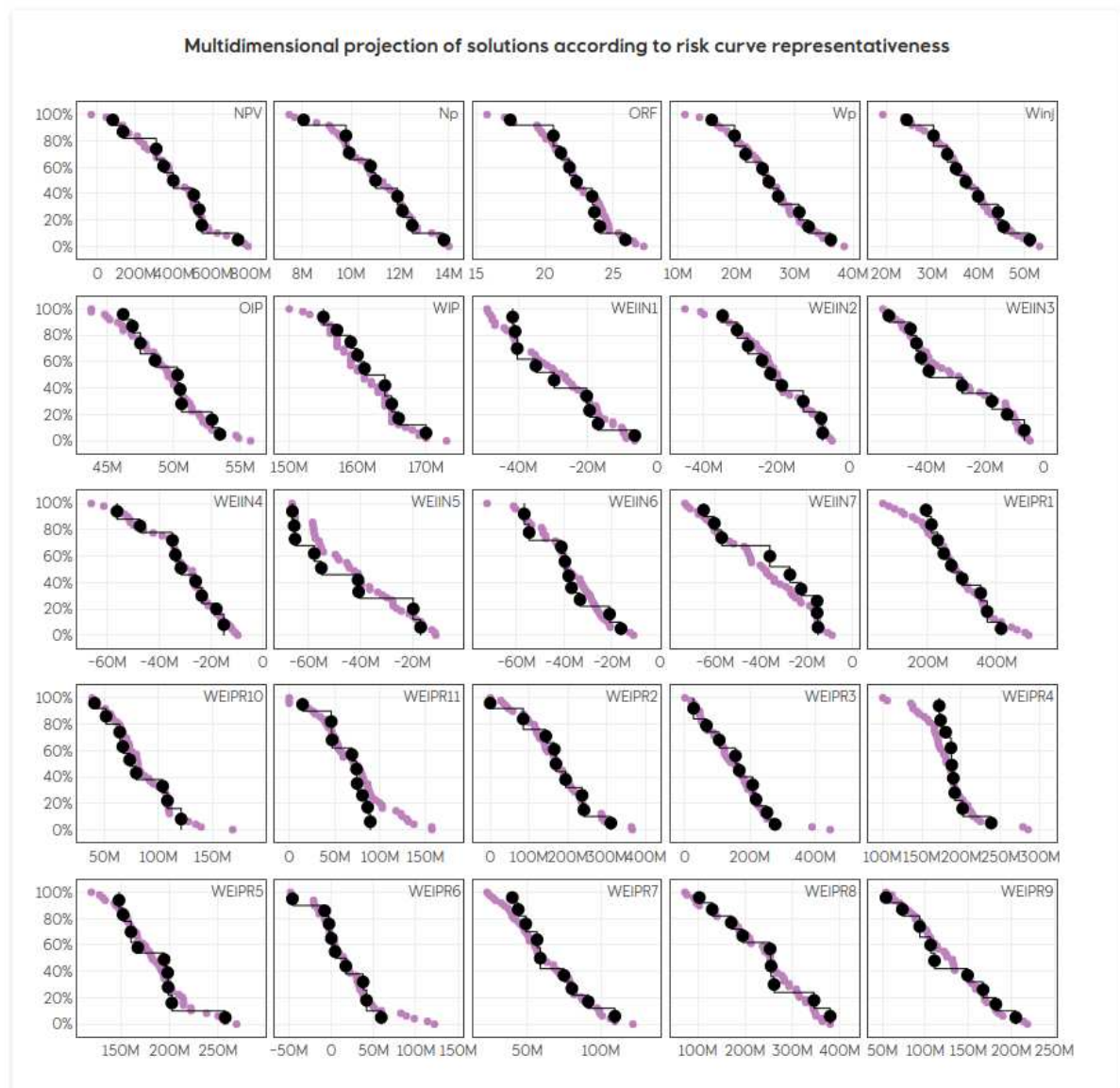
Na Figura 6.27a, é possível visualizar que a curva de risco do conjunto de todos os modelos da variável *WEIPR4* (a mesma destacada nesta seção anteriormente ao analisar as curvas de risco sumarizadas por envoltório) não está sendo sobreposta pelos pontos pretos e linhas pretas da solução 2 da execução 7, que tem um conjunto de sete MRs. Dessa forma, a curva de risco dessa solução não é próxima à curva de risco de todos os modelos para essa variável de saída, ou seja, não há uma representatividade adequada. Essa afirmação pode ser estendida à curva de risco da variável de saída *WEIPR4* da solução 5 da execução 9 (com nove MRs), conforme demonstrado na Figura 6.27b. Essa solução também não apresenta uma representatividade adequada para essa variável de saída. No entanto, na Figura 6.27c, nota-se um comportamento diferente: a curva de risco de todos os modelos da variável de saída *WEIPR4* (pontos roxos) é parcialmente sobreposta pelos pontos e linha preta da solução 6 da execução 9 (com nove MRs), o que torna essa solução mais representativa em relação às outras para essa variável de saída. Essa comparação, associada à mesma conclusão, pode ser constatada também ao analisar as variáveis de saída *WEIPR8* e *WEIPR11* (índice econômico dos poços produtores número 8 e 11), por exemplo.



(a)

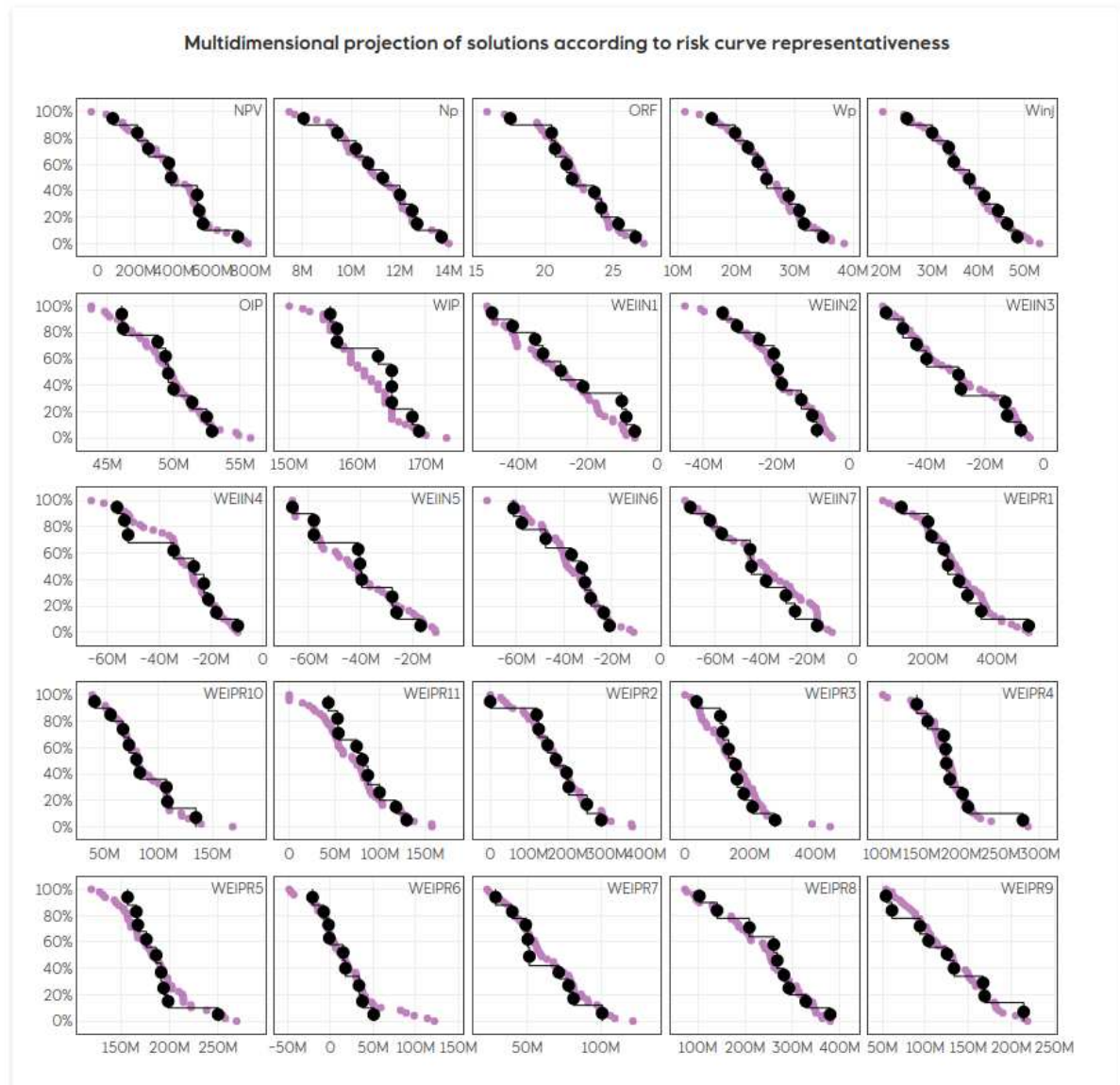
Figura 6.27: Caso de uso do gráfico curvas de risco justapostas com curvas de risco sobrepostas, usando dados do caso de estudo OLYMPUS: (a) solução 2 da execução 7, tendo um conjunto de sete MRs).

Fonte: Dados originais da pesquisa.



(b)

Figura 6.27: Caso de uso do gráfico curvas de risco justapostas e sobrepostas, usando dados do caso de estudo OLYMPUS: (b) solução 5 da execução 9, tendo um conjunto de nove MRs. Fonte: Dados originais da pesquisa.



(c)

Figura 6.27: Caso de uso do gráfico curvas de risco justapostas e sobrepostas, usando dados do caso de estudo OLYMPUS: (c) solução 6 da execução 9, tendo um conjunto de nove MRs. Fonte: Dados originais da pesquisa.

Sendo assim, dentre as 90 soluções analisadas dos dados do caso de estudo OLYMPUS, respeitando a premissa de considerar somente as execuções com um conjunto de até nove MRs, após uma análise das representações visuais de visão geral e detalhes sob demanda, pode-se afirmar que a solução 6 da execução 9, com um conjunto de nove MRs, apresentou a melhor “representatividade baseada em curvas de risco” em relação a todas as variáveis de saída em comparação com as outras soluções analisadas. Essa afirmação pode ser reforçada ao destacá-la nos gráficos de visão geral.

No gráfico projeção multidimensional de curvas de risco, a solução 6 da execução 9 (quadrado vermelho na Figura 6.28a) é a mais próxima da curva de risco do conjunto de modelos originais (marca de maior tamanho na Figura 6.28a). A solução 6 da execução 9 (retângulo vermelho na Figura 6.28b) apresenta valores baixos (células verdes) para todas as variáveis de saída, ou seja, esta solução está entre uma fração significativa das soluções (aproximadamente um terço, conforme assinalado no retângulo vermelho na Figura 6.21) que apresentam uma boa “representatividade baseada em curvas de risco” em relação a todas as variáveis de saída. Na Figura 6.27c, essa representatividade pode ser vista em maiores detalhes, comparando a semelhança entre a curva dos modelos originais e a curva dos MRs dessa solução. Por fim, na Figura 6.28c, é possível observar os MRs da solução 6 da execução 9 (retângulo vermelho na Figura 6.28c). Em particular, nota-se que esses modelos (com exceção do modelo 3) não fazem parte do subconjunto comum de MRs (47, 2, 16, 3 e 11), conforme ilustrado na Figura 6.22, presentes no conjunto de MRs de aproximadamente dois terços das soluções selecionadas. Ou seja, essa seria uma solução mais atípica do que outras desse conjunto de soluções, o que poderia aumentar a variabilidade das soluções selecionadas, caso mais de uma solução fosse escolhida.

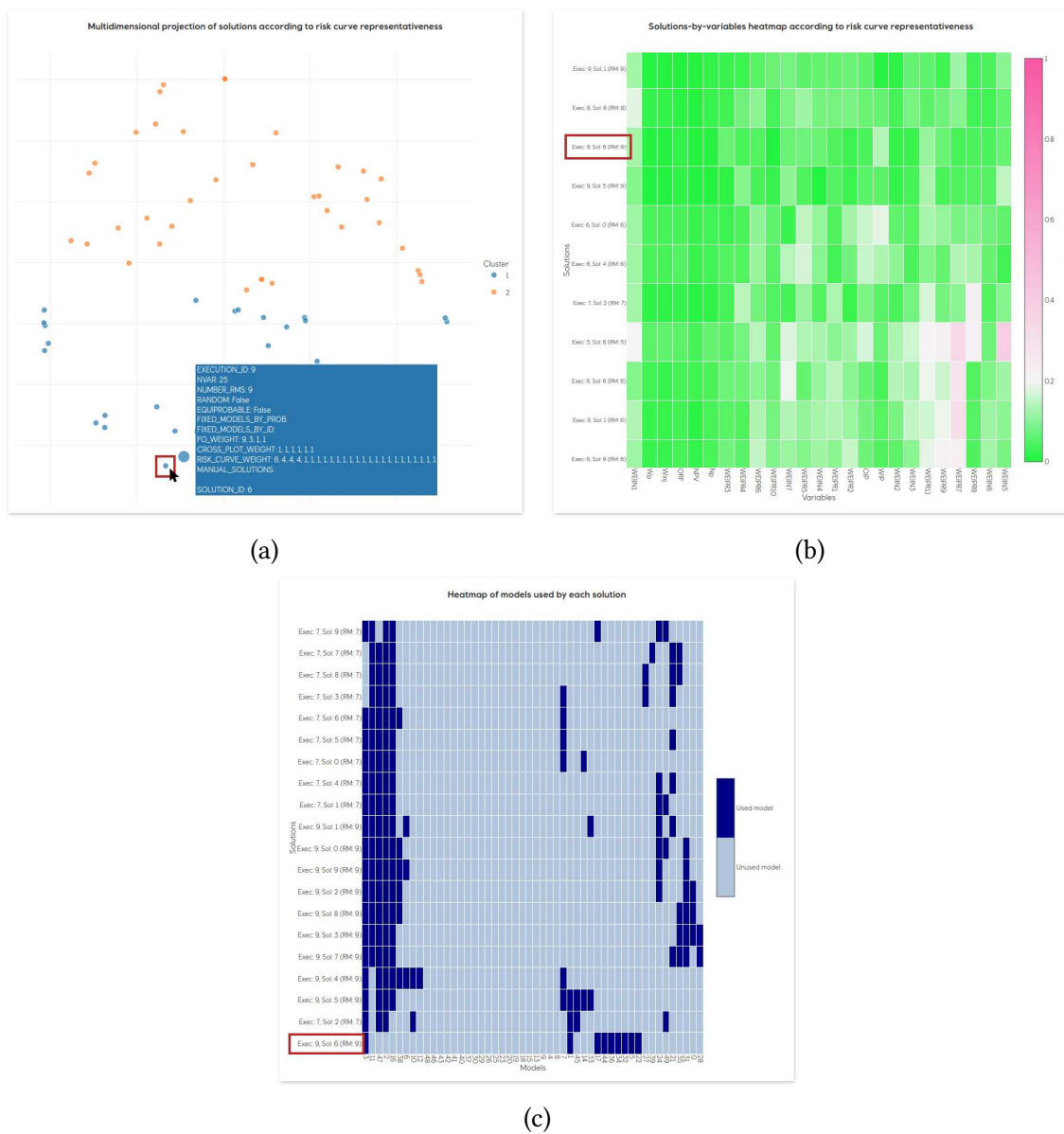


Figura 6.28: Gráficos de visão geral destacando a solução 6 e execução 9, usando dados do caso de estudo OLYMPUS: (a) projeção multidimensional de curvas de risco; (b) *heatmap* reordenável de variáveis por soluções; e (c) *heatmap* de modelos utilizados por solução.

Fonte: Dados originais da pesquisa.

6.3 Considerações finais

Este capítulo apresentou a implementação do sistema de visualização interativo proposto. Na Seção 6.1, descreveu-se detalhadamente o desenvolvimento desse sistema: a escolha dos algoritmos utilizados, tais como t-SNE, *X-Means* e *Optimal Leaf Ordering* para projeção multidimensional, clusterização e reordenação de matriz, respectivamente; os mapeamentos visuais, dando enfoque na representação matemática; os recursos interativos; e as tarefas

associadas a cada representação visual. Em seguida, na Seção 6.2, foi apresentado um caso de uso realista com o objetivo de selecionar uma solução do RMFinder com base na representatividade em curvas de risco por meio do sistema de visualização interativo.

Capítulo 7

Avaliação do sistema de visualização

Nessa etapa do projeto, realizou-se um teste remoto com usuários; nesse teste, os usuários executaram no sistema de visualização interativo, desenvolvido no Capítulo 6, algumas atividades relativas à escolha de MRs, como forma de avaliar se as visualizações propostas auxiliam, do ponto de vista desses usuários, nesse processo de escolha. Suas opiniões foram registradas, analisadas e sumarizadas, visando aprimoramento das visualizações.

Na Seção 7.1 constam os detalhes da preparação do teste remoto com os usuários e, na Seção 7.2, a análise dos resultados dessa avaliação. Em seguida, na Seção 7.3, encontra-se os limites e possibilidades de continuidade desta pesquisa. A Seção 7.4 demonstra como as abordagens técnicas desenvolvidas no sistema de visualização interativo estão relacionadas às tarefas do usuário, bem ao objetivo deste trabalho (Seção 1.3). A última Seção 7.5 apresenta uma comparação entre os trabalhos relacionados na Seção 3.1 e esta pesquisa.

7.1 Preparação da avaliação

Essa etapa de avaliação do sistema de visualização interativo foi submetida para aprovação do CEP/UNICAMP no mesmo projeto de pesquisa elencado na Seção 5.1. Além disso, por conta da necessidade de entrevistas e desenvolvimento do protótipo (Seções 5.1 e 5.2, respectivamente), não teve suas atividades descritas em detalhes nesse projeto de pesquisa. No entanto, foi mencionada com uma breve descrição, relatando os seus objetivos, e como ela seria executada (um teste remoto com usuário).

Com o intuito de avaliar se as tarefas dos usuários, capturadas na etapa de entrevistas (Seção 5.1) e às quais se pretendia dar apoio, estavam contempladas no sistema de

visualização interativo, os participantes envolvidos na etapa de entrevista (Seção 5.1) foram novamente convidados a participar voluntariamente desse teste remoto. Somado a isso, eles foram informados que no sistema a ser avaliado eles poderiam executar suas tarefas usando os casos de estudo OLYMPUS (FONSECA; GEEL; LEEUWENBURGH, 2017), UNISIM-I-D (GASPAR et al., 2015) e UNISIM-II-D (CORREIA et al., 2015); além disso, foi dada a opção de eles enviarem por *e-mail* seus próprios projetos executados no RMFinder (ou seja, novos conjuntos de dados) para serem disponibilizados na aplicação. Sendo assim, dois participantes compartilharam seus próprios projetos: um participante compartilhou quatro projetos e outro participante compartilhou dois projetos, totalizando seis novos projetos.

Após inserir esses novos conjuntos de dados no SGDB do sistema proposto, foi enviado um *e-mail* aos participantes, convidando-os para esta etapa de teste remoto, contendo os seguintes itens:

- (i) antes de iniciar com os testes, os participantes deveriam obter uma cópia do TCLE (Apêndice E) disponibilizado por meio de um *link*, previamente assinado e rubricado pelos pesquisadores, assinar e anexar no retorno do *e-mail*;
- (ii) foi disponibilizado aos participantes um vídeo tutorial sobre o uso da ferramenta e suas respectivas visualizações;
- (iii) essa é uma etapa de teste remoto, em que o usuário deveria acessar e testar a ferramenta implementada; sendo assim foi disponibilizado um acesso à aplicação por meio de um *link* para que ele pudesse avaliar o sistema de visualização. Para disponibilizar esse *link* com acesso à ferramenta, a aplicação foi instalada em uma instância Amazon Elastic Compute (EC2)¹ da família *c5a.2xlarge* na Amazon Web Services (AWS)², rodando um contêiner³ na plataforma *Docker*⁴ com um banco de dados *PostgreSQL* e os serviços *WebViz* da aplicação (Seção 4.2). Os participantes que compartilharam seus projetos receberam um *link* específico contendo seus dados (adicionalmente aos casos de estudo citados), preservando desse modo a confidencialidade da informação ao não compartilhá-la com os outros participantes;

¹Capacidade de computação redimensionável para praticamente qualquer *workload*.

²Plataform de computação em nuvem

³Um único pacote de software, conhecido como um contêiner, agrupa o código de um aplicativo e os respectivos arquivos de configuração e bibliotecas necessários para a execução do aplicativo.

⁴O *Docker* é uma plataforma destinada a ajudar os desenvolvedores a criar, compartilhar e executar aplicativos de contêineres.

(iv) foi ainda disponibilizado um formulário com questionário sobre a ferramenta (Apêndice F), contendo dez afirmações para as quais ele poderia expressar o seu nível de concordância e, ao final, duas perguntas a serem respondidas textualmente de forma livre. O nível de concordância foi expressado por meio de uma escala *Likert*⁵ com cinco níveis; são eles: discordo totalmente, discordo, não estou decidido, concordo e concordo totalmente.

Por fim, foi estipulado um prazo para os participantes finalizarem essa etapa de teste remoto, inicialmente compreendida de 31 de outubro de 2023 a 07 de novembro de 2023. No entanto, dois participantes julgarem o prazo infactível e solicitaram um período maior. Sendo assim, essa etapa finalizou em 13 de novembro de 2023 contendo a avaliação dos quatro participantes convidados (os mesmos envolvidos na etapa de entrevista da Seção 5.1).

7.2 Análise dos resultados da avaliação

As dez afirmações da Tabela 7.1, contidas no formulário detalhado no Apêndice F, foram respondidas pelos quatro participantes convidados de acordo com o seu grau de concordância, de modo a realizar uma avaliação quantitativa do sistema de visualização interativo. Essas afirmações foram organizadas para avaliar se as representações visuais da ferramenta podem ajudar nas tarefas de análise de dados dos usuários (elencadas na Seção 5.1); e se os recursos interativos disponíveis em cada gráfico facilitam a exploração dos dados, auxiliando a revelar relações que poderiam permanecer ocultas em uma visão estática ou inicial.

Desse modo, as afirmações um e dois correspondem ao gráfico de projeção multidimensional de curvas de risco; três e quatro, ao *heatmap* reordenável de variáveis por soluções; cinco, ao *heatmap* reordenável de modelos utilizados por solução; seis e sete, ao *heatmap* reordenável de modelos utilizados por solução; oito, às curvas de risco sumarizadas por envoltório; e nove, às curvas de risco justapostas e sobrepostas. A afirmação dez está intrinsecamente relacionada ao objetivo deste presente documento (Seção 1.3): avaliar se o sistema de visualização interativo ajuda o usuário a efetuar melhores escolhas de conjuntos de modelos representativos de reservatórios de petróleo em cenários com múltiplas soluções e variáveis.

⁵A Escala *Likert* é uma técnica amplamente utilizada na avaliação de opiniões, a qual permite entender o grau de satisfação dos usuários a uma determinada ferramenta, p. ex..

Tabela 7.1: Perguntas para a etapa de teste remoto do sistema de visualização interativo.

1	2	3	4	5	
Discordo totalmente	Discordo	Não estou decidido	Concordo	Concordo totalmente	
Perguntas quantitativas da escala Likert				Graus de concordância	
1. O gráfico “projeção multidimensional de soluções de acordo com a representatividade da curva de risco” possibilita ao usuário identificar a similaridade das soluções em relação ao conjunto de modelos originais (ou melhor, o conjunto de modelos utilizados como entrada do RMFinder).	1	2	3	4	5
2. Os recursos interativos (filtragens, zoom, parâmetros e texto em foco), disponíveis no gráfico de projeção multidimensional, auxiliam o usuário a observar de várias formas as soluções de acordo com a representatividade da curva de risco, em busca de novas informações (como por exemplo descobrir que os dados se relacionam de uma forma que poderia estar oculta em sua representação visual inicial).	1	2	3	4	5
3. O mapa de calor de modelos usados para cada solução contribui para o usuário encontrar padrões relevantes em conjuntos de soluções com relação à sua variabilidade, como soluções com subconjuntos comuns de modelos representativos.	1	2	3	4	5
4. Os recursos interativos (filtragens, parâmetros, texto em foco e reordenação), disponíveis no mapa de calor de modelos usados em cada solução, auxiliam o usuário a observar de várias formas a variabilidade das soluções (como subconjuntos comuns de modelos representativos) em busca de novas informações (como por exemplo descobrir que os dados se relacionam de uma forma que poderia estar oculta em sua representação visual inicial).	1	2	3	4	5
5. O gráfico “mapa de calor de soluções por variáveis de acordo com a representatividade da curva de risco” permite ao usuário analisar a representatividade de uma dada solução com relação a várias variáveis.	1	2	3	4	5
6. O gráfico “mapa de calor de soluções por variáveis de acordo com a representatividade da curva de risco” auxilia o usuário a comparar soluções com números distintos de modelos representativos e observar quais variáveis de produção provocam mais ou menos erro (ou seja, quais variáveis precisam de mais modelos representativos para serem representadas corretamente).	1	2	3	4	5
7. Os recursos interativos (filtragens, zoom, parâmetros, texto em foco e reordenação), disponíveis no mapa de calor de soluções por variáveis de acordo com a representatividade da curva de risco, auxiliam o usuário a observar de várias formas o erro de cada solução quando comparado ao conjunto completo de modelos (ou melhor, o conjunto de modelos utilizados como entrada do RMFinder) em busca de novas informações (como por exemplo descobrir que os dados se relacionam de uma forma que poderia estar oculta em sua representação visual inicial).	1	2	3	4	5
8. Os gráficos de curvas de risco (que mostram uma curva por variável) com intervalos de confiança que resumem o conjunto de curvas de risco de várias soluções junto ao conjunto de modelos originais (ou melhor, o conjunto de modelos utilizados como entrada do RMFinder), ajudam os usuários na análise da representatividade das soluções baseada em sua curva de risco e os seus respectivos detalhes.	1	2	3	4	5
9. Os recursos interativos (filtragens, zoom, parâmetros, texto em foco e seleção de pontos entre gráficos), disponíveis nos gráficos de curvas de risco justapostas e sobrepostas, auxiliam o usuário de várias formas na seleção do melhor conjunto de modelos representativos em busca de novas informações (como por exemplo descobrir que os dados se relacionam de uma forma que poderia estar oculta em sua representação visual inicial).	1	2	3	4	5
10. O conjunto de visualizações interativas de múltiplas curvas de risco do protótipo avaliado ajuda o usuário a efetuar melhores escolhas de conjuntos de modelos representativos de reservatórios de petróleo em cenários com múltiplas soluções e variáveis.	1	2	3	4	5

Na Figura 7.1 são ilustradas as respostas consolidadas dos quatro participantes para as dez afirmações por meio de um gráfico de barras horizontal empilhado. Cada barra horizontal no eixo Y representa uma questão da Tabela 7.1, a qual é segregada no eixo X pela frequência relativa dos cinco graus de concordância, identificados por suas cores.

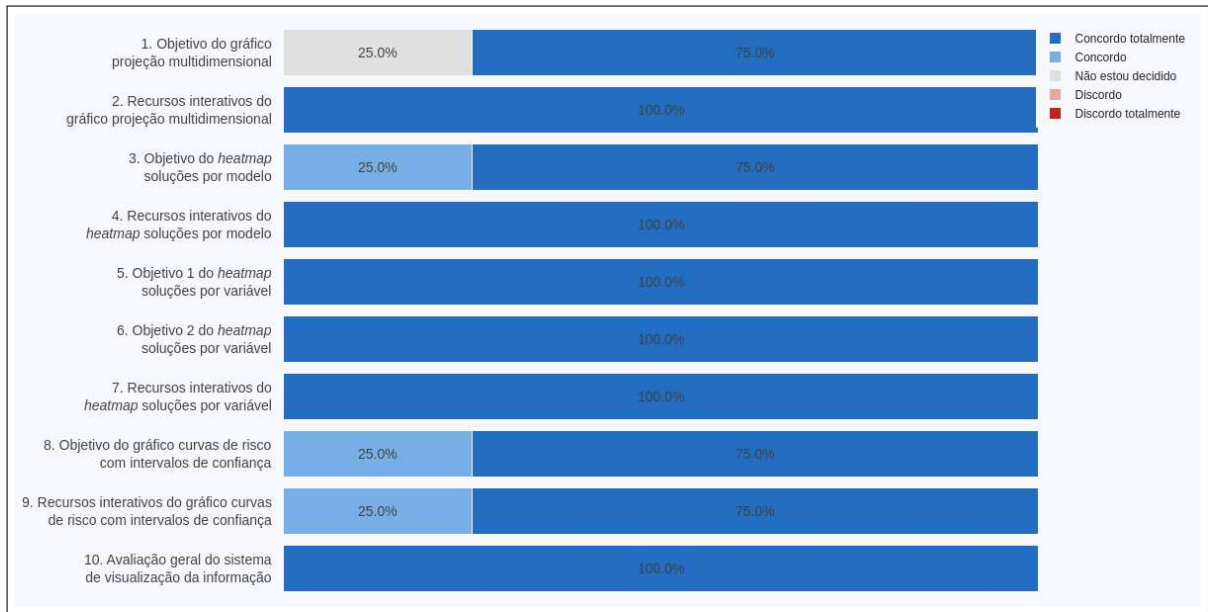


Figura 7.1: Análise dos resultados da avaliação do sistema de visualização interativo.

Fonte: Dados originais da pesquisa.

Observa-se na Figura 7.1 que nenhum participante atribuiu as categorias “discordo totalmente” ou “discordo” para umas das dez afirmações; portanto, o resultado quantitativo da avaliação do sistema de visualização interativo, via etapa de teste remoto com o usuário, não apresentou nenhum nível de discordância. Somado a isso, nota-se um alto percentual de concordância (categorias “concordo totalmente” e “concordo”) nas opiniões quantitativas dos participantes (com exceção de um participante não estar decidido se o gráfico projeção multidimensional de curvas de risco possibilita identificar a similaridade das soluções em relação ao conjunto de modelos originais).

Além disso, com intuito de também realizar uma avaliação qualitativa (Seção 2.1.6), duas perguntas dissertativas foram acrescentadas ao final do formulário do Apêndice F: uma para explorar os pontos positivos do sistema de visualização interativo, e a outra os pontos a melhorar; são elas:

- Escreva sobre os pontos positivos que avaliou com pontuação igual ou maior que 4; e
- Escreva sobre os pontos a melhorar que avaliou com pontuação igual ou menor que 3

A análise dos resultados qualitativos dos pontos positivos é descrita a seguir:

- um participante destacou “que o gráfico de projeção multidimensional de soluções de acordo com a representatividade da curva de risco sugere a afinidade dos MRs em relação

ao conjunto base, permitindo assim a eliminação de soluções ruins e a rápida escolha das melhores soluções que ainda podem passar pelas demais avaliações de acordo com o interesse do tomador de decisão”. Esse comentário reforça a hipótese da projeção multidimensional de curvas de risco (Seção 5.2): possibilitar ao usuário a identificação da similaridade entre cada solução e o conjunto de todos os modelos (ou “conjunto base”);

- (ii) outro participante observou que o *heatmap* reordenável de variáveis por soluções de acordo com a representatividade da curva de risco pode “auxiliar na definição mais apropriada do número de MRs para um dado modelo avaliado, visto que auxilia na visualização de subconjuntos de MRs”. Essa tarefa não foi associada ao *heatmap* de soluções por variáveis na Seção 5.2 (ou a outra representação visual), sendo uma tarefa registrada via etapa de teste remoto com o usuário;
- (iii) as curvas de risco sumarizadas por envoltório, segundo um participante, permitiram “identificar o *range* de estocasticidade do método e traçar a confiabilidade dos MRs selecionados segundo uma dada variável, bem como identificar os percentis menos ou mais representativos”;
- (iv) o recurso de reordenação de matriz (presente nos *heatmaps* de variáveis por soluções e modelos utilizados por solução) guiou um participante a identificar “quais modelos seriam os mais comuns entre as diferentes soluções” no *heatmap* de modelos utilizados por solução, por exemplo; essa descoberta, “caso não houvesse esta opção, não ficaria tão clara”;
- (v) os recursos interativos do sistema de visualização foram amplamente mencionados; são eles: as técnicas de interação de “filtragens, *zoom*, parâmetros, texto em foco e seleção de pontos entre gráficos permitem a rápida identificação de um dado MR ou solução, e auxiliam na tomada de decisão frente a outras soluções”; outro ponto mencionado foi a funcionalidade de exportação para PNG (Figura 6.18); o passeio pelo sistema de visualização (Figura 6.19) “ajudou bastante (algumas funções que não são tão intuitivas ficaram claras com o tutorial — como, por exemplo, selecionar uma região dos mapas e gerar as curvas de risco somente das soluções dessa região)”; e os filtros para “selecionar somente algumas execuções ajudaram bastante a analisar com mais detalhes” as soluções; e

- (vi) em geral o sistema de visualização foi avaliado como “intuitivo e fácil de usar”, “interface bastante amigável” e as representações visuais ajudam “bastante no processo de seleção dos MRs”.

Os pontos a melhorar vieram na forma de sugestões a serem implementadas no sistema de visualização e dificuldades enfrentadas durante o teste remoto. A análise desses resultados qualitativos são descritas a seguir:

- (i) “não foi possível visualizar um número grande de curvas de risco no protótipo para identificar a relação entre os dados observados no gráfico de projeção multidimensional, impossibilitando assim a concordância sobre o assunto. Porém, havendo relação é uma excelente ferramenta para eliminar soluções ruins e restringir o espaço de busca pela melhor solução”. Esse comentário foi realizado por um participante (o mesmo que respondeu não estar decidido na afirmação um da Tabela 7.1), o qual compartilhou três projetos do RMFinder com uma execução e uma solução cada (i. e., conjuntos de dados sem a presença de múltiplas soluções), conforme exemplo ilustrado na Figura 7.2. Neste contexto, a projeção multidimensional de curvas de risco não deve auxiliar o usuário a identificar soluções adequadas de acordo com a representatividade da curva de risco;
- (ii) além do passeio pelo sistema de visualização (Figura 6.19), um dos participantes sugeriu um “botão de interrogação em cada um dos tópicos (*execution*, *parameters*, etc.), para tirar dúvidas específicas de cada um, sem ter que passar pelo *tour*”;
- (iii) dois participantes relataram dificuldades como lentidão e travamento para carregarem as curvas de risco justapostas e sobrepostas “dependendo da quantidade de soluções a serem geradas”. Em particular, essa lentidão e/ou travamento nas curvas de risco justapostas e sobrepostas podem ser resolvidas delegando parte do processamento para o *client-side*⁶; e
- (iv) outro participante relatou que a “seleção de pontos entre gráficos nos recursos interativos (gráficos de curvas de risco justapostas e sobrepostas) ficou pouco destacada”. Sendo assim, sugeriu usar “uma outra cor para identificar os pontos selecionados entre gráficos”.

⁶O *client-side* é todo processamento em um aplicativo *web* que ocorre no dispositivo do cliente (usuário final), incluindo imagem, texto, etc.

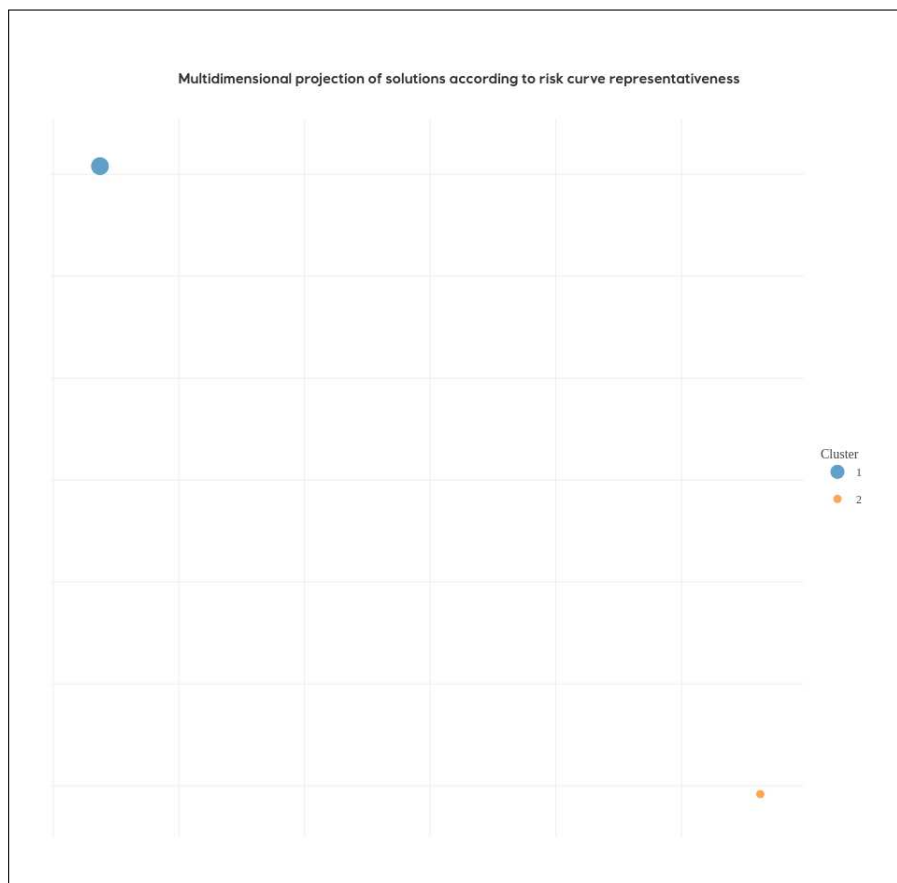


Figura 7.2: Gráfico de projeção multidimensional de curvas de risco com apenas uma solução.
Fonte: Dados originais da pesquisa.

A análise do resultado quantitativo (alto percentual de concordância e nenhuma discordância), somada às respostas qualitativas dos pontos positivos, dão fortes indícios de que as representações visuais e os recursos interativos, disponíveis no sistema de visualização, são capazes de ajudar os usuários envolvidos na etapa de teste remoto a efetuarem melhores escolhas de conjuntos de MRs (ou soluções) de reservatórios de petróleo, além de possibilitar comparações de mais alto nível de abstração por meio das visualizações de caráter mais geral, permitindo comparar cenários com diversas variáveis e soluções. Os pontos a melhorar trouxeram sugestões relevantes a serem aplicadas no sistema de visualização. No entanto, o grupo de participantes é pequeno (apenas quatro); nesse sentido, os resultados positivos obtidos neste trabalho são um indicativo de que as soluções apresentadas são promissoras, e novos testes, envolvendo outros usuários do RMFinder ou *software* similar, devem ser conduzidas. Esse aspecto foi mencionado por um participante quando comentou que “todos os gráficos são úteis para auxiliar os usuários a identificar e

melhorar o conjunto de MRs, embora mais pesquisa e desenvolvimento ainda deverão ser feitos”.

7.3 Limites e possibilidades de continuidade da pesquisa

É importante assinalar alguns pontos que podem caracterizar os limites desta pesquisa, bem como as oportunidades de trabalhos futuros. Nesse sentido, por conta da indisponibilidade de tempo no cronograma deste trabalho, realizou-se somente um ciclo de desenvolvimento iterativo, conforme metodologia detalhada na Seção 4.3. Sendo assim, as melhorias assinaladas pelos participantes na etapa de teste remoto não foram implementadas no sistema de visualização interativo, podendo ser incorporadas na continuidade deste trabalho (i. e., em um novo ciclo de desenvolvimento iterativo).

Além disso, o RMFinder possui quatro funções objetivo multicritério (Seção 2.3.1) e o escopo deste trabalho contempla somente uma delas: a função objetivo curva de risco. Outro pesquisador do SEIS, sob a liderança do mesmo orientador deste trabalho, tem desenvolvido visualizações associadas à função objetivo *cross-plot*. Em virtude disso, o desenvolvimento de visualizações relacionadas às demais funções objetivo do RMFinder, bem como a integração entre elas, podem ser considerados em trabalhos futuros.

Por fim, indica-se para trabalhos futuros avaliar se os algoritmos t-SNE (MAATEN; HINTON, 2008) para projeção multidimensional, o *X-Means* (PELLEG; MOORE, 2000) para clusterização e o algoritmo *Optimal Leaf Ordering* (BAR-JOSEPH; GIFFORD; JAAKKOLA, 2001) para reordenação de matriz são os mais adequados para o conjunto de dados do RMFinder e as visualizações nos quais foram aplicados. Neste trabalho, foram mencionados os artigos de Nonato e Aupetit (2019) e Baroni e Silva (2022) onde são elencadas recomendações para a escolha de uma das diversas técnicas de projeção multidimensional e reordenação de matriz, respectivamente. Em particular, no *heatmap* reordenável de modelos utilizados por solução (Seção 6.1.3), de modo a reordenar os modelos utilizados por solução, além da reordenação de matriz, também seria possível implementar uma ordenação simples, na qual os modelos seriam ordenados de acordo com o número de soluções que os utilizam. Além disso, é recomendável avaliar se é possível realizar o agrupamento por meio de um algoritmo de *clustering* diretamente no domínio multidimensional original dos dados, ao invés de fazê-lo diretamente no resultado da projeção (neste trabalho, foram utilizadas as

técnicas *X-Means* (PELLEG; MOORE, 2000) e t-SNE (MAATEN; HINTON, 2008) para o agrupamento e projeção, respectivamente).

7.4 Relacionamento entre as tarefas do usuário e as abordagens técnicas

Inspirado na estrutura de quatro níveis de validação de Munzner (2014), o intuito dessa seção é demonstrar o relacionamento em cascata entre o objetivo da pesquisa, tarefas dos usuários, mapeamento visual e interações, e algoritmos implementados no sistema de visualização. O nível de tarefas dos usuários está relacionado às questões de “por que e o quê”, enquanto os níveis de mapeamento visual, interações e algoritmos estão relacionados à questão de “como”. Além disso, essa relação em quatro níveis pode auxiliar na compreensão de como cada técnica empregada neste trabalho está relacionada à tarefa do usuário e ao objetivo de pesquisa (Seção 1.3). Dessa forma, o relacionamento entre as tarefas do usuário e as abordagens técnicas utilizadas neste trabalho é descrito a seguir:

1. **Objetivo:** visualizações de caráter mais geral, possibilitando comparações de mais alto nível de abstração em cenários com diversas variáveis de saída e soluções.
 - 1.1. **Tarefa:** identificar similaridade entre pares de soluções e também a similaridade entre cada solução e o conjunto de todos os modelos.
 - 1.1.1. **Mapeamento visual:** projeção multidimensional de curvas de risco.
 - 1.1.1.1. Projeção multidimensional via algoritmo t-SNE, utilizando as medidas de distância “área entre as curvas de risco” e o teste não-paramétrico proposto por Kolmogorov-Smirnov.
 - 1.1.1.2. Cada ponto representa uma solução com múltiplas variáveis de saída, considerando a média entre as matrizes de distância (uma para cada variável de saída).
 - 1.1.1.3. Ponto com um tamanho maior identifica o conjunto de todos os modelos.
 - 1.1.1.4. Grupo de soluções semelhantes agrupados com cores diferentes, utilizando o algoritmo *clustering X-means*.
 - 1.2. **Tarefa:** analisar a representatividade de uma dada solução com relação a várias variáveis de saída.

1.2.1. **Mapeamento visual:** *Heatmap* reordenável de variáveis por soluções.

1.2.1.1. Técnica baseada em região (*heatmap*).

1.2.1.2. Variáveis de saída e soluções nos eixos X e Y , respectivamente.

1.2.1.3. As células indicam a “representatividade baseada em curvas de risco” de uma variável de saída para uma solução. Essa representatividade é definida a partir de uma medida de similaridade (“área entre as curvas de risco” ou o teste não-paramétrico proposto por Kolmogorov-Smirnov) entre a curva de risco do conjunto de todos os modelos e a curva de risco de uma solução.

1.2.1.4. A escala de cores segue um padrão divergente que se inicia em verde (ausência de erro), passando pelo branco (valor neutro) e seguindo até roxo (erro máximo no conjunto de dados).

1.2.1.5. Reordenação de matriz por meio do algoritmo *Optimal Leaf Ordering*, com o objetivo de identificar quais variáveis de saída provocam mais ou menos erros; ou identificar a existência de grupo de soluções com boa “representatividade baseada em curvas de risco” em relação a todas as variáveis de saída.

1.3. **Tarefa:** comparar soluções com números distintos de MRs e observar quais variáveis de produção provocam mais ou menos erro (ou seja, quais variáveis precisam de mais MRs para serem representadas corretamente).

1.3.1. **Mapeamento visual:** *Heatmap* reordenável de variáveis por soluções.

1.3.1.1. Técnica baseada em região (*heatmap*).

1.3.1.2. Variáveis de saída e soluções nos eixos X e Y , respectivamente.

1.3.1.3. As células indicam a “representatividade baseada em curvas de risco” de uma variável de saída para uma solução.

1.3.1.4. A escala de cores segue um padrão divergente que se inicia em verde (ausência de erro), passando pelo branco (valor neutro) e seguindo até roxo (erro máximo no conjunto de dados).

1.3.1.5. Reordenação de matriz por meio do algoritmo *Optimal Leaf Ordering*, com o objetivo de identificar quais variáveis de saída provocam mais ou menos erros em grupo de soluções com números distintos de MRs.

1.4. **Tarefa:** encontrar padrões relevantes em conjuntos de soluções com relação à sua variabilidade, como soluções com subconjuntos comuns de MRs.

1.4.1. **Mapeamento visual:** *Heatmap* reordenável de modelos utilizados por solução

1.4.1.1. Técnica baseada em região (*heatmap*).

1.4.1.2. A célula azul-escuro representa o modelo utilizado na solução como MR, enquanto a cor azul-clara indica a ausência do modelo na solução. A cor vermelha representa o modelo fixado na configuração dos parâmetros de execução do RMFinder.

1.4.1.3. Modelos e soluções nos eixos *X* e *Y*, respectivamente.

1.4.1.4. Reordenação de matriz usando o algoritmo *Optimal Leaf Ordering*, com o objetivo de agrupar soluções similares e modelos que foram usados nas mesmas soluções.

1.5. **Tarefa:** analisar a representatividade das soluções baseada em sua curva de risco.

1.5.1. **Mapeamento visual:** curvas de risco sumarizadas por envoltório.

1.5.1.1. Um conjunto de curvas de risco (ou de soluções) sumarizado por um envoltório cinza (similar a uma “sombra”), fornecendo uma visão geral da similaridade das soluções em relação às suas curvas de risco para cada variável de saída.

1.5.1.2. O envoltório é composto por duas curvas de risco: uma que representa o limite superior e a outra que representa o limite inferior do conjunto de curvas de risco em questão.

2. **Objetivo:** melhorias nas visualizações atualmente utilizadas pelo *software* RMFinder.

2.1. **Tarefa:** analisar as curvas de risco das soluções e os seus respectivos detalhes.

2.1.1. **Interações:** curvas de risco justapostas e sobrepostas.

2.1.1.1. Filtragem por soluções e variáveis de saída.

2.1.1.2. Texto em destaque com os valores do modelo (variável de saída e percentual acumulado) ao posicionar o ponteiro do *mouse* na marca da curva de risco de todos os modelos.

- 2.1.1.3. Texto em destaque com os valores do modelo (variável de saída e percentual acumulado) ao posicionar o ponteiro do mouse no ponto (ou MR) da curva de risco de uma solução.
- 2.1.1.4. Texto em destaque com o detalhamento da execução ao posicionar o ponteiro do *mouse* na curva de risco de uma solução.
- 2.1.1.5. Seleção de modelos (incluindo os MRs de uma solução) nas curvas de risco de uma variável de saída, ao clicar com o botão esquerdo do mouse (utilizando tecla SHIFT para multi-seleção), e destacá-los nas curvas de risco das outras variáveis de saída (recurso interativo de *brushing* e *linking*, destacado na Seção 2.1.5).
- 2.1.1.6. Seleção das curvas de risco justapostas e sobrepostas com a oclusão da curva de risco de todos os modelos.

7.5 Comparação entre os trabalhos relacionados

A Tabela 7.2 apresenta a mesma comparação realizada na Tabela 3.2, entre os trabalhos relacionados à InfoVis aplicados na seleção de MRs; no entanto, essa nova comparação inclui a presente pesquisa (última linha da Tabela 7.2) com o objetivo de posicioná-la em relação à literatura atual, identificada por meio do processo de revisão sistemática descrito no Apêndice A.

Tabela 7.2: Comparação entre a literatura atual e a presente pesquisa.

Artigos	Projeção Multidimensional			Técnicas de Agrupamento		Medida de Distância		Heatmaps de Modelos		CCDF			Cenários com múltiplas		Mapas		Outros	
	PCA	MDS	t-SNE	K-means	X-means	Área	K-S	Representatividade	Variabilidade	Sobreposta	Justa-posta	Sumarizada	Soluções	Variáveis	2D	3D	Reorientação de Matriz	Interação
Anupam e Tewari, 2021	X		X	X														
Amaral, 2021										X	X							X
Mahjour et al., 2020							X			X								
Silva, Santos e Schiozer, 2019					X	X		X					X	X	X	X	X	
Barbiero et al., 2019				X														
Sahaf et al., 2019		X		X												X		X
Sahaf et al., 2018		X		X												X		X
Zheng et al., 2018		X		X														
Meira et al., 2017										X	X							
Schiozer et al., 2004										X								
Este presente trabalho			X		X	X	X	X	X	X	X	X	X	X			X	X

7.6 Considerações finais

O Capítulo 7 apresentou a avaliação do sistema de visualização interativo. Na Seção 7.1 dissertou-se sobre os preparativos da avaliação, tais como: seleção dos participantes voluntários, disponibilização do sistema ao usuário e questionário sobre a ferramenta. Em seguida, na Seção 7.2, discutiu-se o teste remoto com usuários, detalhando a avaliação quantitativa, pontos positivos e a melhorar. A Seção 7.3 tratou dos limites desta pesquisa e das oportunidades futuras, dando enfoque à continuidade desta pesquisa. A Seção 7.4 apresentou um mapeamento entre as abordagens técnicas empregas no sistema de

visualização e as tarefas do usuário. Por fim, na Seção 7.5, é apresentada uma comparação entre este trabalho e a literatura atual.

Capítulo 8

Conclusão

A simulação computacional de modelos de reservatórios é uma ferramenta indispensável para o desenvolvimento e gerenciamento de reservatórios de petróleo. Os engenheiros de petróleo desenvolvem uma grande variedade de modelos para representar as incertezas de um determinado reservatório de petróleo. No entanto, simular o comportamento de todos esses modelos pode ser difícil em algumas etapas da metodologia, ou inapropriada a depender do alto custo computacional. Nesse sentido, o RMFinder é uma metodologia e um *software* para reduzir os cenários possíveis (ou seleção de MRs), sendo uma estratégia para contornar essa limitação. Este trabalho respondeu ao problema de pesquisa de que é possível aprimorar os recursos de visualização de dados de probabilidade acumulada de riscos atualmente utilizados para auxiliar engenheiros de petróleo na escolha de MRs em processo de desenvolvimento e gerenciamento de reservatórios de petróleo com foco nas curvas de risco geradas pelo RMFinder.

Os mapeamentos visuais apresentados neste trabalho têm como objetivo auxiliar o engenheiro de petróleo a analisar a representatividade de uma curva de risco e comparar um conjunto de curvas de risco para encontrar a mais representativa. Esses mapeamentos visuais foram disponibilizados por meio de um sistema de visualização interativo. Nesse sistema, estão disponíveis gráficos de visão geral, como: projeção multidimensional de curvas de risco, *heatmap* de variáveis por soluções e *heatmap* de modelos utilizados por solução. Além disso, esse sistema possui recursos interativos para filtrar e selecionar as curvas de risco e detalhá-las por meio do gráfico curvas de risco sumarizadas por envoltório e, sob-demanda, avaliar o último nível de detalhamento, as curvas de risco justapostas e sobrepostas.

O objetivo geral deste trabalho de propor e/ou aprimorar visualizações interativas de múltiplas curvas de risco para ajudar os analistas de petróleo a efetuarem melhores escolhas de conjuntos de MRs de reservatórios de petróleo em cenários com múltiplas soluções e variáveis foi alcançado, com boas avaliações no teste remoto com usuários, sinalizando que as representações visuais apresentadas neste trabalho são promissoras.

Além disso, como objetivo específico, este trabalho conseguiu focar na proposição de melhorias para as visualizações atualmente utilizadas pelo *software* RMFinder, bem como na proposição de visualizações de caráter mais geral, atualmente ausentes nesse sistema. Essas visualizações foram capazes de possibilitar comparações de mais alto nível de abstração, permitindo comparar cenários com diversas variáveis e soluções, conforme as avaliações registradas por meio do teste remoto com usuários.

Apesar de o objetivo inicial deste trabalho ter sido alcançado, ainda existem diversas oportunidades para aprimoramento em trabalhos futuros, tais como: avaliar se os algoritmos utilizados para projeção multidimensional, clusterização e reordenação de matriz são os mais adequados para os conjuntos de dados do RMFinder; propor um sistema de visualização interativo contemplando as quatro funções objetivo multicritério do RMFinder e integrações entre elas; aprimorar o desempenho do sistema ao carregar no navegador do usuário as curvas de risco justapostas e sobrepostas de várias soluções; implementar uma ordenação simples no *heatmap* de modelos utilizados por solução, na qual os modelos seriam ordenados de acordo com o número de soluções que os utilizam; verificar se é possível realizar a clusterização no domínio multidimensional original dos dados, em vez de aplicar diretamente no resultado da projeção; e aperfeiçoar a documentação do sistema para esclarecer dúvidas específicas sobre cada um de seus componentes de modo a guiar de forma padronizada a seleção de MRs. A implementação desses trabalhos futuros tornará o sistema de visualização interativo mais completo e relevante para a tomada de decisões.

Referências bibliográficas

AMARAL, L. D. B. **Visualização de informação aplicada à análise de resultados de processo de seleção de modelos representativos de reservatórios de petróleo**. 2021. Trabalho de Conclusão de Curso – Faculdade de Tecnologia, Universidade Estadual de Campinas, Limiera.

ANUPAM, A.; TEWARI, R. Easy Steps in Probabilistic Dynamic Modelling in FDP. In: INTERNATIONAL Petroleum Technology Conference. Virtual: [s.n.], mar. 2021. DOI: 10.2523/IPTC-21323-MS.

BAR-JOSEPH, Z.; GIFFORD, D. K.; JAAKKOLA, T. S. Fast optimal leaf ordering for hierarchical clustering. **Bioinformatics**, v. 17, suppl 1, s22–s29, jun. 2001. ISSN 1367-4803. DOI: 10.1093/bioinformatics/17.suppl_1.S22. Disponível em: <https://doi.org/10.1093/bioinformatics/17.suppl_1.S22>.

BARBIERO, M.; TURRI, F.; ANASTASI, P.; ROSSA, E. Screening of Geological Uncertainty on Reservoir Dynamic Behavior with Statistical Learning Techniques. In: SPE Reservoir Characterisation and Simulation Conference and Exhibition. Abu Dhabi, UAE: Society of Petroleum Engineers (SPE), set. 2019. spe-196712-ms. DOI: 10.2118/196712-MS.

BARONI, M. P.; SILVA, C. G. A comparative analysis of matrix reordering algorithms regarding canonical data patterns. **Information Visualization**, SAGE Publications Ltd, v. 21, p. 321–332, 3 jul. 2022. ISSN 14738724. DOI: 10.1177/14738716221091487.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. **Readings in Information Visualization: Using Vision to Think**. Londres, UK: Morgan Kaufmann Publ., 1999. (The Morgan Kaufmann series in interactive technologies). ISBN 9780613915861.

CARPENDALE, S. Evaluating Information Visualizations. In: **Information Visualization: Human-Centered Issues and Perspectives**. Edição: Andreas Kerren, John T. Stasko, Jean-Daniel Fekete e Chris North. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. P. 19–45. ISBN 978-3-540-70956-5. DOI: 10.1007/978-3-540-70956-5_2. Disponível em: <https://doi.org/10.1007/978-3-540-70956-5_2>.

CORREIA, M.; HOHENDORFF, J. C. V.; GASPAR, A.; SCHIOZER, D. UNISIM-II-D: Benchmark Case Proposal Based on a Carbonate Reservoir. In: SPE Latin American and Caribbean Petroleum Engineering Conference. Quito, Equador: Society of Petroleum Engineer, nov. 2015. P. 21. DOI: 10.2118/177140-MS.

FONSECA, R.; GEEL, C.; LEEUWENBURGH, O. **Description of OLYMPUS reservoir model for optimization challenge**. 2017. Disponível em: <<http://www.isapp2.com/downloads/olympus-reservoir-model.pdf>>. Acesso em: 2 set. 2022.

GASPAR; AVANSI, G.; SANTOS, A.; HOHENDORFF, J. C. V.; SCHIOZER, D. UNISIM-I-D. benchmark studies for oil field development and production strategy selection. **International Journal of Modeling and Simulation for the Petroleum Industry**, v. 9, p. 47–55, jan. 2015.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data Clustering: A Review. Association for Computing Machinery, New York, NY, USA, v. 31, n. 3, p. 264–323, set. 1999. ISSN 0360-0300. DOI: 10 . 1145 / 331499 . 331504. Disponível em: <<https://doi.org/10.1145/331499.331504>>.

KEIM, D. A. Information Visualization and Visual Data Mining. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Educational Activities Department, USA, v. 8, n. 1, p. 1–8, jan. 2002. ISSN 1077-2626. DOI: 10 . 1109 / 2945 . 981847. Disponível em: <<https://doi.org/10.1109/2945.981847>>.

KERREN, A.; STASKO, J. T.; FEKETE, J.-D.; NORTH, C. **Information Visualization: Human-Centered Issues and Perspectives**. 1. ed. Berlin, Heidelberg: Springer, 2008. 177 p. (Lecture Notes in Computer Science). ISBN 978-3-540-70955-8. DOI: 10 . 1007 / 978 - 3 - 540 - 70956 - 5.

KRUSKAL, J. B.; LANDWEHR, J. M. Icicle Plots: Better Displays for Hierarchical Clustering. **Source: The American Statistician**, Taylor and Francis, Ltd. on behalf of the American Statistical Association, v. 37, p. 162–168, 2 1983.

KULYK, O.; KOSARA, R.; URQUIZA, J.; WASSINK, I. Human-Centered Aspects. In: KERREN, A.; EBERT, A.; MEYER, J. (Ed.). **Human-Centered Visualization Environments**. Berlin: Springer, 2017. cap. 2, p. 13–75.

MAATEN, L. van der; HINTON, G. Visualizing Data using t-SNE. **Journal of Machine Learning Research**, v. 9, n. 86, p. 2579–2605, 2008.

MACQUEEN, J. B. Some Methods for Classification and Analysis of MultiVariate Observations. In: CAM, L. M. L.; NEYMAN, J. (Ed.). **Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability**. [S.l.]: University of California Press, 1967. v. 1, p. 281–297.

MAHJOUR, S. K.; SANTOS, A. A. S.; CORREIA, M. G.; SCHIOZER, D. J. Developing a workflow to select representative reservoir models combining distance-based clustering and data assimilation for decision making process. **Journal of Petroleum Science and Engineering**, Elsevier B.V., v. 190, jul. 2020. ISSN 09204105. DOI: 10.1016/j.petrol.2020.107078.

MAZZA, R. **Introduction to Information Visualization**. London: Springer Science e Business Media, 2009. DOI: 10 . 1007 / 978 - 1 - 84800 - 219 - 7.

MEIRA, L. A. A.; COELHO, G. P.; SANTOS, A. A. S.; SCHIOZER, D. J. Selection of Representative Models for Decision Analysis Under Uncertainty. **Computers and Geosciences**, Elsevier Ltd, v. 88, p. 67–82, mar. 2016. ISSN 00983004. DOI: 10.1016/j.cageo.2015.11.012.

MEIRA, L. A. A.; COELHO, G. P.; SILVA, C. G.; SCHIOZER, D. J. RMFinder 2.0: An Improved Interactive Multi-Criteria Scenario Reduction Methodology. In: SPE Latin America and

Caribbean Petroleum Engineering Conference. Society of Petroleum Engineers. Buenos Aires, Argentina: [s.n.], mai. 2017. DOI: 10.2118/185502-MS.

MEIRA, L. A. A. et al. Improving representativeness in a scenario reduction process to aid decision making in petroleum fields. **Journal of Petroleum Science and Engineering**, Elsevier B.V., v. 184, p. 1–19, jan. 2020. ISSN 09204105. DOI: 10.1016/j.petrol.2019.106398.

MIRZAEI-PAIAMAN, A.; SANTOS, S. M. G.; SCHIOZER, D. J. A review on closed-loop field development and management. **Journal of Petroleum Science and Engineering**, Elsevier B.V., v. 201, jun. 2021. ISSN 09204105. DOI: 10.1016/j.petrol.2021.108457.

MOHER, D.; LIBERATI, A.; TETZLAFF, J.; ALTMAN, D. G. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. **International Journal of Surgery**, v. 8, p. 336–341, 5 2010. ISSN 17439191. DOI: 10.1016/j.ijvsu.2010.02.007.

MUNZNER, T. **Visualization Analysis and Design**. [S.l.]: CRC Press, 2014. (AK Peters Visualization Series). ISBN 9781466508934.

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**, v. 25, n. 8, p. 2650–2673, 2019. DOI: 10.1109/TVCG.2018.2846735.

PELLEG, D.; MOORE, A. W. X-Means: Extending K-Means with Efficient Estimation of the Number of Clusters. In: PROCEEDINGS of the Seventeenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000. (ICML '00), p. 727–734. ISBN 1558607072.

PÓVOA, M. A. M.; DA SILVA, S. O.; SILVA, C. G. Prototyping visualizations as a support for selecting representative models of petroleum reservoirs. In: 17TH International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing (CGVCVIP). Porto, Portugal: [s.n.], jun. 2023. DOI: 10.13140/RG.2.2.31233.02408.

SAHAF, Z.; MOTA, R. C.; HAMDI, H.; SOUSA, M. C.; MAURER, F. A visual analytics framework for exploring uncertainties in reservoir models. In: VISIGRAPP 2018 - Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. Funchal, Madeira, Portugal: SciTePress, 2018. v. 3, p. 74–84. ISBN 9789897582899. DOI: 10.5220/0006608500740084.

SAHAF, Z.; MOTA, R. C.; HAMDI, H.; SOUSA, M. C.; MAURER, F. An ROI visual-analytical approach for exploring uncertainty in reservoir models. In: COMMUNICATIONS in Computer and Information Science. Funchal, Madeira, Portugal: Springer Verlag, 2019. v. 997, p. 121–142. ISBN 9783030267551. DOI: 10.1007/978-3-030-26756-8_6.

SANTOS, S. M. G.; SANTOS, A. A.; BALDON, S.; GABRIEL, S.; SCHIOZER, D. J. **Representative models and RMFinder Program**. Campinas: Technical Report. UNISIM - CEPETRO/UNICAMP, 6 out. 2021.

SANTOS, S. M. G.; SANTOS, A. A. S.; SCHIOZER, D. J. Selecting Representative Models for Ensemble-Based Production Optimization in Carbonate Reservoirs with Intelligent Wells and WAG Injection. In: ECMOR XVII – 17th European Conference on the Mathematics of Oil

Recovery. Online Event: European Association of Geoscientists e Engineers, set. 2020. P. 14–17. DOI: 10.3997/2214-4609.202035041.

SANTOS, S. M.; GASPAR, A. T. F.; SCHIOZER, D. J. Risk management in petroleum development projects: Technical and economic indicators to define a robust production strategy. **Journal of Petroleum Science and Engineering**, Elsevier B.V., v. 151, p. 116–127, 2017. ISSN 09204105. DOI: 10.1016/j.petrol.2017.01.035.

SANTOS, S. M. G.; GASPAR, A. T. F. S.; SCHIOZER, D. J. Managing reservoir uncertainty in petroleum field development: Defining a flexible production strategy from a set of rigid candidate strategies. **Journal of Petroleum Science and Engineering**, Elsevier B.V., v. 171, p. 516–528, dez. 2018. ISSN 09204105. DOI: 10.1016/j.petrol.2018.07.048.

SCHIOZER, D. J.; LIGERO, E. L.; SUSLICK, S. B.; COSTA, A. P. A.; SANTOS, J. A. M. Use of representative models in the integration of risk analysis and production strategy definition. **Journal of Petroleum Science and Engineering**, v. 44, p. 131–141, 1-2 out. 2004. ISSN 09204105. DOI: 10.1016/j.petrol.2004.02.010.

SCHIOZER, D. J.; SANTOS, A. A. S.; DRUMOND, P. S. Integrated Model Based Decision Analysis in Twelve Steps Applied to Petroleum Fields Development and Management. In: EUROPEC. Madri, Espanha: Society of Petroleum Engineer, jul. 2015. DOI: 10.2118/174370-MS.

SCHIOZER, D. J.; SANTOS, A. A. S.; SANTOS, S. M. G.; FILHO, J. C. V. H. Model-based decision analysis applied to petroleum field development and management. **Oil and Gas Science and Technology**, Editions Technip, v. 74, 2019. ISSN 19538189. DOI: 10.2516/ogst/2019019.

SHNEIDERMAN, B. The eyes have it: a task by data type taxonomy for information visualizations. In: PROCEEDINGS 1996 IEEE Symposium on Visual Languages. [S.l.: s.n.], 1996. P. 336–343. DOI: 10.1109/VL.1996.545307.

SILVA, C. G. Visualização de Informação: introdução e influências de IHC. In: BOSCARIOLI, C.; BIM, S. A.; LEITÃO, C.; MACIEL, C. (Ed.). **Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais – IHC 2014**. 13. ed. Foz do Iguaçu, PR: Sociedade Brasileira de Computação - SBC, 2014. cap. 4, p. 79–108. ISBN 978-85-7669-295-9.

SILVA, C. G.; MEIRA, L. A. A.; SANTOS, A. A. S.; SCHIOZER, D. J. Using Reorderable Matrices to Compare Risk Curves of Representative Models in Oil Reservoir Development and Management Activities. In: 4TH International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing. Zagreb, Croacia: International Association for Development of the Information Society, jul. 2020. P. 19–26. ISBN 978-989-8704-21-4. DOI: 10.33965/cgv2020_202011L003.

SILVA, C. G. et al. An improved visualization-based approach for project portfolio selection. **Computers in Human Behavior**, Elsevier Ltd, v. 73, p. 685–696, ago. 2017. ISSN 07475632. DOI: 10.1016/j.chb.2016.12.083.

SILVA, C. G. **Visualization research at the Software Engineering and Information Systems Laboratory (SEIS)**. Limeira, São Paulo: School of Technology – University of Campinas, 6 out. 2021. Disponível em: <<https://www.ft.unicamp.br/seis>>.

SILVA, C. G.; SANTOS, A. A. S.; SANTOS; SCHIOZER, D. Visualization of Ensembles of Oil Reservoir Models Based on Pixelization, Small Multiples and Reservoir Similarities. In: OFFSHORE Technology Conference Brasil. Rio de Janeiro, BR: Offshore Technology Conference, out. 2019. DOI: 10.4043/29788-MS.

SPENCE, R. **Information Visualization: An Introduction**. 3. ed. Londres, UK: Springer International Publishing, 2014. 321 p. ISBN 9783319073415. DOI: 10.1007/978-3-319-07341-5.

STASKO, J.; ZHANG, E. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In: IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings. Salt Lake City, UT, USA: IEEE, 2000. P. 57–65. ISBN 0-7695-0804-9. DOI: 10.1109/INFVIS.2000.885091.

STEAGALL, D.; SCHIOZER, D. Uncertainty Analysis In Reservoir Production Forecasts During Appraisal And Pilot Production Phases. In: SPE Reservoir Simulation Symposium. Houston, Texas: Society of Petroleum Engineers, fev. 2001. DOI: 10.2118/66399-MS.

WARD, M.; GRINSTEIN, G.; KEIM, D. **Interactive Data Visualization: Foundations, Techniques, and Applications, Second Edition**. 2. ed. Boca Raton, FL: CRC Press, 2015. 548 p. (360 Degree Business). ISBN 9781482257380.

ZHENG, B.; SADLO, F. On the Visualization of Hierarchical Multivariate Data. In: 2021 IEEE 14th Pacific Visualization Symposium (PacificVis). Tianjin, China: IEEE, 2021. P. 136–145. DOI: 10.1109/PacificVis52677.2021.00026.

ZHENG, J.; LEUNG, J. Y.; SAWATZKY, R. P.; ALVAREZ, J. M. A Cluster-Based Approach for Visualizing and Quantifying the Uncertainty in the Impacts of Uncertain Shale Barrier Configurations on SAGD Production. In: Day 2 Wed, March 14, 2018. (SPE Canada Heavy Oil Conference). DOI: 10.2118/189753-MS.

Apêndice A

Processo de revisão sistemática

Questão 1. Quais são os recursos de visualização de função de distribuição cumulativa complementar utilizados para auxiliar engenheiros de petróleo na escolha de modelos representativos em processo de desenvolvimento e gerenciamento de reservatórios de petróleo?

Questão 2. Quais são os recursos de visualização utilizados para auxiliar engenheiros de petróleo na escolha de modelos representativos em processo de desenvolvimento e gerenciamento de reservatórios de petróleo?

Questão 3. Quais são os recursos de visualização de dados de função de distribuição cumulativa complementar utilizados para auxiliar analistas na escolha de modelos representativos em outras áreas (por exemplo, previsão de tempo e mineração)?

Base de dados:

- IEEE Xplore Digital Library: <http://ieeexplore.ieee.org/>;
- ACM Digital Library: <https://dl.acm.org/> e;
- OnePetro: <https://onepetro.org/>.

Intervalo: De 2017 a 2023.

Palavras-chave ou termos:

- Sinônimos de visualização: visualization, dashboard, graph;
- Sinônimos de petróleo: petroleum, oil;
- Sinônimos de reservatório: reservoir, field;

- Sinônimos de curva de risco: risk curve, culmulative distribution function, CDF, complementary cumulative distribution functions, CCDF e;
- Outros substantivos: representative models, model.

Busca:

- Questão 1:
 - (visualization OR dashboard OR graph) AND (“representative models”) AND (“risk curve” OR ccdf OR (“culmulative distribution function”) AND (petroleum OR oil) AND (reservoir OR field)
- Questão 2:
 - (visualization OR dashboard OR graph) AND (“representative models”) AND (petroleum OR oil) AND (reservoir OR field)
- Questão 3:
 - (visualization OR dashboard OR graph) AND (“representative models”) AND (“risk curve” OR ccdf OR “culmulative distribution function”)

Resultados:

- IEEE Xplore Digital Library: 0;
- ACM Digital Library: 0 e;
- OnePetro: 5.

Apêndice B

Artigos resultantes da revisão sistemática

A Tabela B.1 apresenta os 5 artigos resultantes do processo de revisão sistemática, detalhado no Apêndice A.

Tabela B.1: Artigos resultantes da revisão sistemática.

Artigos
Easy Steps in Probabilistic Dynamic Modelling in FDP (Anupam e Tewari, 2021)
Visualization of Ensembles of Oil Reservoir Models Based on Pixelization, Small Multiples and Reservoir Similarities (Silva, Santos e Schiozer, 2019)
Screening of Geological Uncertainty on Reservoir Dynamic Behavior with Statistical Learning Techniques (Barbiero et al., 2019)
A Cluster-Based Approach for Visualizing and Quantifying the Uncertainty in the Impacts of Uncertain Shale Barrier Configurations on SAGD Production (Zheng et al., 2018)
RMFinder 2.0: An Improved Interactive Multi-Criteria Scenario Reduction Methodology (Meira et al., 2017)

Apêndice C

Termo de consentimento livre e esclarecido (etapa de entrevistas)

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Visualização de múltiplas curvas de risco e de variáveis de produção de campos e poços como auxílio na seleção de modelos representativos de reservatórios de petróleo (Etapa de entrevistas)

Celmar Guimarães da Silva, Miguel Angelo Mena Póvoa, Samuel Oliveira da Silva

Número do CAAE: 68642523.9.0000.5404

Você está sendo convidado a participar como voluntário de uma pesquisa. Este documento, chamado Termo de Consentimento Livre e Esclarecido, visa assegurar seus direitos como participante e é elaborado em duas vias, uma que deverá ficar com você e outra com o pesquisador.

Por favor, leia com atenção e calma, aproveitando para esclarecer suas dúvidas. Se houver perguntas antes ou mesmo depois de assiná-lo, você poderá esclarecê-las com o pesquisador. Se preferir, pode levar este Termo para casa e consultar seus familiares ou outras pessoas antes de decidir participar. Não haverá nenhum tipo de penalização ou prejuízo se você não aceitar participar ou retirar sua autorização em qualquer momento.

Justificativa e objetivos:

Esse projeto visa propor e/ou aprimorar visualizações interativas de múltiplas curvas de risco e de variáveis de produção para ajudar os analistas de petróleo a efetuarem melhores escolhas de conjuntos de modelos representativos (MRs) de reservatórios de petróleo em

cenários com múltiplas estratégias, soluções e variáveis.

Procedimentos

Ao concordar em participar do estudo, você está sendo convidado a colaborar em entrevistas presenciais nas quais serão observados aspectos que são considerados relevantes no processo de escolha da qualidade dos MRs na execução da ferramenta RMFinder 3.0. Esta etapa de entrevistas será conduzida com base em tópicos relacionados às atividades dos participantes, no formato de uma entrevista semiestruturada. Nesta etapa pretende-se apresentar visualizações preliminares aos participantes, que reflitam as primeiras ideias dos pesquisadores deste projeto para auxiliá-los em suas tarefas. A intenção é que essas visualizações incentivem os participantes a observarem o potencial do uso de gráficos interativos na tentativa de apoio às suas tarefas, e a colaborarem com ideias.

Ainda nesta etapa algumas questões serão levantadas quanto a como os participantes tomam suas decisões na escolha dos MRs. Algumas questões serão levantadas por meio de tópicos.

Cada participante participará de uma entrevista de aproximadamente 1h30, feita em local conveniente para o entrevistado. Se for assinado pelo participante, a entrevista terá o áudio gravado para facilitar posterior análise pelos pesquisadores, sendo este áudio descartado após 5 anos. A gravação do áudio não é essencial para a participação na pesquisa (ou seja, é possível não gravar e ainda assim participar); a opção de consentimento está disponível na seção “**Gravação de áudio**” deste documento.

Desconfortos e riscos:

A pesquisa não apresenta riscos previsíveis. Caso haja algum tipo de dúvida referente ao material, as dúvidas poderão ser esclarecidas no momento das entrevistas. As entrevistas serão feitas seguindo o ritmo de conversa do participante entrevistado, podendo ser interrompidas a qualquer momento.

São critérios de exclusão desta pesquisa não ter conhecimento prévio sobre o uso de modelos representativos para o desenvolvimento e gerenciamento de campos de petróleo, bem como não ter utilizado previamente o software RMFinder. Usuários com deficiências visuais comprovadas que comprometam o uso das soluções (cegueira, baixa visão). Os participantes serão questionados sobre isso antes do agendamento da entrevista, visando

evitar situações desconfortáveis.

Benefícios:

O experimento a ser realizado neste projeto permitirá aos pesquisadores validarem (ainda que parcialmente) as visualizações propostas, que posteriormente serão colocadas à disposição dos profissionais da área de Engenharia de Petróleo para uso na seleção de modelos representativos.

Especialistas e pesquisadores em Engenharia de Petróleo serão beneficiados diretamente pelo resultado desta pesquisa, pois poderão passar a utilizar as visualizações em suas avaliações na seleção de modelos representativos. Ao cumprirem o papel de facilitar essa seleção de modelos representativos, as visualizações contribuirão com fases de processos de desenvolvimento e gerenciamento de campos de petróleo; em especial, uma escolha adequada de modelos representativos possibilitará escolhas mais adequadas de estratégias de exploração usando menores quantidades de dados (poucos modelos, e modelos que bem representam o conjunto total de modelos do reservatório em questão), colaborando para uma maior agilidade no processo.

Acompanhamento e assistência:

Você tem o direito à assistência integral e gratuita devido a danos diretos e indiretos, imediatos e tardios, pelo tempo que for necessário. Você poderá entrar em contato com os pesquisadores a qualquer momento através dos dados disponibilizados na seção “**Contato**” deste documento.

Sigilo e privacidade:

Você tem a garantia de que sua identidade será mantida em sigilo e nenhuma informação será dada a outras pessoas que não façam parte da equipe de pesquisadores. Na divulgação dos resultados desse estudo, seu nome não será citado.

Gravação de áudio:

A gravação do áudio não é essencial para a participação na pesquisa (é possível não gravar e ainda assim participar).

- () AUTORIZO o registro de gravação de áudio, para uso restrito nesta pesquisa.

Ressarcimento e indenização:

A participação neste trabalho é de caráter totalmente voluntário e será realizada em horário que você julgue conveniente. Você não terá qualquer despesa por participar desta pesquisa. Caso tenha gastos para participar desta pesquisa fora da sua rotina você será ressarcido integralmente de suas despesas.

Tratamento dos dados:

Esta pesquisa prevê o armazenamento dos dados coletados em repositório de dados, em local virtual de acesso público, com o objetivo de possível reutilização, verificação e compartilhamento em trabalhos de colaboração científica com outros grupos de pesquisa.

Sua identidade não será revelada nesses dados, pois os dados só serão armazenados de forma anônima (isto é, os dados não terão identificação), utilizando mecanismos que impeçam a possibilidade de associação, direta ou indireta com você. Cabe ressaltar que quem compartilhar os dados também não terá possibilidade de identificação dos participantes de quem os dados se originaram. Sendo assim, não haverá possibilidade de reversão da anonimização.

Contato:

Em caso de dúvidas sobre a pesquisa, você poderá entrar em contato com os pesquisadores:

- Prof. Dr. Celmar Guimarães da Silva, no endereço profissional Rua Paschoal Marmo, 1888 – CEP: 13484-332 – Jd. Nova Itália – Limeira, SP – Brasil – Laboratório Sistemas de Informação e Engenharia de *Software* – SIES, cujo telefone é (19) 2113-3439 e e-mail celmar@ft.unicamp.br.
- Miguel Angelo Mena Póvoa, no mesmo endereço profissional acima, cujo telefone pessoal é (11) 95083-3534 e e-mail m016954@dac.unicamp.br.
- Samuel Oliveira da Silva, no mesmo endereço profissional acima, cujo telefone pessoal é (19) 98456-0024 e e-mail s231294@dac.unicamp.br.

Em caso de denúncias ou reclamações sobre sua participação e sobre questões éticas do estudo, você poderá entrar em contato com a secretaria do Comitê de Ética em Pesquisa

(CEP) da UNICAMP das 8h30 às 11h30 e das 13h às 17h na Rua Tessália Vieira de Camargo, 126; CEP 13083-887 Campinas — SP; telefone (19) 3521-8936 ou (19) 3521-7187; e-mail: cep@fcm.unicamp.br.

O Comitê de Ética em Pesquisa (CEP)

O papel do CEP é avaliar e acompanhar os aspectos éticos de todas as pesquisas envolvendo seres humanos. A Comissão Nacional de Ética em Pesquisa (CONEP) tem por objetivo desenvolver a regulamentação sobre proteção dos seres humanos envolvidos nas pesquisas. Desempenha um papel coordenador da rede de Comitês de Ética em Pesquisa (CEPs) das instituições, além de assumir a função de órgão consultor na área de ética em pesquisas.

Consentimento livre e esclarecido:

Após ter recebido esclarecimentos sobre a natureza da pesquisa, seus objetivos, métodos, benefícios previstos, potenciais riscos e o incômodo que esta possa acarretar, aceito participar:

Nome do(a) participante da pesquisa: _____

(Assinatura do participante)

Data

Responsabilidades dos Pesquisadores:

Asseguramos ter cumprido as exigências da resolução 466/2012 CNS/MS e complementares na elaboração do protocolo e na obtenção deste Termo de Consentimento Livre e Esclarecido. Asseguramos, também, ter explicado e fornecido uma via deste documento ao participante da pesquisa. Informamos que o estudo foi aprovado pelo CEP perante o qual o projeto foi apresentado e pela CONEP, quando pertinente. Comprometemo-nos a utilizar o material e os dados obtidos nesta pesquisa exclusivamente para as finalidades previstas neste documento ou conforme o consentimento dado pelo participante.

Prof. Dr. Celmar Guimarães da Silva	Data
-------------------------------------	------

Miguel Angelo Mena Póvoa	Data
--------------------------	------

Samuel Oliveira da Silva	Data
--------------------------	------

Apêndice D

Perguntas para guiar a entrevista semiestruturada

Segue abaixo as perguntas para guiar a entrevista semiestruturada sobre o processo de seleção de modelos representativos baseado na metodologia *software* RMFinder 3.0.

1. Como costuma ser seu uso do RMFinder para encontrar bons modelos representativos?
2. Como costuma analisar se uma solução é boa? Por onde começa? Quais os passos? Onde você foca inicialmente sua análise? E quais os passos seguintes?
3. Faz várias execuções? Em que você se foca em cada execução? Costuma atribuir pesos para gráficos (crossplots ou curvas de risco) nessas execuções? Como costuma fazer para decidir sobre como usar esses pesos? Como costuma definir se não são necessárias mais execuções (ou seja, como você entende que o resultado obtido na última execução está de acordo com o que você esperava)?
4. Você costuma analisar cenários com muitas variáveis? Quantas? O que são “muitas variáveis” para você?
5. Como costuma fazer sua análise em cenários em que há muitas variáveis?
6. Você compara muitas soluções em uma mesma execução? Quantas? Se sim, qual seu objetivo quando faz isso? O que são “muitas soluções” para você? Como você costuma fazer para analisar muitas soluções? Uma vez que fez a análise de muitas soluções, que tomada de decisão você faz e em que você a baseia?

7. Que deficiências acha que o RMFinder tem para apoiar tomadas de decisão envolvendo muitas variáveis, considerando a sua forma de usá-lo?
8. Que deficiências acha que o RMFinder tem para apoiar tomadas de decisão envolvendo muitas soluções, considerando a sua forma de usá-lo?
9. Que deficiências acha que o RMFinder tem para apoiar tomadas de decisão envolvendo muitas execuções, considerando a sua forma de usá-lo?
10. Você utiliza o relatório gerado pelo RMFinder? Tem sido útil?

Apêndice E

Termo de consentimento livre e esclarecido (etapa teste remoto)

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Visualização de múltiplas curvas de risco e de variáveis de produção de campos e poços como auxílio na seleção de modelos representativos de reservatórios de petróleo (Etapa de teste remoto de sistema)

Celmar Guimarães da Silva, Miguel Angelo Mena Póvoa, Samuel Oliveira da Silva

Número do CAAE: 68642523.9.0000.5404

APRESENTAÇÃO DA PESQUISA:

Você está sendo convidado a participar como voluntário de uma pesquisa. Este documento, chamado Termo de Consentimento Livre e Esclarecido, visa assegurar seus direitos como participante e é elaborado em duas vias, uma que deverá ficar com você e outra com o pesquisador.

Por favor, leia com atenção e calma, aproveitando para esclarecer suas dúvidas. Se houver perguntas antes ou mesmo depois de assiná-lo, você poderá esclarecê-las com o pesquisador. Se preferir, pode levar este Termo para casa e consultar seus familiares ou outras pessoas antes de decidir participar. Não haverá nenhum tipo de penalização ou prejuízo se você não aceitar participar ou retirar sua autorização em qualquer momento.

INFORMAÇÕES SOBRE ESTA PESQUISA:

Objetivos:

Esse projeto visa propor e/ou aprimorar visualizações interativas de múltiplas curvas de risco e de variáveis de produção para ajudar os analistas de petróleo a efetuarem melhores escolhas de conjuntos de modelos representativos (MRs) de reservatórios de petróleo em cenários com múltiplas estratégias, soluções e variáveis.

Importância do estudo:

Trabalhar com modelos representativos de cenários, em vez de trabalhar com um conjunto completo de centenas de modelos, tem por objetivo tornar viável a simulação computacional de diferentes cenários e estratégias de desenvolvimento e gerenciamento de reservatórios de petróleo. Porém, se a escolha não for adequada, ela pode conduzir a tomadas de decisão inapropriadas, que afetarão a forma como se utilizará esses reservatórios. O projeto em questão se foca justamente em prover um suporte adequado para possibilitar melhores tomadas de decisão, considerando a quantidade de dados envolvida para isso.

Dessa maneira, entende-se como relevância social deste projeto que o apoio a melhores tomadas de decisão na escolha de modelos representativos impactará positivamente em processos de desenvolvimento e gerenciamento de reservatórios de petróleo, implicando uma melhor utilização dos recursos não renováveis (petróleo e gás) disponíveis nesses reservatórios.

Procedimentos e metodologias:

Ao concordar em participar do estudo, você colaborará com um teste remoto. Nesse teste os usuários utilizarão um protótipo de software para executar algumas tarefas relativas à escolha de modelos representativos, como forma de avaliar se as visualizações propostas auxiliam, do ponto de vista desses usuários, esse processo de escolha. Suas opiniões serão registradas, analisadas e sumarizadas, visando aprimoramento das visualizações.

Tratamento dos dados:

Esta pesquisa prevê o armazenamento dos dados coletados em repositório de dados, em local virtual de acesso público, com o objetivo de possível reutilização, verificação e compartilhamento em trabalhos de colaboração científica com outros grupos de pesquisa.

Sua identidade não será revelada nesses dados, pois os dados só serão armazenados de forma anônima (isto é, os dados não terão identificação), utilizando mecanismos que

impeçam a possibilidade de associação, direta ou indireta, com você. Cabe ressaltar que quem compartilhar os dados também não terá possibilidade de identificação dos participantes de quem os dados se originaram. Sendo assim, não haverá possibilidade de reversão da anonimização.

Desconfortos e riscos previstos:

Não foram identificados riscos à saúde mental ou física dos participantes ou dos pesquisadores. Caso haja algum tipo de dúvida referente ao material, a comunicação entre participante e pesquisador se resumirá a mensagens enviadas por e-mail, o que pode prolongar o processo.

Pode ocorrer de haver falta de tempo do participante para a participação na pesquisa, em especial frente ao prazo de entrega estipulado. Caso isso ocorra, será solicitado que o participante entre em contato com os pesquisadores para verificar a possibilidade de estender o prazo.

Benefícios:

O experimento a ser realizado neste projeto permitirá aos pesquisadores validarem (ainda que parcialmente) as visualizações propostas, que posteriormente serão colocadas à disposição dos profissionais da área de Engenharia de Petróleo para uso na seleção de modelos representativos.

Especialistas e pesquisadores em Engenharia de Petróleo serão beneficiados diretamente pelo resultado desta pesquisa, pois poderão passar a utilizar as visualizações em suas avaliações na seleção de modelos representativos. Ao cumprirem o papel de facilitar essa seleção de modelos representativos, as visualizações contribuirão com fases de processos de desenvolvimento e gerenciamento de campos de petróleo; em especial, uma escolha adequada de modelos representativos possibilitará escolhas mais adequadas de estratégias de exploração usando menores quantidades de dados (poucos modelos, e modelos que bem representam o conjunto total de modelos do reservatório em questão), colaborando para uma maior agilidade no processo.

Acompanhamento e assistência:

Você tem o direito à assistência integral e gratuita devido a danos diretos e indiretos, imediatos e tardios, pelo tempo que for necessário. Você poderá entrar em contato com os pesquisadores a qualquer momento através dos dados disponibilizados na seção **“Forma de contato com os pesquisadores”** deste documento.

Forma de contato com os pesquisadores:

Em caso de dúvidas sobre a pesquisa, você poderá entrar em contato com os pesquisadores:

- Prof. Dr. Celmar Guimarães da Silva, no endereço profissional Rua Paschoal Marmo, 1888 – CEP: 13484-332 – Jd. Nova Itália – Limeira, SP – Brasil – Laboratório Sistemas de Informação e Engenharia de *Software* – SIES, cujo telefone é (19) 2113-3439 e e-mail celmar@ft.unicamp.br.
- Miguel Angelo Mena Póvoa, no mesmo endereço profissional acima, cujo telefone pessoal é (11) 95083-3534 e e-mail m016954@dac.unicamp.br.
- Samuel Oliveira da Silva, no mesmo endereço profissional acima, cujo telefone pessoal é (19) 98456-0024 e e-mail s231294@dac.unicamp.br.

Em caso de denúncias ou reclamações sobre sua participação e sobre questões éticas do estudo, você poderá entrar em contato com a secretaria do Comitê de Ética em Pesquisa (CEP) da UNICAMP das 8h30 às 11h30 e das 13h às 17h na Rua Tessália Vieira de Camargo, 126; CEP 13083-887 Campinas – SP; telefone (19) 3521-8936 ou (19) 3521-7187; e-mail: cep@fcm.unicamp.br.

O Comitê de Ética em Pesquisa (CEP)

O papel do CEP é avaliar e acompanhar os aspectos éticos de pesquisas envolvendo seres humanos, protegendo o(a)s participantes em seus direitos e dignidade. Em caso de dúvidas, denúncias ou reclamações sobre aspectos éticos de sua participação e sobre seus direitos como participante da pesquisa, entre em contato com a secretaria do Comitê de Ética em Pesquisa da UNICAMP (CEP-UNICAMP), de segunda a sexta-feira, das 08:00hs às 11:30hs e das 13:00hs às 17:30hs à Rua: Tessália Vieira de Camargo, 126; CEP 13083-887 Campinas – SP; telefones (19) 3521-8936 e (19) 3521-7187; e-mail: cep@fcm.unicamp.br. Se houver necessidade da intermediação da comunicação em Libras, você pode fazer contato com a Central TILS da Unicamp no site <https://www.prg.unicamp.br/tils> e solicitar ajuda para

comunicação com o CEP.

GARANTIAS AOS PARTICIPANTES:

Esclarecimentos:

Você será informado(a) e esclarecido(a) sobre os aspectos relevantes da pesquisa, antes, durante e depois da pesquisa, mesmo se esta informação causar sua recusa na participação ou sua saída da pesquisa.

Direito de recusa a participar e direito de retirada do consentimento:

Você tem direito de se recusar a participar da pesquisa e de desistir e retirar o seu consentimento em qualquer momento da pesquisa sem que isto traga qualquer penalidade ou represálias de qualquer natureza e sem que haja prejuízo para você.

Sigilo e privacidade:

Você tem a garantia de que sua identidade será mantida em sigilo e nenhuma informação será dada a outras pessoas que não façam parte da equipe de pesquisadores. Na divulgação dos resultados desse estudo, seu nome não será citado.

Ressarcimento:

Você não terá qualquer despesa por participar desta pesquisa. Caso tenha gastos para participar desta pesquisa fora da sua rotina você será ressarcido integralmente de suas despesas.

Assistência, indenização e medidas de reparação:

Você tem direito de buscar indenização e reparação de danos se sentir-se prejudicado(a) pela participação nesta pesquisa, mesmo se já tiver concordado em participar da pesquisa e assinado o TCLE.

Entrega do TCLE:

Você receberá este Termo assinado e rubricado pelos pesquisadores por meio do acesso
ao link:

<https://drive.google.com/drive/folders/1hxZoLevRQYtvRqJx47nCCSyz-zvyCbcJ?usp=sharing>.

CONSENTIMENTO LIVRE E ESCLARECIDO:

Após ter recebido esclarecimentos sobre a natureza da pesquisa, seus objetivos, métodos, benefícios previstos, potenciais riscos e o incômodo que esta possa acarretar, aceito participar e declaro ter recebido este documento assinado pelos pesquisadores por meio do link <https://drive.google.com/drive/folders/1hxZoLevRQYtvRqJx47nCCSyz-zvyCbcJ?usp=sharing> e que manifesto a minha concordância em participar desta pesquisa:

Nome do(a) participante da pesquisa: _____

Contato telefônico/e-mail: _____

Data: _____

Responsabilidades dos Pesquisadores:

Asseguramos ter cumprido as exigências da resolução 466/2012 CNS/MS e complementares na elaboração do protocolo e na obtenção deste Termo de Consentimento Livre e Esclarecido. Asseguramos, também, ter explicado e fornecido uma via deste documento ao participante da pesquisa. Informamos que o estudo foi aprovado pelo CEP perante o qual o projeto foi apresentado e pela CONEP, quando pertinente. Comprometemo-nos a utilizar o material e os dados obtidos nesta pesquisa exclusivamente para as finalidades previstas neste documento ou conforme o consentimento dado pelo participante.

Prof. Dr. Celmar Guimarães da Silva	Data
-------------------------------------	------

Miguel Angelo Mena Póvoa	Data
--------------------------	------

Samuel Oliveira da Silva	Data
--------------------------	------

Apêndice F

Formulário com questionários sobre a ferramenta

Segue abaixo o formulário com questionários sobre a ferramenta como forma de avaliar se as visualizações propostas auxiliam, do ponto de vista dos usuários, no processo de seleção de modelos representativos.

Visualização de múltiplas curvas de risco de campos e poços como auxílio na seleção de modelos representativos de reservatórios de petróleo (Etapa de teste remoto de sistema)

Contexto:

Esse projeto visa propor e/ou aprimorar visualizações interativas de múltiplas curvas de risco para ajudar os analistas de petróleo a efetuarem melhores escolhas de conjuntos de modelos representativos (MRs) de reservatórios de petróleo em cenários com múltiplas estratégias, soluções e variáveis.

Trabalhar com modelos representativos de cenários, em vez de trabalhar com um conjunto completo de centenas de modelos, tem por objetivo tornar viável a simulação computacional de diferentes cenários e estratégias de desenvolvimento e gerenciamento de reservatórios de petróleo. Porém, se a escolha não for adequada, ela pode conduzir a tomadas de decisão inapropriadas, que afetarão a forma como se utilizará esses reservatórios. O projeto em questão se foca justamente em prover um suporte adequado para possibilitar melhores tomadas de decisão, considerando a quantidade de dados envolvida para isso

Ao concordar em participar do estudo, você colaborará com um teste remoto. Nesse teste os usuários utilizarão um protótipo de software para executar algumas tarefas relativas à

escolha de modelos representativos, como forma de avaliar se as visualizações propostas auxiliam, do ponto de vista desses usuários, esse processo de escolha. Suas opiniões serão registradas, analisadas e sumarizadas, visando aprimoramento das visualizações.

Identificação:

* Indica uma pergunta obrigatória

Nome: * _____

E-mail: * _____

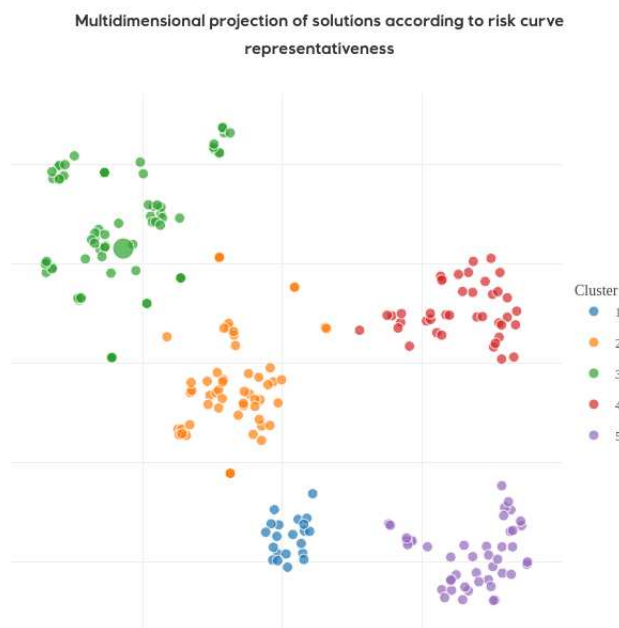
Questionário:

Os usuários executarão no protótipo desenvolvido algumas atividades relativas à escolha de modelos representativos, como forma de avaliar se as visualizações propostas auxiliam, do ponto de vista desses usuários, nesse processo de escolha. Suas opiniões serão registradas, analisadas e sumarizadas, visando aprimoramento das visualizações.

A seguir são apresentadas algumas afirmações para as quais você pode expressar o seu nível de concordância. Ao final do formulário, existem perguntas a serem respondidas textualmente de forma livre.

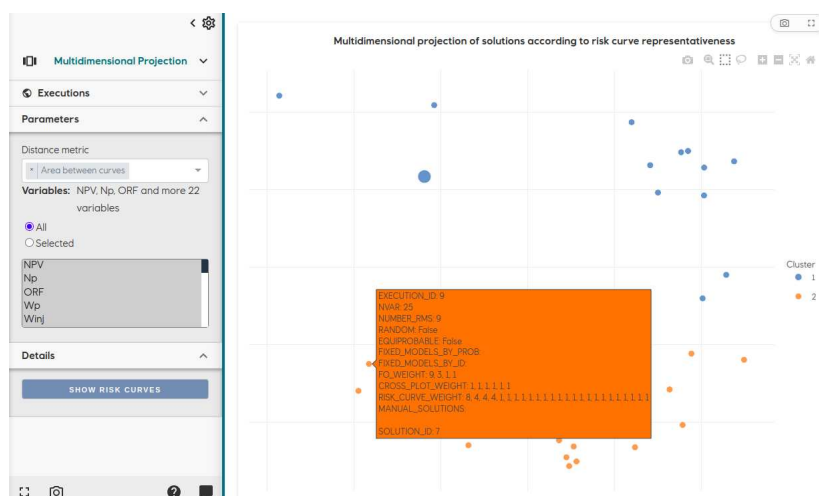
* Indica uma pergunta obrigatória

1. O gráfico “projeção multidimensional de soluções de acordo com a representatividade da curva de risco” possibilita ao usuário identificar a similaridade das soluções em relação ao conjunto de modelos originais (ou melhor, o conjunto de modelos utilizados como entrada do RMFinder). *



Discordo totalmente 1 2 3 4 5 Concordo totalmente

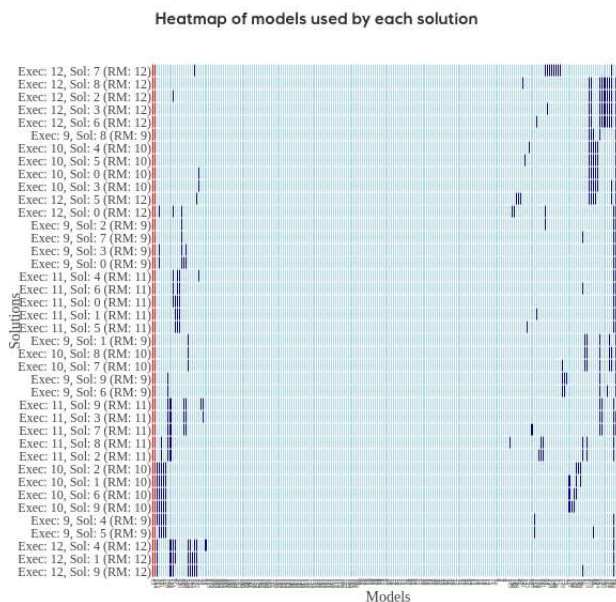
2. Os recursos interativos (filtragens, zoom, parâmetros e texto em foco), disponíveis no gráfico de projeção multidimensional, auxiliam o usuário a observar de várias formas as soluções de acordo com a representatividade da curva de risco, em busca de novas informações (como por exemplo descobrir que os dados se relacionam de uma forma que poderia estar oculta em sua representação visual inicial). *



1 2 3 4 5

Discordo totalmente ○ ○ ○ ○ ○ Concordo totalmente

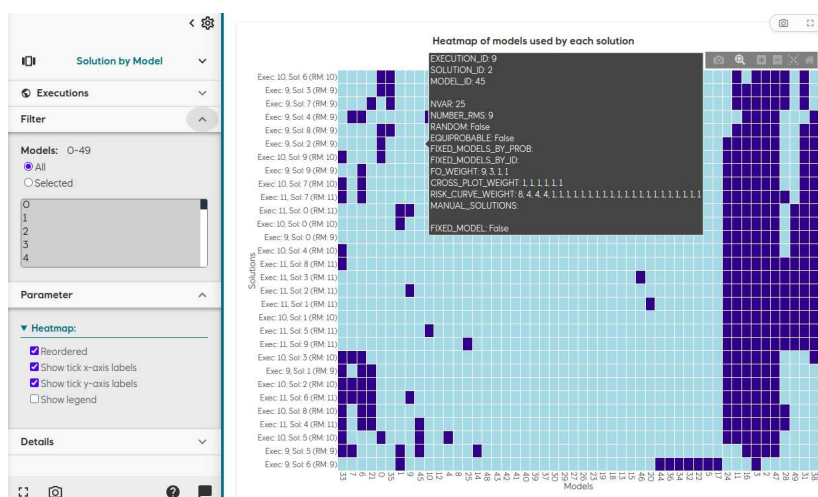
3. O mapa de calor de modelos usados para cada solução contribui para o usuário encontrar padrões relevantes em conjuntos de soluções com relação à sua variabilidade, como soluções com subconjuntos comuns de modelos representativos. *



1 2 3 4 5

Discordo totalmente ○ ○ ○ ○ ○ Concordo totalmente

4. Os recursos interativos (filtragens, *zoom*, parâmetros, texto em foco e reordenação), disponíveis no mapa de calor de modelos usados em cada solução, auxiliam o usuário a observar de várias formas a variabilidade das soluções (como subconjuntos comuns de modelos representativos) em busca de novas informações (como por exemplo descobrir que os dados se relacionam de uma forma que poderia estar oculta em sua representação visual inicial). *



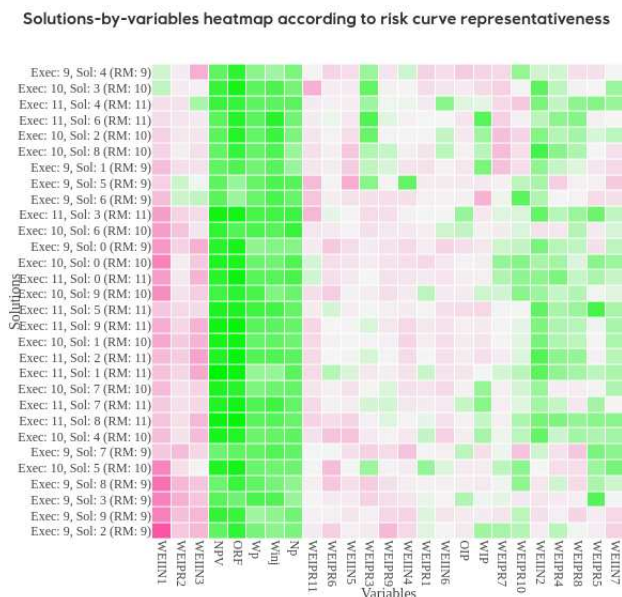
1 2 3 4 5
Discordo totalmente ○ ○ ○ ○ ○ Concordo totalmente

5. O gráfico “mapa de calor de soluções por variáveis de acordo com a representatividade da curva de risco” permite ao usuário analisar a representatividade de uma dada solução com relação a várias variáveis. *



1 2 3 4 5
Discordo totalmente ○ ○ ○ ○ ○ Concordo totalmente

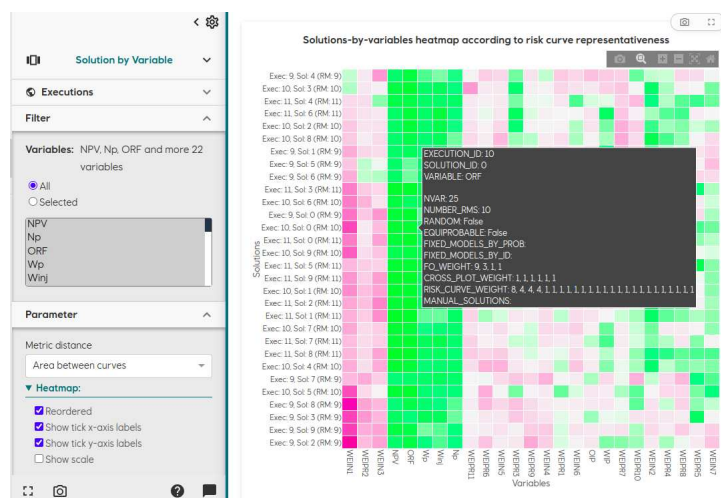
6. O gráfico “mapa de calor de soluções por variáveis de acordo com a representatividade da curva de risco” auxilia o usuário a comparar soluções com números distintos de modelos representativos e observar quais variáveis de produção provocam mais ou menos erro (ou seja, quais variáveis precisam de mais modelos representativos para serem representadas corretamente). *



1 2 3 4 5

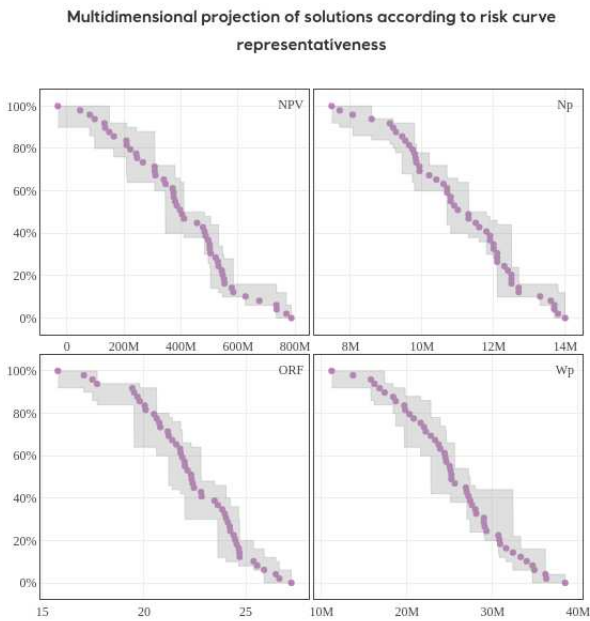
Discordo totalmente ○ ○ ○ ○ ○ Concordo totalmente

7. Os recursos interativos (filtragens, *zoom*, parâmetros, texto em foco e reordenação), disponíveis no mapa de calor de soluções por variáveis de acordo com a representatividade da curva de risco, auxiliam o usuário a observar de várias formas o erro de cada solução quando comparado ao conjunto completo de modelos (ou melhor, o conjunto de modelos utilizados como entrada do RMFinder) em busca de novas informações (como por exemplo descobrir que os dados se relacionam de uma forma que poderia estar oculta em sua representação visual inicial). *



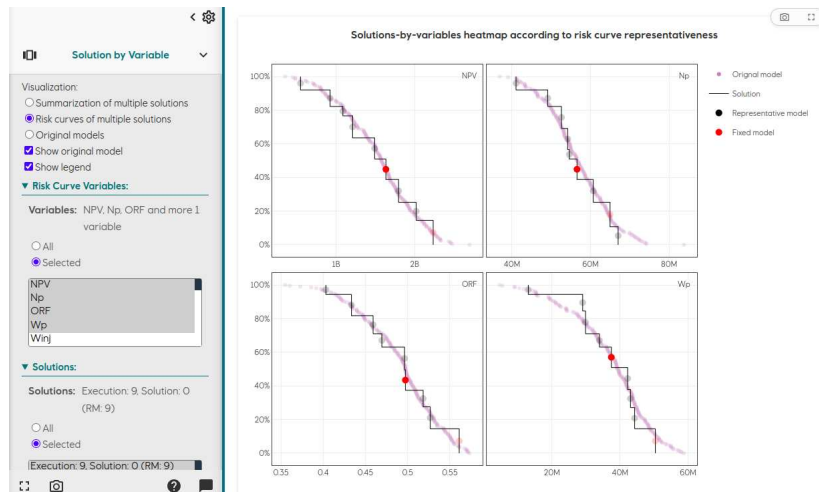
1 2 3 4 5
Discordo totalmente ○ ○ ○ ○ ○ Concordo totalmente

8. Os gráficos de curvas de risco (que mostram uma curva por variável) com intervalos de confiança que resumem o conjunto de curvas de risco de várias soluções junto ao conjunto de modelos originais (ou melhor, o conjunto de modelos utilizados como entrada do RMFinder), ajudam os usuários na análise da representatividade das soluções baseada em sua curva de risco e os seus respectivos detalhes. *



1 2 3 4 5
Discordo totalmente ○ ○ ○ ○ ○ Concordo totalmente

9. Os recursos interativos (filtragens, *zoom*, parâmetros, texto em foco e seleção de pontos entre gráficos), disponíveis nos gráficos de curvas de risco justapostas e sobrepostas, auxiliam o usuário de várias formas na seleção do melhor conjunto de modelos representativos em busca de novas informações (como por exemplo descobrir que os dados se relacionam de uma forma que poderia estar oculta em sua representação visual inicial). *



1 2 3 4 5

Discordo totalmente ☐ ☐ ☐ ☐ ☐ Concordo totalmente

10. O conjunto de visualizações interativas de múltiplas curvas de risco do protótipo avaliado ajuda o usuário a efetuar melhores escolhas de conjuntos de modelos representativos de reservatórios de petróleo em cenários com múltiplas estratégias, soluções e variáveis. *

1 2 3 4 5

Discordo totalmente ☐ ☐ ☐ ☐ ☐ Concordo totalmente

11. Escreva sobre os pontos positivos que avaliou com pontuação igual ou maior que 4:

12. Escreva sobre os pontos a melhorar que avaliou com pontuação igual ou menor que 3:

Apêndice G

Modelo Entidade-Relacionamento

Na Figura G.1 é apresentado o MER que foi desenvolvido para armazenar os dados brutos dos conjuntos de dados de entrada e saída do RMFinder, com intuito de transformá-los em tabela de dados utilizadas nas estruturas visuais desse trabalho.

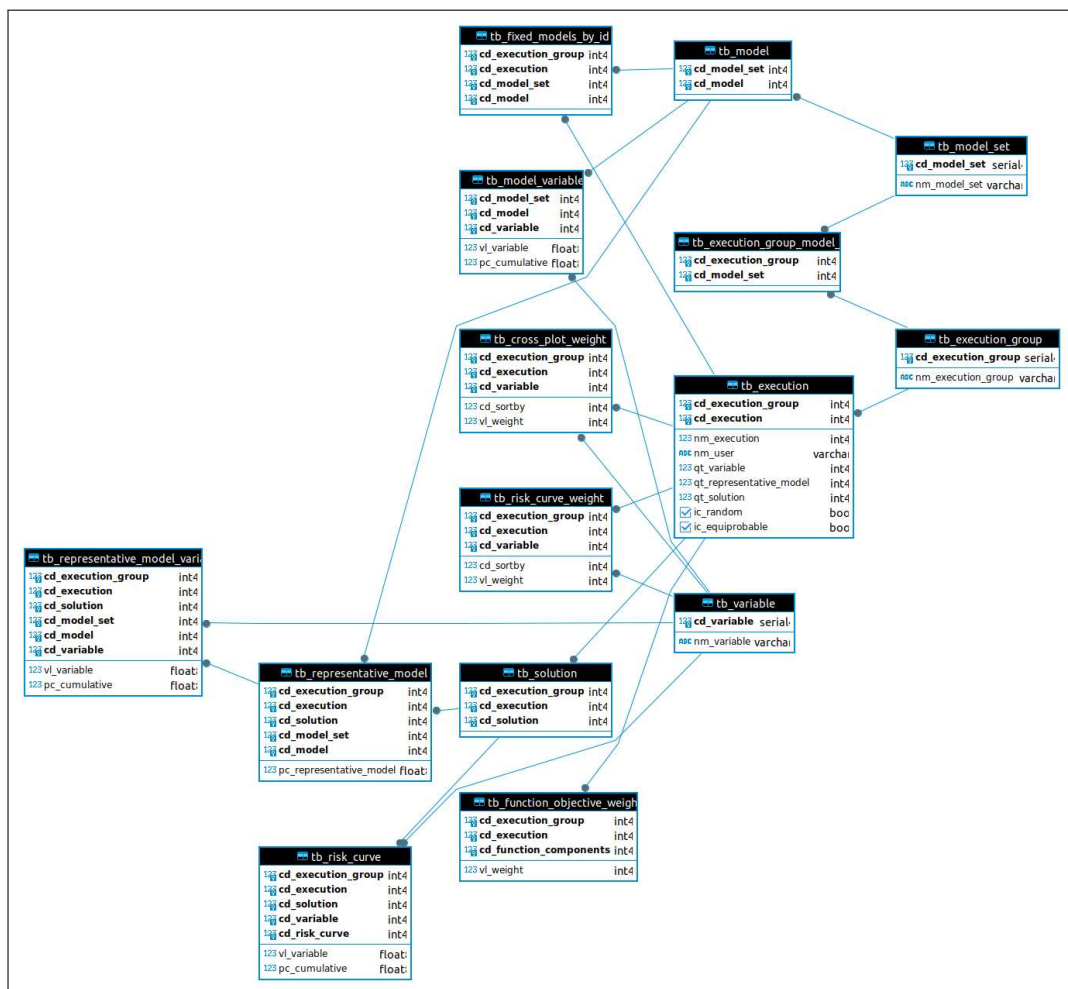


Figura G.1: MER dos conjuntos de dados de entrada e saída do RMFinder.

Fonte: Dados originais da pesquisa.

Apêndice H

t-SNE

A Tabela H.1 apresenta os parâmetros utilizados neste trabalho para o algoritmo de projeção multidimensional t-SNE.

Tabela H.1: Parâmetros do algoritmo de projeção multidimensional t-SNE.

Parâmetro	Descrição	Valor
n_components	Dimensão do espaço incorporado.	2
init	Inicialização da incorporação.	random
n_iter	Número máximo de iterações para a otimização.	500
n_iter_without_progress	Número máximo de iterações sem progresso antes de interromper a otimização.	150
n_jobs	O número de processos paralelos a serem executados para a pesquisa de vizinhos.	2
metric	A métrica a ser usada ao calcular a distância entre os componentes em uma matriz numérica. Se a métrica for “precomputed”, os dados de entrada devem representar uma matriz de distância.	precomputed
random_state	Determina o gerador de números aleatórios.	0
perplexity	A perplexidade está relacionada ao número de vizinhos mais próximos. A perplexidade deve ser menor que a quantidade total de amostras.	O menor valor entre a quantidade total de amostras do conjunto de dados e o valor 30.

Apêndice I

X-means

A Tabela I.1 apresenta os parâmetros utilizados neste trabalho para o algoritmo de *clustering X-means*.

Tabela I.1: Parâmetros do algoritmo de agrupamento *X-means*.

Parâmetro	Descrição	Valor
initial_centers	Coordenadas iniciais dos centros dos clusters.	O algoritmo de clustering K-means, com k=2 (quantidade de centros iniciais), foi utilizado para determinar as coordenadas dos centros dos clusters a partir dos quais o X-Means iniciará a análise.
kmax	Número máximo de clusters que podem ser alocados (ou divididos).	10
tolerance	Condição de parada para cada iteração: se o valor máximo da alteração dos centros dos clusters for menor que a tolerância, o algoritmo interromperá o processamento.	0,001
criterion	Tipo de criação de divisões (ou partições).	O Bayesian information criterion (BIC) foi utilizado para aproximar o número correto de clusters.