



Isabella Fischer Guindani Vieira

**Redução de Cenários com Formulação de
Cobertura de Conjuntos: Uma Aplicação na
Indústria de Petróleo**

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Engenharia de Produção, do Departamento de Engenharia Industrial da PUC-Rio.

Orientador: Prof. Rafael Martinelli Pinto

Rio de Janeiro,
abril de 2021



Isabella Fischer Guindani Vieira

**Redução de Cenários com Formulação de
Cobertura de Conjuntos: Uma Aplicação na
Indústria de Petróleo**

Dissertação apresentada como requisito parcial para
obtenção do grau de Mestre pelo Programa de Pós-
graduação em Engenharia de Produção da PUC-Rio.
Aprovada pela Comissão Examinadora abaixo.

Prof. Rafael Martinelli Pinto

Orientador

Departamento de Engenharia Industrial - PUC-Rio

Prof. Silvio Hamacher

Departamento de Engenharia Industrial - PUC-Rio

Prof. Fernando Luiz Cyrino Oliveira

Departamento de Engenharia Industrial - PUC-Rio

Rio de Janeiro, 12 de abril de 2021

Todos os direitos reservados. A reprodução, total ou parcial, do trabalho é proibida sem autorização da universidade, da autora e do orientador.

Isabella Fischer Guindani Vieira

Graduou-se em Engenharia de Produção pela Universidade Federal do Rio de Janeiro (UFRJ), em 2018. No mesmo ano, ingressou no Programa de Pós-Graduação em Engenharia Industrial com ênfase em Transporte e Logística da PUC-Rio, para obtenção do título de Mestre.

Atualmente trabalha no setor de Tecnologia de Informação do Instituto Tecgraf – PUC-Rio, mais especificamente na área de desenvolvimento de software.

Ficha Catalográfica

Vieira, Isabella Fischer Guindani

Redução de cenários com formulação de cobertura de conjuntos: uma aplicação na indústria de petróleo / Isabella Fischer Guindani Vieira ; orientador: Rafael Martinelli Pinto. – 2021.

97 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2021.

Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Seleção de cenários. 3. Modelo de cobertura de conjuntos. 4. Programação inteira. I. Pinto, Rafael Martinelli. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. III. Título.

CDD: 658.5

À minha mãe Flávia
e minha avó Edléa, por me
ensinaram tudo que
verdadeiramente importa.

Agradecimentos

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

Ao meu orientador Professor Rafael Martinelli pelo estímulo e parceria para a realização deste trabalho.

À Capes e à PUC-Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

À minha família e amigos, em especial minha mãe Flávia pela paciência e compreensão nos meus momentos de ausência, e ao Junior pela orientação, auxílio técnico e, principalmente, pelo apoio nos momentos difíceis na pandemia.

Aos meus colegas da PUC-Rio e do Instituto Tecgraf.

Aos professores que participaram da Comissão examinadora.

A todos os professores e funcionários do Departamento de Engenharia Industrial da PUC-Rio e da Molde University College, pelos ensinamentos e pela ajuda.

A todos os amigos e familiares que de uma forma ou de outra me estimularam ou me ajudaram a alcançar mais este objetivo.

Resumo

Vieira, Isabella Fischer Guindani; Pinto, Rafael Martinelli. **Redução de Cenários com Formulação de Cobertura de Conjuntos: Uma Aplicação na Indústria de Petróleo**. Rio de Janeiro, 2021. 97p. Dissertação de Mestrado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

As técnicas de agrupamentos aplicadas a um grande número de cenários de incerteza permitem a escolha de um conjunto reduzido, porém, representativo da população de cenários completa. Em outras palavras, selecionar uma amostra que contenha uma quantidade menor de elementos a ponto de reduzir suficientemente o volume total de dados e obter ganhos significativos de eficiência no processamento dos dados. Esta amostra deve, sobretudo, conseguir preservar as características do processo estocástico que o originou. Com este intuito, o presente trabalho propõe uma metodologia de seleção de cenários estocásticos utilizando o modelo clássico de Cobertura de Conjuntos, inspirada no método *forward selection* proposto por Heitsch & Romisch (2003). Aplicada na etapa de cálculo de demanda estocástica de ferramentas e serviços para construção de poços marítimos de exploração de petróleo, esta abordagem apresenta uma concepção de cenário diferente da usada pelos autores. O conjunto de cenários consiste em cronogramas de atividades gerados a partir da introdução de incertezas no planejamento de cada atividade, sendo eles estáticos, independentes e com múltiplos atributos. Uma análise de sensibilidade compara os resultados das demandas calculadas com os cenários selecionados pelo Problema de Cobertura de Conjuntos (PCC) e a demanda calculada com o conjunto universo de cenários. O PCC foi solucionado, nesta aplicação, em sua versão clássica da literatura a partir de um algoritmo exato e um heurístico. Os resultados apontam diferenças pouco representativas no resultado final das demandas calculadas com cenários reduzidos e com o total de cenários. A heurística, ainda que seja *first solution*, apresentou um resultado satisfatório em relação ao ganho de desempenho *versus* confiabilidade, e indica o potencial do método se aplicado em conjunto com algoritmos de metaheurística e busca local.

Palavras-chave

Seleção de Cenários; Modelo de Cobertura de Conjuntos; Programação Inteira.

Abstract

Vieira, Isabella Fischer Guindani; Pinto, Rafael Martinelli (Advisor). **Scenario Reduction with Set Covering Formulation: An Application in the Oil Industry**. Rio de Janeiro, 2021. 97p. Dissertação de Mestrado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Clustering techniques applied to a large number of scenarios under uncertainty allows the selection of a reduced, however, representative set of the complete set of scenarios. In other words, it allows to select a sample that contains a smaller amount of elements to the point of sufficiently reducing the total data volume and obtaining efficiency gains in data processing. The challenge is that the sample must, above all, be able to preserve the characteristics of the stochastic process that originated it. To this end, this study proposes a methodology for selecting stochastic scenarios using the classic Set Covering model, inspired by the forward selection method proposed by Heitsch & Romisch (2003). Applied in the calculating of stochastic demand for tools and services for the construction of offshore oil exploration wells, this approach presents a different scenario conception from the one used by the authors. The set of scenarios consists of activity schedules generated from the introduction of uncertainties in the planning of each activity, which are static, independent and with multiple attributes. A sensitivity analysis compares the results of the demands calculated with the scenarios selected by the Set Covering Problem (SCP) and the demand calculated with all the universe of scenarios. The SCP was solved, in this application, in its classic version using an exact algorithm and a heuristic algorithm. The results appoint na unexpressive loss in the final result of the demand calculated with reduced scenarios and with the complete set of scenarios. The simple first solution heuristic presented a satisfactory result in relation to the performance gain versus reliability, and indicates the potential of the method if solved with metaheuristic and local search algorithms.

Keywords

Scenario Selection; Set Covering Model; Integer Programming.

Sumário

1	Introdução.....	14
1.1	Objetivo.....	16
1.2	Estrutura da dissertação.....	17
2	Revisão Bibliográfica	18
2.1	Seleção de cenários.....	18
2.1.1	Conceito de Incerteza e Risco	18
2.1.2	Características Gerais	20
2.1.3	Métodos Hierárquicos	22
2.1.4	Métodos Particionais.....	24
2.2	Problema da Cobertura de Conjuntos	29
2.2.1	Definição matemática	29
2.2.2	Métodos Heurísticos para solução do PCC.....	31
3	Aplicação	34
3.1	Análise preliminar do problema	34
3.1.1	Planejamento de operações para construção de poços marítimos de produção de petróleo.....	34
3.1.2	Cálculo da demanda por ferramentas e serviços para construção de poços.....	37
3.1.3	Análise da demanda frente à quantidade de cenários.....	39
3.2	Caracterização genérica do problema	44
3.3	Trabalhos Relacionados.....	46
3.4	A metodologia aplicada	49
3.5	Limitações da Pesquisa.....	56
4	Análise de Resultados	57
4.1	Construção computacional	57
4.2	Método exato	58

4.3 Método heurístico.....	64
4.4 Comparação entre os métodos	68
5 Considerações Finais	74
5.1 Sugestões para desenvolvimentos futuros	75
Bibliografia.....	77
Apêndice	82

Lista de Figuras

Figura 1 - Diagrama dos métodos de agrupamento hierárquicos	23
Figura 2 - Dendograma dos estados americanos utilizando ligação simples. Fonte: https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/	24
Figura 3 - Estágios do algoritmo K-means. Fonte: Chen & Lai (2018)	27
Figura 4 - Esquemático de um cronograma determinístico.	35
Figura 5 – Esquemático de um cronograma determinístico e cronogramas simulados gerados a partir da aplicação das incertezas nas operações da careira de uma sonda de perfuração	38
Figura 6 - Estrutura de <i>fan</i> convertida em árvore de cenários. Fonte: Hamacher & Oliveira, 2013.....	48
Figura 7 - Etapas da metodologia utilizada no estudo.....	49
Figura 8 - Diagrama da intensidade da poda frente ao valor do cutoff	52
Figura 9 - <i>Subsets</i> de cada elemento para diferentes valores de <i>cutoff</i>	52
Figura 10 - <i>Subsets</i> dos elementos do grafo podado	53
Figura 11 - Elementos representativos do conjunto total de vértices.....	55
Figura 12 - Demandas médias diárias calculadas com os cenários selecionados por cada <i>cutoff</i> , para a ferramenta A.....	60
Figura 13 - Desvios padrões diários das demandas divididos pelas demandas médias diárias calculadas com os cenários selecionados por cada <i>cutoff</i> , para a ferramenta A.....	61
Figura 14 - Distâncias absolutas das demandas médias diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> de valor até 0.1 e a demanda média diária calculada com todos os cenários, no método exato, para a ferramenta A	62
Figura 15 - Distâncias relativas das demandas médias diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> de valor até 0.1 e a demanda média diária calculada com todos os cenários, no método exato, para a ferramenta A	63
Figura 16 - Demandas médias diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> até 0.1 no método heurístico, para a ferramenta A.....	65
Figura 17 - Desvio padrão dividido pela média diário das demandas calculadas com os cenários selecionados pelos <i>cutoffs</i> até 0.1 no método heurístico, para a ferramenta A	66

Figura 18 - Distâncias absolutas das demandas médias diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> de valor até 0.1 e a demanda média diária calculada com todos os cenários, no método heurístico, para a ferramenta A	66
Figura 19 - Distâncias relativas das demandas médias diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> de valor até 0.1 e a demanda média diária calculada com todos os cenários, no método heurístico, para a ferramenta A	67
Figura 20 - Demanda média encontrada em cada método, para cutoff = 0.025	68
Figura 21 - Demanda média encontrada em cada método, para cutoff = 0.05	68
Figura 22 - Demanda média encontrada em cada método, para cutoff = 0.075	69
Figura 23 - Demanda média encontrada em cada método, para cutoff = 0.1	69
Figura 24 - Demanda média encontrada em cada método, para cutoff = 0.15	69
Figura 25 - Demanda média encontrada em cada método, para cutoff = 0.2	69
Figura 26 - Demanda média encontrada em cada método, para cutoff = 0.25	69
Figura 27 - Distâncias relativas entre as médias das demandas calculadas para cada <i>cutoff</i> em cada método e a média da demanda calculada com todos os cenários	71
Figura 28 - Demanda média diária entre os cronogramas, por ferramenta, por quantidade de cronogramas	83
Figura 29 - Desvio padrão médio diário dividido pela média diária entre os cronogramas por ferramenta, por quantidade de cronogramas	84
Figura 30 - Distância absoluta entre as demandas diárias médias de todas as quantidades de cronogramas candidatas e a demanda diária média de 3000 cronogramas	85
Figura 31 - Distância relativa entre as demandas diárias médias de todas as quantidades de cronogramas candidatas e a demanda diária média de 3000 cronogramas	86
Figura 32 - Média diária das 10 demandas gerados com amostras aleatórias de 500 cronogramas	87
Figura 33 - Gráficos das demandas médias diárias calculadas com os cenários selecionados por cada <i>cutoff</i> no método exato, por ferramenta	88
Figura 34 - Gráficos dos desvios padrões divididos pelas médias diárias das demandas calculadas com os cenários selecionados por cada <i>cutoff</i> no método exato, por ferramenta.	89
Figura 35 - Demandas médias diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> de valor até 0.1 no método exato, para a ferramenta A	90
Figura 36 - Desvio padrão médio das demandas calculadas com os cenários selecionados pelos <i>cutoffs</i> de valor até 0.1 no método exato, para a ferramenta A	90
Figura 37 - Desvio padrão dividido pela média diário das demandas calculadas com os cenários selecionados pelos <i>cutoffs</i> de valor até 0.1 no método exato, para ferramenta A	90

Figura 38 - Percentil 10 das demandas diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> de valor até 0.1 no método exato, para a ferramenta A	91
Figura 39 - Mediana (percentil 50) das demandas diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> de valor até 0.1 no método exato, para a ferramenta A	91
Figura 40 - Percentil 90 das demandas diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> de valor até 0.1 no método exato, para a ferramenta A	91
Figura 41 - Variâncias das demandas diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> de valor até 0.1, para a ferramenta A.....	92
Figura 42 - Gráficos das demandas médias diárias calculadas com os cenários selecionados por cada <i>cutoff</i> no método heurístico, por ferramenta	93
Figura 43 - Gráficos dos desvios padrões diários dividido pelas médias diárias das demandas calculadas com os cenários selecionados por cada <i>cutoff</i> no método heurístico, por ferramenta.	94
Figura 44 - Desvio padrão médio das demandas calculadas com os cenários selecionados pelos <i>cutoffs</i> até 0.1 no método heurístico, para a ferramenta A	95
Figura 45 - Percentil 10 das demandas diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> até 0.1 no método heurístico, para a ferramenta A	95
Figura 46 - Mediana (percentil 50) das demandas diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> até 0.1 no método heurístico, para a ferramenta A	96
Figura 47 - Percentil 90 das demandas diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> até 0.1 no método heurístico, para a ferramenta A	96
Figura 48 - Variâncias das demandas diárias calculadas com os cenários selecionados pelos <i>cutoffs</i> até 0.1 no método heurístico, para a ferramenta A	96

Lista de Tabelas

Tabela 1 - Distâncias relativas médias das demandas com diferentes quantidades de cronogramas para todo o horizonte de simulação, por indicador.	42
Tabela 2 - Distâncias relativas da média diária entre cada par de demandas geradas a partir das 10 amostras de 500 cronogramas aleatórios	43
Tabela 3 - Palavras-chaves da busca por trabalhos correlatos	46
Tabela 4 - Resultados da busca por trabalhos correlatos no repositório de trabalhos científicos Scopus (realizada em 04/08/2021).....	47
Tabela 5 - Execuções dos algoritmos com os valores de <i>cutoff</i>	58
Tabela 6 - Quantidades de cenários selecionados pelo algoritmo exato para cada <i>cutoff</i> e seus tempos de execução	59
Tabela 7 - Distâncias absolutas médias das curvas referentes a cada indicador, por <i>cutoff</i> no método exato, para a ferramenta A	62
Tabela 8 - Distâncias relativas médias das curvas referentes a cada indicador, por <i>cutoff</i> no método exato, para a ferramenta A.....	63
Tabela 9 - Quantidades de cenários selecionados pelo algoritmo heurístico para cada <i>cutoff</i> e seus tempos de execução	64
Tabela 10 - Distâncias absolutas médias das curvas referentes a cada indicador, por <i>cutoff</i> no método heurístico, para a ferramenta A	66
Tabela 11 - Distâncias relativas médias das curvas referentes a cada indicador, por <i>cutoff</i> no método heurístico, para a ferramenta A	67
Tabela 12 - Médias totais dos resultados dos indicadores da Ferramenta A em cada método aplicado e demanda de referência, por valor de <i>cutoff</i>	70
Tabela 13 - Médias totais das distâncias absolutas e relativas dos indicadores média, desvio padrão, desvio padrão dividido pela média e variância para cada <i>cutoff</i> , separados por método aplicado.	72

1

Introdução

O planejamento de operações refere-se ao processo de projetar no futuro o objetivo desejado e determinar quais atividades e recursos são necessários para alcançá-lo, baseando-se somente em informações e premissas disponíveis no momento da decisão de planejar. No entanto, existem riscos externos envolvidos no processo que não podem ser mapeados de maneira determinística e incrementados ao processo, podendo ameaçar o sucesso do planejamento. Esses riscos podem ser eliminados, mitigados ou contingenciados através de uma gestão de riscos apropriada (VALLE et al., 2014). Há riscos difíceis de serem eliminados, como por exemplo incertezas quanto às durações efetivas de cada atividade planejada, exigindo a adoção de estratégias de mitigação. Uma das formas de minimizar alguns riscos e ainda trazer flexibilidades gerenciais é a aplicação de análises estocásticas de cenários de incerteza.

Por isso, o tratamento da estocasticidade está cada vez mais presente nos diversos ramos industriais, tais como petróleo, energia, telecomunicações, meteorologia, finanças, dentre outras (RASMUSSEN & CLAUSEN, 2007). Nas últimas décadas, foi crescente a quantidade de temas relacionados a incertezas que surgiram na indústria, sendo a otimização estocástica uma das mais importantes técnicas.

Os métodos de otimização estocástica têm por objetivo modelar incertezas de modo a fornecer subsídios para uma melhor tomada de decisão diante de um cenário. As incertezas são modeladas através da inserção de variáveis aleatórias em um modelo de otimização, e quase sempre se faz necessário algum tipo de simplificação do problema. Assim, para obter uma solução numérica aproximada, a função de densidade de probabilidade é discretizada e o processo estocástico original aproximado por uma árvore de cenários, um conjunto de possíveis realizações futuras do evento incerto. Em linhas gerais, para que as incertezas sejam melhor reproduzidas, aumenta-se o número de cenários criados, mas em contrapartida eleva-se muito o tempo de processamento computacional

(MACEIRA et al., 2005). Sendo a capacidade computacional um recurso limitado em qualquer situação real, este *tradeoff* é o grande desafio na aplicação da técnica.

Por não escaparem ao desafio de manter o equilíbrio entre a representatividade do processo estocástico e esforço computacional na busca por boas soluções, a construção de árvores de cenários visando a utilização de um número menor de cenários se tornou um assunto fundamental em otimização estocástica e são encontradas diferentes estratégias de solução na literatura. Uma opção clássica consiste em realizar uma simulação baseada em técnicas de amostragem como o *Latin Hypercube Sampling* (MCKAY et al., 1979), *Quasi-Monte Carlo* (NIEDERREITER, 1992), Monte Carlo (CHIRALAKSANAKUL, 2003), e Amostragem Seletiva (PENNA et al., 2011). Outra opção é o uso de métricas probabilísticas como a distâncias de Wasserstein/Kantorovich (DUPAČOVÁ et al., 2003) e a distância aninhada (PFLUG, 2009).

As técnicas mais encontradas na literatura para problemas de seleção de cenários, são os métodos de agrupamento hierárquicos, como o de Ligação, Centroide e Ward, e métodos de agrupamento particionais, como o de clusterização através do algoritmo K-Means e K-Medoids. Neles, os elementos são divididos em *clusters* (grupos), sendo cada elemento pertencente a um único *cluster*, com um ponto central ou elemento central representante dos demais. Assim, a quantidade de representantes do conjunto completo será igual a quantidade de *clusters* criados, podendo ser um número bem reduzido quando comparado ao todo, à critério de quem aplica o método.

Heitsch & Romisch (2003) apresenta 4 métodos, também seguindo o objetivo de redução de cenários. Os resultados identificados pelos autores como mais promissores são da aplicação da metodologia que utiliza o clássico problema de Cobertura se Conjuntos (Set Covering), da programação inteira. Neste método, foram calculadas as distâncias entre cada par de cenários e selecionado o menor subconjunto de cenários que cobre o conjunto universo de cenários, dada uma distância mínima admissível. Um ponto positivo deste método é que o cálculo da distância pode ser feito com diversos algoritmos e o problema de otimização combinatória pode ser solucionado através de heurísticas e metaheurísticas dos mais diversos tipos, atribuindo uma grande capacidade de personalização do método e utilização em diversas abordagens.

Aproveitando desta capacidade de personalização, é utilizado neste estudo a metodologia para seleção de cenários estocásticos utilizando o modelo de Cobertura de Conjuntos apresentada por Heitsch & Romisch (2003), com um conceito de cenário diferente do utilizado pelos autores. São utilizados como cenários os cronogramas estocásticos de operações de construção de poços marítimos e almeja-se selecionar o menor subconjunto que represente o conjunto completo de cronogramas. A capacidade de representação de um cronograma por outro é condicionada à semelhança entre os cronogramas, dada pela distância calculada entre eles e um valor de corte que define se um par é próximo ou distante, semelhante ou diferente.

Para validar a metodologia é realizado um estudo de caso para a etapa de cálculo de demanda de ferramentas e serviços para construção de poços marítimos de exploração de petróleo, finalidade principal da criação dos cronogramas estocásticos. Uma análise de sensibilidade compara os resultados das demandas calculadas a partir dos cronogramas selecionados em uma heurística construtiva para o problema de Cobertura de Conjuntos e a partir da resolução matemática do mesmo modelo, em sua versão clássica da literatura, em face à demanda gerada com o conjunto completo de cronogramas.

1.1

Objetivo

O objetivo geral deste trabalho é aplicar a metodologia de seleção de cenários através da resolução de um Problema de Cobertura de Conjuntos, proposta por Heitsch & Romisch (2003) no conjunto de cronogramas estocásticos gerados para o cálculo de demanda de ferramentas e serviços para construção de poços marítimos.

Para tal, são traçados objetivos específicos, etapas menores que compõem o todo. São eles:

- Seleção e adequação do modelo matemático de PCC para o contexto de cenários de incerteza;

- Implementação computacional da metodologia, com os modelos exato e heurístico, e diferentes quantidades de cenários, de forma que a variação dos resultados funcione como uma análise de sensibilidade;
- Avaliação da metodologia proposta a partir dos resultados computacionais como ferramenta de pré-processamento no cálculo da demanda estocástica de ferramentas e serviços para construção de poços marítimos de exploração de petróleo;
- Comparação dos resultados com a demanda gerada a partir do conjunto completo de cenários.

1.2

Estrutura da dissertação

Este trabalho está organizado como descrito na sequência. O capítulo 2 introduz os conceitos de Seleção de Cenários e Problema de Cobertura de Conjuntos. Também discorre sobre o processo de construção dos cronogramas estocásticos para cálculo da demanda de ferramentas e serviços para construção de poços marítimos.

O capítulo 3 mostra em detalhes a análise preliminar do caso do cálculo da demanda estocástica e seus resultados, define o problema genericamente para a decisão dos métodos adotados e apresenta as etapas e os detalhes da implementação computacional da metodologia proposta.

No capítulo 4 são apresentados os resultados da aplicação metodológica com uma análise de sensibilidade e a comparação do resultado final da demanda com a demanda calculada com o conjunto universo de cenários.

Finalizando o trabalho, no capítulo 5 são descritas as principais considerações sobre o trabalho e sugestões para desenvolvimentos futuros.

2

Revisão Bibliográfica

Este capítulo trará uma breve discussão sobre o conceito de riscos e incertezas, introduzindo a abordagem da demanda estocástica e cronogramas estocásticos. Discorrerá sobre a seleção de cenários e os métodos de agrupamentos hierárquicos e Particionais. Também trará uma visão geral sobre o Problema de Cobertura de Conjuntos.

2.1

Seleção de cenários

Neste tópico serão apresentados os métodos para seleção de cenários encontrados mais recorrentes na literatura e que contribuiram para o entendimento do problema e para a escolha dos métodos utilizado neste estudo.

2.1.1

Conceito de Incerteza e Risco

Há muito é debatido na literatura a diferença entre riscos e incertezas em um projeto. Já dizia Knight (1921), que o risco está presente quando os eventos futuros ocorrem com probabilidade mensurável e a incerteza está presente quando a probabilidade de eventos futuros é indefinida ou incalculável.

Em Lanz e Tomei (2015), risco é definido como a possibilidade de ocorrência de um determinado evento ou resultado futuro, incerto, mas previsível, que pode afetar o agente. O risco pode ser quantificado em termos de probabilidade de ocorrência e impacto caso ocorra e, na maioria das vezes, independe da vontade do agente para sua ocorrência. Para Tenenbaum (2012), O risco é tangível, a incerteza não. Pode-se definir o risco, mas mal se consegue delinear as camadas exteriores de incerteza. O risco pode ser tornado concreto e mensurado, a incerteza não pode.

A partir de Gray e Larson (2000) podemos identificar 3 categorias de risco em projetos: (i) riscos de projeto, ligados a execução do projeto em si. Tem impacto sobre requisitos, tempo, custo, qualidade, recursos e pessoal; (ii) riscos técnicos ou de produto, ligados ao produto a ser desenvolvido, sua qualidade, design; e (iii) riscos de negócio, ligados a viabilidade de negócios do projeto e produto, associados à demanda no mercado, à concorrência, à disponibilidade de recursos e à estratégia da organização.

Na indústria de óleo e gás, uma característica importante dos problemas de planejamento, programação e *scheduling* é o alto grau de incerteza intrínseco ao negócio (Lababidi et al., 2004). Como citado em Ribas (2008 apud Ruczyński, 1997), neste setor existem inúmeros fatores cujos valores são incertos, e ignorá-los pode levar a uma decisão ruim ou, ainda, errada. Nesta dissertação, todos os três tipos de riscos mencionados acima serão considerados, e as incertezas serão tratadas como inerentes ao processo de determinação e cálculo dos riscos.

Definindo a demanda determinística como o caso em que as demandas de um conjunto de clientes são conhecidas a priori, pode-se chamar de demanda estocástica, resumidamente, os casos em que estas informações não são conhecidas ou não podem ser determinadas, preliminarmente. Este é o tipo de demanda que mais se observa na literatura quando consideradas aplicações reais, onde dificilmente será encontrado um problema em que se possa ignorar a variabilidade do sistema, seja qual for.

Laporte et al. (2002) apresentam um Problema de Roteamento de Veículos capacitado com demandas estocásticas (SVRP - do inglês *Stochastic Vehicle Routing Problem*), assim como em Shanmugam et al. (2011). He et al. (2009) investigam a questão da coordenação de canais para uma cadeia de suprimentos que enfrenta demanda estocástica sensível ao esforço de vendas e ao preço de varejo. Já no contexto da indústria petrolífera, Ribas et al. (2010) propõem um modelo de planejamento estratégico para uma cadeia de suprimentos integrada considerando demanda variável de derivados de petróleo, dentre outras fontes de incerteza.

Kisialiou et al. (2019) utilizam várias *recourse actions* executadas simultaneamente em várias combinações e destinadas a garantir a viabilidade do cronograma de embarcações de suprimentos offshore, que é constantemente afetado por condições meteorológicas incertas e demanda incerta nas instalações. Uma vez que estas incertezas resultam na reprogramação das atividades e que, no pior caso,

uma embarcação extra pode ser necessária e contratada a um custo mais alto, seu objetivo é definir uma composição ideal da frota e um plano semanal de menor custo usado repetidamente ao longo de uma temporada. Para tal, os autores apresentam uma metodologia de otimização-simulação para a construção de cronogramas de embarcações de suprimento e um algoritmo heurístico capaz de gerar soluções com algum nível de confiabilidade contra demanda e clima incertos, combinado com uma simulação de evento discreto para avaliar o desempenho do cronograma e calcular o custo esperado do recurso.

Salavati-Khoshghalb et al. (2018) consideram o problema de roteamento de veículos com demandas estocásticas (VRPSD - do inglês *Vehicle Routing Problem with Stochastic Demands*). As falhas ocorrem sempre que a capacidade residual do veículo é insuficiente para atender à demanda observada de um cliente, implicando na utilização de *recourse actions* para recuperar a viabilidade da rota. Um algoritmo de solução exata é desenvolvido para resolver o VRPSD e é realizado um extenso estudo computacional, que mostra que soluções significativamente melhores podem ser obtidas ao usar as políticas propostas em comparação com a solução do problema na definição clássica de recurso.

Heitsch & Romisch (2003) e Dupačová et al. (2003), abordam métodos de solução para a programação estocástica. O primeiro propõe os métodos *forward selection* e a *backward reduction* para redução de cenários. Em especial o primeiro método é de vital importância para a definição da metodologia aplicada neste estudo e será detalhado no capítulo 3 de aplicação.

2.1.2

Características Gerais

A primeira publicação registrada sobre um método de clusterização foi Sorensen (1948), que apresentou o Método Hierárquico de Ligação Completa. Desde então muitos métodos distintos de agrupamento de dados foram criados e geraram milhares de aplicações das mais diversas áreas. Qualquer método de agrupamento ou *clustering* consiste em agrupar um conjunto de observações de modo que as observações que pertençam a um mesmo grupo sejam parecidas ao máximo entre si e diferentes o suficiente das dos demais grupos para justificar não estarem todas em um único conjunto. As técnicas podem ser usadas para descobrir padrões e grupos naturais em conjuntos de dados sem ter qualquer conhecimento

prévio das características dos dados. Ou mesmo para descobrir estruturas em dados sem fornecer uma explicação/interpretação (KOGAN, 2007).

As técnicas de agrupamentos objetivam reduzir a cardinalidade dos dados ao agrupar um conjunto em subconjuntos, partindo de um critério fixado que pode variar ligeiramente em virtude do método de agrupamento utilizado, como classificação dos dados. O conceito de análise de agrupamentos reúne diferentes métodos e algoritmos para agrupar objetos do mesmo tipo nas respectivas categorias.

Segundo Hair (1998), os grupos obtidos devem apresentar tanto uma homogeneidade interna, dentro de cada grupo, como uma grande heterogeneidade externa, entre grupos. Portanto, se a agregação for bem-sucedida, ao serem representados graficamente, os elementos dentro dos grupos estarão muito próximos, e afastados dos elementos de grupos distintos.

Visto que o objetivo da análise de cluster é agrupar elementos semelhantes, é necessário um parâmetro de semelhança. Comumente este parâmetro é compreendido como uma medida da distância entre os elementos. Aqueles com menor distância entre si são mais semelhantes, e por isso são agregados em um mesmo grupo. E os mais distantes compõem grupos distintos. Existem várias formas de medir a distância entre os objetos, porém, a mais utilizada é a distância euclidiana.

Em análise de *clustering*, quando a representação de dados é feita com poucos clusters, perde-se muitos detalhes finos, semelhantemente ao processo de perda de dados por compressão, mas consegue-se simplificação, pois é possível representar muitos dados em um pequeno número de grupos. Pelo contrário, quando há um número grande de clusters, é feita uma categorização mais rica dos dados, porém a distância entre clusters pode ser tão sutil a ponto de a divisão dos clusters ser pouco representativa na análise (BERKHIN, 2002). Por este motivo, o número de clusters k é de vital importância para a definição do método, mesmo que, na maioria dos casos, o número ideal ou desejado de clusters é desconhecido.

Conforme Malhotra (2001), a aplicação da análise de agrupamentos é dividida em etapas, sendo a primeira delas a definição do problema de aglomeração e as variáveis a serem tratadas. Escolhe-se, então, uma medida de distância dos grupos. Em seguida, define-se o processo de agregação que dependerá das variáveis em estudo e do problema em foco. Os grupos resultantes devem ser interpretados

em termos das variáveis usadas para constituí-los e de outras variáveis adicionais importantes. Por fim, o processo de agregação deve ser validado.

Na classificação clássica, os métodos de agrupamento são comumente divididos em duas classes: hierárquicos e particionais. A seguir é apresentada uma breve descrição destes métodos.

2.1.3 Métodos Hierárquicos

Os métodos de agrupamento hierárquicos podem ser classificados em aglomerativos e divisivos. A abordagem aglomerativa parte de elementos individuais, agrupando-os conforme a medida de similaridade escolhida para refletir a proximidade entre os objetos. A cada etapa novos grupos são formados e por fim tem-se um único grupo, o nível mais alto da hierarquia, cabendo ao pesquisador definir um critério de parada do agrupamento conforme seu interesse. Na abordagem divisiva o processo é inverso: parte-se de um único grupo de elementos e, também segundo a medida de similaridade, segue-se separando os objetos em grupos até que cada grupo contenha somente um elemento, sendo também necessário que o pesquisador defina um critério de parada do agrupamento, que pode ser um número máximo de grupos, por exemplo.

Os métodos hierárquicos mais utilizados são:

- Ligação simples (ou método do vizinho mais próximo)
- Ligação completa (ou método do vizinho mais distante)
- Ligação média
- Centróide
- *Ward*

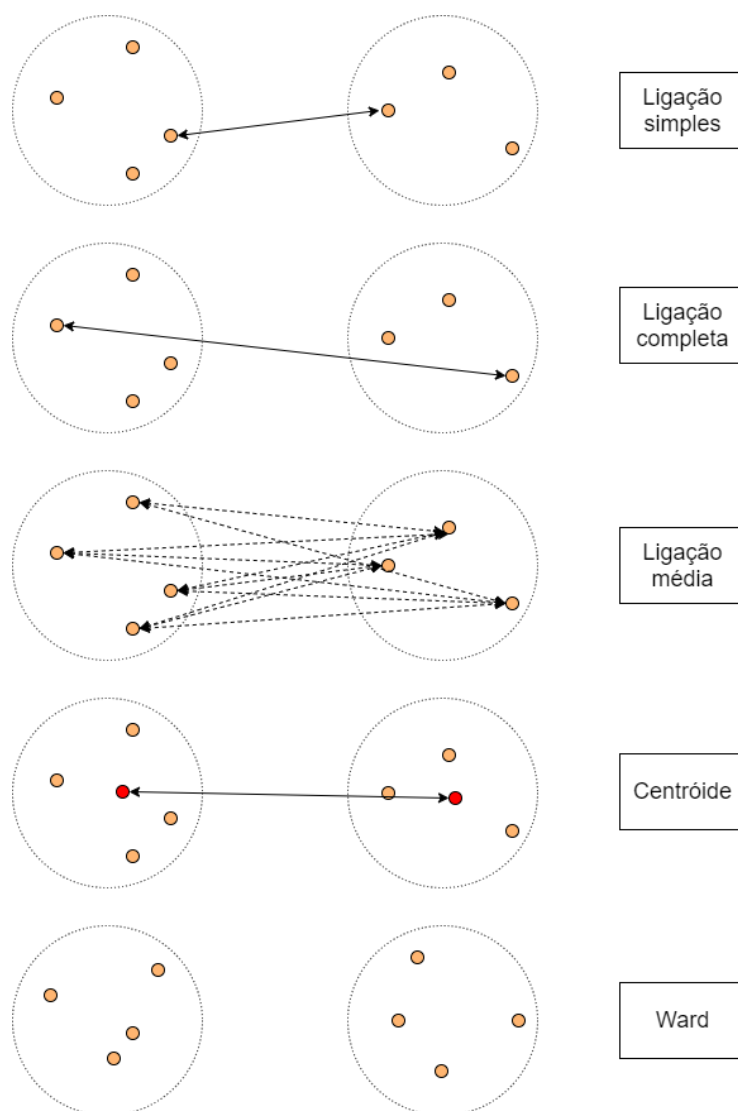


Figura 1 - Diagrama dos métodos de agrupamento hierárquicos

Segundo Jain (1999), os métodos de ligação simples e ligação completa são os mais frequentemente utilizados. Esses dois métodos diferem na maneira de caracterizar a similaridade entre pares de grupos. No primeiro, a distância entre dois grupos é a distância mínima entre todos os pares de elementos pertencentes a estes grupos (vizinho mais próximo). Já no método da ligação completa, é a máxima distância entre todos os pares de elementos que pertencem a estes grupos (vizinho mais distante). A Figura 1 ilustra essas diferenças na definição da similaridade.

Os métodos hierárquicos são assim chamados devido ao fato de que quando atribuído um elemento a um grupo, ele não poderá mais ser removido deste, mesmo que constatado nas iterações seguintes a sua semelhança com outro grupo, o que é uma desvantagem do método. Outra desvantagem é a necessidade de construção e armazenamento da matriz de similaridade, o que pode ser um fator limitante para a

utilização do método em aplicações com computadores comerciais, sem a necessidade recursos computacionais de alta capacidade. Devido a estas desvantagens, em geral, os métodos hierárquicos não são recomendados para problemas envolvendo grandes massas de dados.

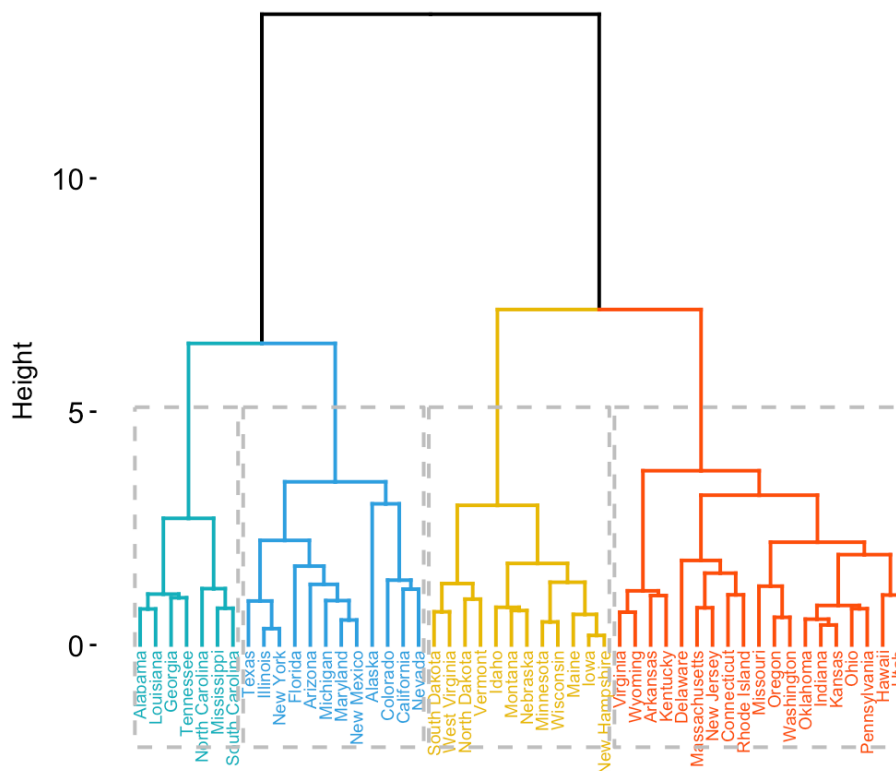


Figura 2 - Dendrograma dos estados americanos utilizando ligação simples. Fonte:

<https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/>

Os resultados desta classe de métodos são representados graficamente por um gráfico chamado dendrograma, onde observar-se as etapas de construção dos agrupamentos e a hierarquia produzida entre os grupos, bem como visualizar o valor da distância que fez com que esse agrupamento ocorresse. Na Figura 2 - **Dendrograma dos estados americanos utilizando ligação simples.** Fonte: <https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/> é mostrado um dendrograma dos estados americanos que representa as etapas de agrupamento utilizando ligação simples (vizinho mais próximo).

2.1.4 Métodos Particionais

Os métodos particionais são também conhecidos como métodos não hierárquicos, porque em qualquer etapa do processo de agrupamento, pode-se

realocar um elemento de um grupo para outro, caso este possua maior proximidade com o grupo vizinho. Os grupos são formados de forma para que exista a máxima semelhança entre elementos do mesmo grupo e não haja outros elementos significativamente semelhantes atribuídos a outros grupos. Tecnicamente falando isso se traduz em coesão interna e isolamentos dos grupos, uma vantagem em relação aos métodos hierárquicos.

Diferentemente dos métodos hierárquicos, os particionais não necessitam de determinação prévia e armazenamento da matriz de similaridade, que tem dimensões proporcionais ao número de elementos. Por isso, esses métodos são indicados quando é necessário reduzir a cardinalidade de um espaço amostral com muitos elementos. Em contrapartida, os métodos não-hierárquicos precisam da definição do número de grupos a priori, ou seja, não é capaz de determinar o número ótimo de agrupamentos para um determinado conjunto de dados. Esse problema, no entanto, pode ser resolvido utilizando as chamadas técnicas de validação de agrupamentos, que avaliam a qualidade dos conglomerados formados segundo um determinado índice ou fator. (KRANEN, 2004).

O caminho mais natural para encontrar a melhor partição é checar todas as partições possíveis dos conjuntos de dados. No entanto, o número de combinações possíveis é muito grande, da ordem de K^{n-1} , onde K é o número de grupos e n o número de elementos que se deseja agrupar (BUSSAB, 1990). Muitos estudos vêm sendo feitos, desde as primeiras publicações sobre métodos de clusterização, a fim de desenvolver procedimentos heurísticos que analisam algumas partições especiais e encontrar a melhor ou uma quase-ótima, visto que analisar todas as partições possíveis é inviável computacionalmente para a maioria das aplicações em casos reais.

Um desses procedimentos heurísticos desenvolvidos é o K-means. Este método é muito popular e vastamente aplicado em trabalhos com grandes conjuntos de dados. Proposto há 65 anos por Steinhaus (1956), ele ainda é um dos algoritmos mais utilizados para clusterização devido à sua simplicidade e facilidade de implementação, além de apresentar em muitos das aplicações encontradas na literatura alta eficiência e sucesso empírico. O algoritmo possui muitas variações e por isso pode ser usado com dados diversos (JAIN, 2009).

K-means

Entre os algoritmos de agrupamento de partição, o K-means é destacadamente o mais importante, originalmente proposto por MacQueen (1967). O algoritmo K-means é amplamente difundido, utilizado nas mais diversas aplicações. A função ser minimizada é definida em função da distância dos elementos em relação aos centros dos agrupamentos. O elemento representativo de um cluster é o seu centróide que, no K-means, é o valor médio para os atributos considerados, relativos a todos os elementos do grupo.

A partir de uma estimativa inicial das coordenadas dos centros dos agrupamentos (centróides), o algoritmo calcula a distância entre todos os pontos do conjunto e estes centróides. A seguir, o algoritmo aloca cada elemento do conjunto em um grupo com a menor distância ao centróide correspondente. A nova estimativa das coordenadas dos centróides é calculada pela média aritmética das coordenadas dos pontos associados a cada grupo e o processo se repete iterativamente. A cada iteração, os elementos são agrupados em função do centróide mais próximo, realocando os objetos entre os grupos, até que os centróides não sejam mais deslocados, ou seja, até que haja uma convergência da sequência de centróides.

De forma resumida, o algoritmo, em suas diversas formas, possui a seguinte estrutura de passos, segundo Xavier(2012):

- I. Seleção de k pontos para sementes iniciais dos centróides dos k grupos.
- II. Cada observação é associada a um cluster, para o qual a dissimilaridade entre o objeto e o centro do cluster é menor que as demais.
- III. Os centros dos clusters são recalculados, redefinindo cada um em função da média de todas as observações atribuídas ao grupo.
- IV. Retorna ao passo 2 até que os centros dos clusters se estabilizem.

A Figura 3 abaixo ilustra o processo.

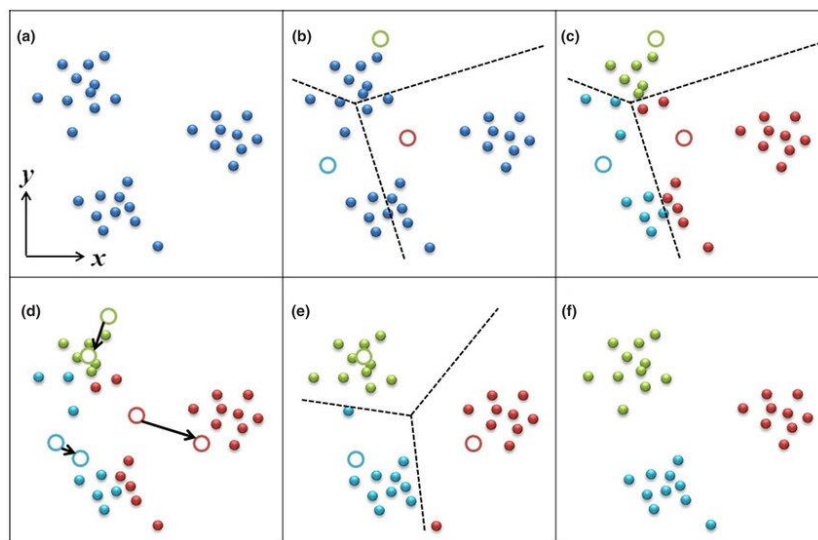


Figura 3 - Estágios do algoritmo K-means. Fonte: Chen & Lai (2018)

O algoritmo é normalmente bastante sensível ao particionamento inicial realizado, em virtude da escolha das coordenadas dos K centróides dos grupos ser feita inicialmente de forma aleatória, sendo a convergência garantida apenas para um ótimo local. Existem várias técnicas para a seleção destas sementes iniciais, sendo a mais simples a amostragem aleatória do conjunto de dados para compor os centroides iniciais.

Esse algoritmo é muito sensível a ruídos, uma vez que o método considera a soma dos desvios ao quadrado em relação aos centros de gravidade. Além disto, é adequado apenas para descobrir grupos com formas esféricas, convexas ou de tamanhos similares.

K-medoids

A diferença básica em relação ao K-means está na utilização de uma das observações do conjunto original como centroide (aqui chamado *medoide*) ao invés da tradicional escolha do centro de gravidade do grupo e trata-se, portanto, de uma abordagem mais realista para o K-means. Análogamente ao K-means, nos algoritmos K-medoids os grupos são definidos como subconjuntos de pontos que estão mais próximos dos seus medoides. O medoide pode ser definido como o elemento do grupo, cuja soma das dissimilaridades a todos os elementos do mesmo grupo seja mínima, o que é equivalente à média dessa soma ser mínima.

Assim como o K-means, esse método é adequado na hipótese de os grupos serem esféricos, ocupando cada medoide uma observação mais central do grupo.

Entretanto, esse método é menos sensível a ruídos e outliers de que o K-means, pois não avalia os desvios das observações aos centróides ao quadrado, como é feito no K-means. Além disso, possui a característica de ser capaz de lidar com qualquer tipo de atributo, constantes ou categóricos.

O algoritmo *Partitioning Around Medoids (PAM)* é um algoritmo clássico da família dos métodos K-medoids, proposto por Kaufman e Rousseeuw (1989). Em linhas gerais, esta heurística efetua a construção dos agrupamentos com a aplicação de dois procedimentos determinísticos, *Build* e *Swap*. O primeiro procedimento é responsável pela construção de uma solução viável, e o segundo procedimento tem caráter similar a uma busca local que atua sobre a solução produzida pelo procedimento Build. Uma breve descrição de seus passos são apresentados, adaptado de Xavier(2012):

- I. Seleciona-se aleatoriamente k das n observações do conjunto de dados $X = \{x_1, x_2, \dots, x_n\}$ para compor o conjunto de medoides $M = \{x_{m1}, x_{m2}, \dots, x_{mk}\}$;
- II. Associa-se as $(n - k)$ observações restantes do conjunto de dados ao grupo G_i ($i = 1, \dots, k$) cujo medoide esteja mais próximo usando qualquer distância métrica válida;
- III. Em cada grupo, a cada elemento não-medoide: calcula-se o custo total da configuração do grupo como se esta observação fosse o medoide
- IV. Seleciona-se a configuração com o menor custo a fim de minimizar a função objetivo definida pela média (por grupos) das distâncias de todos os elementos aos seus respectivos medoides:

$$f = \sum_{i=1}^k \sum_{\forall x_j \in G_i} \frac{d_{x_{mi}x_j}}{|G_i|} \quad (1)$$

- V. Repete-se os passos 2-5 até que não haja mudança nos medoide.

Em relação ao tempo de processamento, o K-Medoid é menos eficiente do que o K-means, pois o cálculo do medoide é mais custoso computacionalmente do que o cálculo do centro de gravidade, gerando um maior tempo de execução do cálculo.

2.2

Problema da Cobertura de Conjuntos

Uma área de pesquisa não intuitivamente relacionada ao estudo de seleção de cenários é a Programação Inteira (PI), que consiste em um problema de programação de otimização matemática em que tanto a função-objetivo quanto as restrições são lineares, porém uma ou mais variáveis de decisão podem apenas assumir valores inteiros [WOLSEY, 1998]. Programação Inteira também é frequentemente utilizada para resolver problemas de otimização combinatória, que é o ramo de Ciência da Computação que estuda problemas de otimização em conjuntos. O maior obstáculo para estes tipos de problema é o fato deles pertencerem a uma classe de problemas chamados de “não polinomiais” (NP) [GAREY & JOHNSON, 1979].

O Problema de Cobertura de Conjuntos (PCC), também conhecido pelo nome em inglês Set Covering Problem (SCP), é um dos 21 problemas NP-completos de Karp (1972), de complexidade NP-difícil (GAREY & JOHNSON, 1979) que já se tornou um problema de otimização combinatória popular. Ele pode ser resolvido através de métodos de programação inteira 0-1 [VIANNA, 2004], é vastamente difundido na literatura e possui diversas aplicações reais, incluindo a programação de operações, manufatura, problemas de localização de instalações, dentre outros.

O objetivo por trás do problema consiste, dado um universo U de elementos, em selecionar o menor número possível de conjuntos de elementos que, quando unidos, contenham todos os elementos existentes no universo U . Ou seja, todos os elementos contidos no universo devem ser cobertos pelos conjuntos selecionados. O modelo deseja, ainda, alcançar este objetivo minimizando os custos dos conjuntos.

A definição matemática do problema e algumas heurísticas para solução do modelo são apresentadas a seguir.

2.2.1

Definição matemática

O Problema de Cobertura de Conjuntos (PCC) é um problema de programação inteira 0-1, e pode ser definido matematicamente como:

$$\text{Minimizar } z = \sum_{j=1}^n c_j x_j \quad (2)$$

Sujeito a

$$\sum_{j=1}^n a_{ij} x_j \geq 1, \quad i = 1, \dots, m \quad (3)$$

$$x_j \in \{0,1\}, \quad j = 1, \dots, n \quad (4)$$

onde $x_j = 1$ se a coluna j está na solução e $x_j = 0$, em caso contrário, e c_j são os custos das colunas. Se $c_j = 1$, o problema é chamado de não ponderado. A função objetivo (2) minimiza o custo da solução. A inequação (3) é a restrição de cobertura completa. Ela obriga que todos os elementos do universo sejam cobertos e se encerra quando cada linha da matriz é coberta por ao menos uma coluna. Então, se a restrição (2) for satisfeita e a função objetivo for minimizada, a solução cobrirá todos os elementos ao menor custo possível (solução ótima). Se a restrição (2) for relaxada, a função objetivo retornará uma solução vazia, afinal ela possui o menor custo possível (zero). A (4) é a restrição de integralidade. Se a restrição (3) for substituída por:

$$\sum_{j=1}^n a_{ij} x_j = 1, \quad i = 1, \dots, m \quad (5)$$

o problema de programação inteira 0-1 resultante passa a ser definido como Problema de Partição de Conjuntos (PPC).

O PCC pode ser formulado numa notação matricial como abaixo, onde $A = (a_{ij})$ é uma matriz $m \times n$ 0-1, e um vetor coluna de m 1's, e c um vetor linha $1 \times n$ de pesos inteiros positivos.

$$\begin{aligned} \text{(PCC)} \quad & \text{Minimizar} \quad cx \\ & \text{Sujeito a} \quad Ax \geq e \\ & \quad \quad \quad x \in \{0,1\}^n \end{aligned}$$

Dessa forma, o PCC é o problema equivalente a cobertura das m -linhas da matriz $A = (a_{ij})$, por um subconjunto de suas n -colunas com custo mínimo, onde $c_j > 0$ é o custo associado a coluna j .

2.2.2

Métodos Heurísticos para solução do PCC

Os algoritmos heurísticos foram desenvolvidos para encontrar soluções problemas de alta complexidade suficientemente próximas da otimalidade consumindo um tempo computacional razoável, sem, no entanto, conseguir garantir que a solução encontrada é a solução ótima, nem mesmo o quão próxima ela está da solução ótima (VIANNA, 2004).

Existe na literatura inúmeros métodos heurísticos para encontrar soluções viáveis para o PCC. Os métodos gulosos são os métodos mais utilizados de construção de soluções viáveis primais próximas a solução ótima. O Método Guloso é uma forma de se resolver problemas através de uma sequência de passos, com um conjunto de opções a serem escolhidas a cada passo. Um Algoritmo Guloso sempre faz a escolha que parece ser melhor no momento, sem nunca reconsiderar esta decisão. Esta escolha é feita com base nas informações locais, ou seja, o algoritmo faz uma escolha ótima para as condições locais, e conta com a sorte para que a escolha feita leve a uma solução ótima para a situação global, o que não é garantido pelo método (SILVEIRA, 1999).

São descritas abaixo as ideias originais dos algoritmos gulosos que procuram determinar uma solução primal viável para o PPC:

Chvátal – a heurística apresentada por Chvátal (1979) baseia-se no fato de que as chances de uma determinada coluna j^* estar na solução aumentam quando a razão entre o custo da referida coluna j^* pela cardinalidade do conjunto M_{j^*} é menor em relação a outras colunas $j \in N$. Assim, para uma dada iteração t , será incluída no conjunto solução S a coluna com menor valor para a razão custo por cardinalidade.

Uma das deficiências do algoritmo heurístico guloso de Chvátal, assim como de outras heurísticas, é a não garantia de uma cobertura primal não redundante. Isto significa que se pode chegar a soluções que não são mínimas e deve-se ao final retirar as colunas redundantes.

Balas e Ho – apresentada por Balas e Ho (1980), esta heurística melhora o resultado da heurística de Chvátal ao utilizar alternadamente cinco funções de avaliação $f(c_j, k_j)$ listadas a seguir, ao invés de uma única:

$$c_j \quad (6)$$

$$\frac{c_j}{k_j} \quad (7)$$

$$\frac{c_j}{\log_2 k_j} \quad (8)$$

$$\frac{c_j}{k_j \log_2 k_j} \quad (9)$$

$$\frac{c_j}{k_j \ln k_j} \quad (10)$$

Nos casos (8) e (9) $\log_2 k_j$ deve ser substituído por 1 quando $k_j = 1$ e no caso (10), $\ln k_j$ deve ser substituído por 1 quando $k_j = 1$ ou 2.

A função (6) inclui na cobertura S^* a cada iteração, a coluna de menor custo. A função (7) minimiza o custo unitário de cobertura de uma linha não coberta (o que seria a Heurística Gulosa de Chvátal). As funções (8), (9) e (10) selecionam a mesma coluna que (7) sempre que $c_j = 1$ ($j = 1, \dots, n$). Caso contrário, (8) assinala menos peso ao número k_j de linhas cobertas, enquanto (10) atribui mais peso e (9) mais ainda em relação ao custo c , quando comparadas com (7). Os testes de realizados mostraram que nenhuma das funções é significativamente melhor que outra. A heurística inclui ainda um passo adicional ao final do algoritmo, para remover as colunas redundantes da cobertura.

SCHEURI – esta heurística apresentada em Vasko e Wilson (1984), inclui na lista das 5 funções de Balas e Ho mais duas outras funções de avaliação heurísticas, que valorizam ainda mais o número de linhas cobertas:

$$\frac{c_j}{k_j^2} \quad (11)$$

$$\frac{c_j^{1/2}}{k_j^2} \quad (12)$$

Ela acrescenta um passo adicional no algoritmo para realizar uma busca 1-opt em uma vizinhança da solução S^* substituindo uma coluna da solução por uma fora da solução, além da remoção das colunas redundantes. Se uma solução viável com custo menor for encontrada, a permutação é concretizada sobre o conjunto solução.

Os autores realizaram posteriormente uma modificação no algoritmo SCHEURI e o chamaram de SCFUNCITO7. O novo algoritmo mudou o ponto de substituição da função de avaliação para que seja determinada aleatoriamente sempre que uma coluna entra no conjunto solução. Sempre que uma coluna for selecionada para entrar na solução, é gerado um valor aleatório entre 1 e 7, correspondente às funções de avaliação que será utilizada até o próximo ponto de troca. Segundo os autores, o algoritmo melhorado foi estritamente melhor em 50% dos problemas testados.

3

Aplicação

Neste capítulo será apresentada a análise preliminar do problema que inspirou a metodologia, o cálculo da demanda estocástica de ferramentas e serviços para construção de poços marítimos de produção de petróleo. A partir dela, o problema será generalizado para que se possa explorar possíveis soluções e definir a estratégia de atuação. A metodologia aplicada é apresentada, juntamente com detalhes técnicos da implementação computacional.

3.1

Análise preliminar do problema

O processo em análise consiste na contratação de ferramentas e serviços para a construção de poços marítimos de produção de petróleo por parte dos gestores de uma empresa da indústria petrolífera. São muitas as variáveis a serem consideradas no cálculo, incluindo as incertezas inerentes ao negócio, que vão desde riscos políticos, de mudança estratégica, até riscos operacionais como imprevistos logísticos ou condições climáticas.

Por causa destas incertezas, o cálculo da demanda precisa do fator estocástico para considerar possíveis alterações no cronograma determinístico de operações.

3.1.1

Planejamento de operações para construção de poços marítimos de produção de petróleo

A processo de exploração de petróleo no oceano é complexo e envolve diversos riscos ao longo da sua realização. Recentemente estão sendo descobertas reservas petrolíferas em condições cada vez mais desafiadoras, poços cada vez mais distantes da costa, em águas ultra profundas e reservatórios em camadas de rochas abaixo da camada pré-sal, que se tornam obstáculos a serem vencidos pelos

exploradores. Estes desafios se refletem no avanço tecnológico que o setor apresentou nas últimas décadas, com equipamentos cada vez mais modernos, personalizados e robustos, e soluções logísticas cada vez mais heterogêneas.

A etapa de construção de poços offshore para exploração de petróleo e gás não seria diferente. Os diferentes poços, com suas diferentes finalidades, possuem características de pressão, temperatura, formação rochosa, volume de reservatório, presença de gases, dentre outras, que podem ser completamente distintas, e por isso uma operação nunca é exatamente como outra do mesmo tipo. Os poços podem demandar equipamentos e ferramentas com especificações muito diferentes, oferecendo um grande desafio para o planejamento das atividades e para a operação da cadeia de suprimentos.

A construção de um poço é realizada por sondas de perfuração, recurso cujo valor de uma diária pode chegar a custar 179,5 mil dólares (custo da última sonda de perfuração contratada pela Petrobras) (Diário do Porto, 2021). Uma sonda com contrato vigente fica disponível para o contratante em tempo integral, e só retorna para o porto em caso de manutenção. Elas são utilizadas nos 3 principais estágios da construção de um poço: a perfuração, o revestimento/cimentação e a completação, que podem acontecer não necessariamente em seguida (pode haver um tempo de espera entre determinadas operações) nem por uma mesma sonda, mas devem acontecer exatamente nesta ordem. Por isto, quando há algum imprevisto em uma das atividades, que gera atraso na finalização, todas as atividades posteriores no cronograma do poço e da sonda (Figura 4) podem ser afetadas, sendo necessário atrasar, adiar ou cancelar tarefas, ou realocar tarefas para outras sondas. De toda forma, mitigar estes problemas pode gerar uma grande perda financeira para a empresa exploradora.

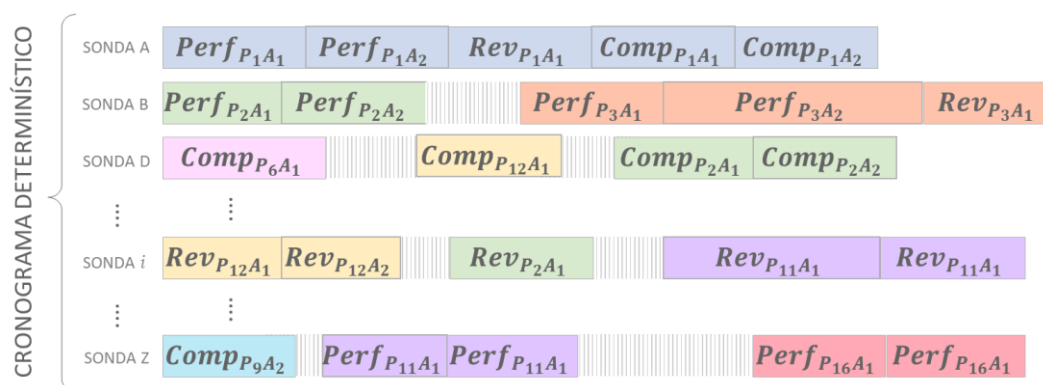


Figura 4 - Esquemático de um cronograma determinístico.

Sejam por fatores controláveis como erros operacionais ou falhas no planejamento e dimensionamento dos materiais e ferramentas, ou por fatores incontroláveis como condições climáticas ou alterações geológicas não mapeadas, as incertezas se mostram inevitáveis, porém podem ser estimadas e incluídas no planejamento, para ter seu impacto financeiro reduzido. Por este motivo, o cronograma determinístico de atividades de todas as sondas é atualizado com frequência, levando em consideração, dentre muitos fatores, o andamento das atividades e as incertezas esperadas para as atividades do futuro próximo, gerando o chamado cronograma operacional.

Para o dimensionamento de ferramentas e serviços, no entanto, o cronograma operacional não possui muita serventia. A fabricação das ferramentas e equipamentos de grande complexidade utilizadas no processo pode chegar a 2 anos, acrescido do período da licitação. Assim sendo, os responsáveis pela contratação destes equipamentos precisam ter uma visão de longo prazo do cronograma e uma previsão das incertezas em um longo horizonte de tempo, de até 5 anos à frente.

Para prever estas incertezas, os dados históricos de cada tipo de operação são analisados e são encontradas as distribuições de probabilidade de falhas no planejamento das operações. A partir delas, é possível estimar a probabilidade de cada tarefa ter sua duração alterada, sua data de início adiantada ou atrasada, ou até mesmo a probabilidade de ser cancelada, mesmo em um horizonte de tempo distante. Com um cronograma estimado de operações, conhecendo as características geológicas esperadas para cada poço, as ferramentas necessárias em cada tipo de operação e a logística de equipamentos, pode-se, então, estimar a demanda de equipamentos no longo prazo para dar início às contratações.

Este é um cálculo complexo, pois envolve muitas regras de negócio e diretrizes técnicas que precisam ser atendidas, e possui um grande volume de dados processados, fazendo necessário o uso de ferramentas computacionais. As etapas e a descrição destas regras são detalhadas no tópico a seguir.

3.1.2

Cálculo da demanda por ferramentas e serviços para construção de poços

A empresa petrolífera analisada utiliza uma ferramenta computacional para realizar o cálculo da demanda por ferramentas e serviços necessários para as operações da construção de poços submarinos. Resumidamente, o cálculo realizado pelo sistema é dividido em duas etapas: a primeira consiste em introduzir as incertezas no cronograma determinístico e, então, aplicar as regras de negócio que definem quais equipamentos serão destinados a cada atividade do cronograma.

A primeira etapa resulta no cronograma estocástico e a forma como ele é gerado é fundamental para este estudo. A geração depende da geração de um cronograma estratégico (determinístico) utilizado como base e da definição das distribuições de probabilidade das incertezas que, por questões internas de confidencialidade e segurança de dados, não são divulgadas nem mesmo aos responsáveis pelas contratações dos equipamentos. São disponibilizados apenas, para cada atividade, em cada poço e sonda, sorteios nas devidas distribuições de probabilidade que indicarão a porcentagem de alteração na duração que a atividade sofrerá, a data de início simulada (que pode ser mantida, atrasada ou adiantada) e se a atividade será mantida no cronograma ou cancelada.

Para que as distribuições de probabilidade sejam corretamente representadas, este sorteio é repetido várias vezes e as taxas aplicadas ao mesmo cronograma estratégico, gerando diferentes cenários do mesmo cronograma. Após a aplicação das incertezas, todos eles passam por um reajuste das atividades, a fim de evitar sobreposição de operações em uma mesma sonda ou a quebra de precedência (ordem em que devem ser realizadas as operações) no poço. O conjunto final de cenários simulados é denominado cronograma estocástico. A Figura 5 ilustra alguns casos de cronogramas estocásticos (1, 2 e 3), para uma única sonda, após a aplicação de incertezas de data de início, duração e permanência das tarefas do cronograma determinístico. Alterações nas atividades podem “puxar”, “empurrar” ou forçar a troca da sonda de atividades seguintes.

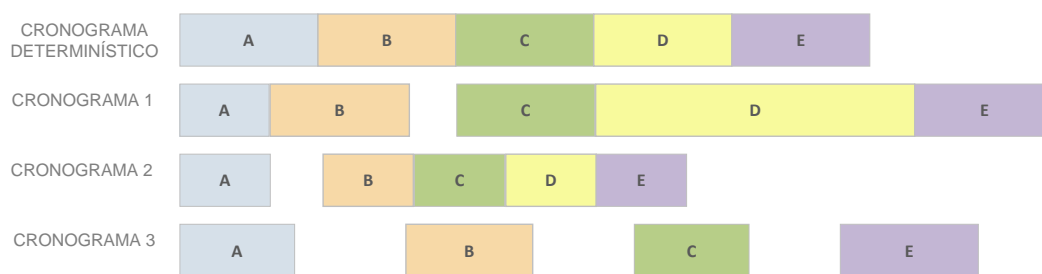


Figura 5 – Esquemático de um cronograma determinístico e cronogramas simulados gerados a partir da aplicação das incertezas nas operações da carreira de uma sonda de perfuração

Apenas esta etapa do cálculo, com o ajuste das regras de sobreposição e precedência, para todas as sondas contratadas, em um horizonte de 5 anos, consome significativo recurso e tempo computacional. E ainda resta a segunda etapa, que se inicia em seguida, e possui diversas regras a serem analisadas.

Cada atividade, de cada sonda, em cada cronograma, possui uma determinada série de ferramentas e serviços necessários para a sua realização. Estas ferramentas devem ter especificações adequadas às características do poço. Por exemplo, se é esperado que em determinada operação uma ferramenta seja submetida a uma pressão de 500Psi, o valor nominal de pressão suportada, garantida pelo fabricante, deve ser igual ou superior a 500Psi. Idealmente, a ferramenta deve ter capacidade próxima a exigida do poço, pois um valor inferior inviabiliza a operação e um valor muito acima do necessário significa desperdício de dinheiro. Analogamente para temperatura, peso e outras características técnicas, os limites das ferramentas nunca devem ser ultrapassados nem serem subutilizados.

Considerando também que ferramentas podem ser reaproveitadas e utilizadas em diversos poços, a logística das ferramentas deve ser levada em conta no dimensionamento. Isso pois, se uma sonda vai operar, em seguida, em dois poços que requerem uma ferramenta de mesmo tipo e finalidade, não é necessário utilizar duas ferramentas e sim uma unidade que atenda a ambos os poços, ou seja, a ferramenta de menor capacidade que seja capaz de atender o poço com características mais restritivas. No entanto, se os poços serão operados pela mesma sonda, mas não em seguida, é preciso avaliar se o tempo entre as operações é inferior ao tempo de desmontagem/desembarque e embarque/montagem da ferramenta. Se for inferior, a ferramenta deve continuar embarcada, pois o tempo em que ficará ociosa ainda será menor que o tempo logístico. Se for maior, valerá

a pena realizar o desembarque da ferramenta para que seja utilizado em outras operações, em outra sonda.

Além disso, é possível definir regras estratégicas do negócio. O próprio gestor define, por exemplo, em que tipos de poços uma ferramenta de determinada especificação poderá ser alocada, ou até mesmo o tempo de espera para reutilização de uma ferramenta e algumas limitações técnicas particulares.

Todas estas regras mencionadas são editáveis no sistema utilizado para o cálculo. Isso porque é frequente a necessidade de ajustes devido a questões contratuais, novas tecnologias contratadas ou simplesmente novas diretrizes definidas pela companhia ou exigidas por órgãos reguladores. Por isso, a cada geração de demanda, mesmo que o cronograma não tenha sido alterado, apenas as regras, ou o contrário, é necessário refazer a segunda etapa do cálculo. Dessa forma, por conta do grande volume de dados processados pelo sistema e a complexidade dos cálculos realizados, fez-se necessário o estudo acerca da quantidade de cronogramas simulados utilizados para a geração do cronograma estocástico, com o objetivo de avaliar uma possível redução desse número. Atualmente, são entregues aos responsáveis pelo dimensionamento de demanda sorteios para 3000 cronogramas simulados e a orientação seguida é de que todos devem ser utilizados, o que gera uma grande ocupação dos servidores e um longo tempo de processamento. A análise de sensibilidade do número de cronogramas e as estatísticas da demanda gerada a partir das diferentes quantidades testadas são apresentadas no tópico a seguir.

3.1.3

Análise da demanda frente à quantidade de cenários

O objetivo deste estudo preliminar é, portanto, avaliar o impacto da redução deste número nos resultados de demanda estocástica diária de cada ferramenta fornecida pelo sistema responsável pelo cálculo. Se possível, identificar a uma quantidade menor de cronogramas que alcance algum ganho computacional no processamento de dados com o mínimo de perda de qualidade na informação gerada. Para tal, foi selecionado um conjunto de ferramentas que abrange os principais tipos de operação, com no mínimo 3 equipamentos para cada. Foram geradas demandas com as diferentes quantidades de cronogramas estocásticos

candidatas selecionadas na ordem em que foram disponibilizadas (50 primeiras, 100 primeiras, assim por diante). Elas variaram de 50 a 1000, além da quantidade de 3000 utilizada atualmente que será a base de comparação dos resultados.

Os indicadores diários considerados, sendo a a quantidade candidata, foram:

- média
- percentis 10, 50 (mediana) e 90
- desvio padrão
- desvio padrão sobre a média

$$\frac{Desvpad(a)}{Média(a)} \quad (13)$$

As estatísticas de comparação com a quantidade atual foram:

- distância absoluta entre os indicadores das quantidades candidatas e os indicadores da quantidade atual (3000 cronogramas)

$$Indicador(a) - Indicador(3000) \quad (14)$$

- distância relativa entre os indicadores das quantidades candidatas e os indicadores da quantidade atual (3000 cronogramas)

$$\frac{Indicador(a) - Indicador(3000)}{Indicador(3000)} \quad (15)$$

Foi considerado um horizonte da simulação de 5 anos, aplicados os riscos de duração e data de início e testadas as quantidades de 50, 100, 200, 500, 1000 e 3000 cronogramas.

Os resultados para a demanda média diária entre os cronogramas encontram-se na Figura 28, agrupados por ferramenta. De forma geral, observa-se que as curvas dos cronogramas de todas as 6 quantidades de cronogramas apresentaram um comportamento muito similar visualmente, o que significa que a demanda média é pouco impactada pela introdução de mais cenários. As demandas com 50, 100 e 200 cronogramas possuem pequenos desvios do conjunto de curvas, mais notáveis nos picos. Já os cronogramas com 500, 1000 e 3000 variam minimamente entre si.

O desvio padrão diário das demandas, por si só, mostra uma variação já esperada devido não apenas a qualidade de cronogramas, mas a quantidade de

elementos da amostra para o cálculo (50 valores por dia, 100 valores por dia, assim por diante). Além disso, possui difícil análise dado o objetivo da pesquisa, visto que as demandas individuais das ferramentas são muito distintas e a comparação entre ferramentas não é possível sem a normalização do dado.

O desvio padrão diário dividido pela média diária (Figura 29), por outro lado, indica a porcentagem da média que os cronogramas variam, e então, com mais clareza, é possível observar o comportamento das demandas com menos cronogramas mais distantes da curva da demanda com 3000 cronogramas.

Até este ponto da pesquisa, as curvas relativas às demandas de 500, 1000 e 3000 cronogramas mal podem ser distinguidas entre si, indicando que os dois primeiros valores são fortes candidatos a uma nova diretriz para alcançar o objetivo proposto.

Para a comparação entre as curvas, foi realizada a análise das distâncias absolutas entre as candidatas (50 a 1000 cronogramas) e a curva realizada com todos os 3000 cronogramas. Este indicador retorna em quantas unidades de ferramenta a demanda diária média da curva candidata difere da curva de 3000, para mais ou para menos. Os resultados são apresentados na Figura 30, onde a curva de referência é a demanda de 3000 cronogramas, é constante e igual a zero.

Os gráficos nos mostram novamente que as curvas de menor número de cronogramas possuem variabilidade maior em relação a curva atual, mas não é possível notar o quão representativa é esta diferença em relação à demanda média. Normalizando, então, os dados, temos a distância relativa entre as curvas candidatas e a atual. Este indicador retorna a porcentagem da demanda calculada com 3000 cronogramas as curvas candidatas diferem. Os resultados são apresentados na Figura 31.

Uma vez normalizado, é possível calcular a média das curvas de distância relativa em todo o horizonte da simulação. Esta aplicação foi feita para os indicadores diários de média, percentil 10, mediana, percentil 90, desvio padrão, desvio padrão sobre a média e variância, e sintetizada na Tabela 1. Foram marcados em vermelho os valores que ultrapassam 10% de diferença da curva de 3000 cronogramas. O resultado confirma, portanto, a hipótese de que as quantidades de 50, 100 e 200 apresentam maior variabilidade em relação a quantidade de 3000 atualmente utilizada, já que este comportamento foi notado em vários indicadores.

As demandas com 500 e 1000 cronogramas, em contrapartida, apresentam demanda média muito similar à demanda completa, com menos de 4% de diferença.

Tabela 1 - Distâncias relativas médias das demandas com diferentes quantidades de cronogramas para todo o horizonte de simulação, por indicador.

Indicador	50	100	200	500	1000
Média	-2,30%	-6,70%	-3,40%	-1,00%	-3,30%
P10	41,60%	24,90%	14,50%	5,00%	3,20%
P50	-3,10%	-3,20%	-2,20%	-0,90%	0,50%
P90	-16,80%	-7,70%	-2,70%	-0,60%	-0,70%
Desvio padrão	-26,10%	-18,70%	-11,40%	-4,50%	-6,50%
Desvio padrão/ Média	-9,40%	-3,90%	-2,10%	-0,70%	-1,70%
Variância	-20,30%	-16,90%	-9,20%	-1,20%	-6,90%

Foram gerados também os indicadores a nível de atividade de cada uma das demandas, por tipo de serviço. Os indicadores são:

- Duração determinística de todas as operações – é esperado que seja a mesma para todas as demandas;
- Média da duração original (apenas das operações que permaneceram no cronograma);
- Média da duração simulada;
- Diferença entre as médias da duração simulada e da original;
- Média da razão entre duração simulada e original de cada atividade;
- Desvio-padrão da razão entre duração simulada e original de cada atividade;
- Média do número total de operações.

Observa-se nos gráficos variações muito pequenas nas médias tanto das durações, quando na média e desvio-padrão da razão entre as durações, entre uma demanda e outra. Nota-se também que a utilização de mais cenários não interferiu o quanto esperado nas durações simuladas. Usando 3000 ou 50 cronogramas, a diferença entre as durações simulada e original médias é alterada, no pior dos casos, em apenas 26 dias.

Devido à escolha dos cronogramas na sequência em que foram disponibilizadas (n primeiros registros) e a falta de informações sobre o método de sorteio das incertezas, não se pode garantir que os valores sorteados são não tendenciosos ou homogeneamente distribuídos. Para refutar qualquer hipótese de

enviesamento dos dados, foram criadas 10 amostras de 500 cronogramas escolhidas aleatoriamente do relatório de Demanda Diária por Cronogramas gerado pelo sistema. Os resultados da demanda média para cada ferramenta são apresentados na Figura 32.

As demandas geradas a partir das 10 amostras foram comparadas entre si através do cálculo da distância relativa, que neste caso consiste simplesmente na diferença ponderada da demanda diária entre cada par, e estão dispostas na Tabela 2. Os resultados confirmam a consistência das análises feitas anteriormente, uma vez definido que uma distância relativa média de 5% seria um valor aceitável para que dois cronogramas sejam considerados próximos e, portanto, semelhantes. A maior distância relativa média encontrada entre as amostras foi de 2,23%, concluindo-se que a utilização de 500 cronogramas na simulação apresenta resultados suficientemente próximos a 3000 cronogramas.

Tabela 2 - Distâncias relativas da média diária entre cada par de demandas geradas a partir das 10 amostras de 500 cronogramas aleatórios

Amostras	1	2	3	4	5	6	7	8	9	10
1	0,00%	-0,11%	-1,28%	-1,01%	-1,55%	-0,56%	-0,48%	-0,19%	-0,38%	0,45%
2	-0,27%	0,00%	-1,45%	-1,13%	-1,80%	-0,62%	-0,53%	-0,24%	-0,35%	0,37%
3	0,99%	0,98%	0,00%	0,18%	-0,36%	0,54%	0,67%	0,92%	0,76%	1,60%
4	0,56%	0,57%	-0,54%	0,00%	-0,71%	0,20%	0,28%	0,53%	0,40%	1,25%
5	1,10%	1,02%	0,02%	0,37%	0,00%	0,64%	0,68%	1,05%	0,82%	1,70%
6	0,22%	0,32%	-0,94%	-0,58%	-1,23%	0,00%	-0,01%	0,16%	0,07%	0,87%
7	0,08%	0,19%	-1,03%	-0,72%	-1,44%	-0,23%	0,00%	0,09%	0,01%	0,79%
8	-0,23%	-0,11%	-1,38%	-1,06%	-1,63%	-0,66%	-0,50%	0,00%	-0,33%	0,47%
9	-0,09%	0,10%	-1,22%	-0,86%	-1,55%	-0,42%	-0,26%	-0,01%	0,00%	0,61%
10	-0,83%	-0,75%	-1,95%	-1,58%	-2,23%	-1,20%	-1,06%	-0,79%	-0,97%	0,00%

Em resumo, os resultados apresentam indícios de que não há diversificação suficiente entre os cronogramas que justifique a utilização dos 3000 cronogramas. Mesmo com uma escolha aleatória de cenários, o número de cronogramas poderia ser reduzido dos atuais 3000 para 500 sem perdas representativas na qualidade dos resultados. Observa-se, dada a simplicidade dos métodos utilizados como critério para seleção dos cronogramas, que algoritmos mais inteligentes de seleção de cenários poderiam apresentar uma quantidade significativamente menor do que os 500 encontrados na análise preliminar, o que contribui fortemente para o objetivo deste estudo e será abordado na metodologia aqui proposta.

3.2

Caracterização genérica do problema

A análise preliminar permitiu compreender melhor o problema estudado. Ela mostra que não apenas é desnecessário utilizar os 3000 cronogramas, como também que os cenários não possuem a variabilidade que parece ser esperada pela empresa ao adotar tal procedimento. De maneira geral, parece haver muitos cronogramas similares, que reduzem o impacto daqueles com alterações, de fato, significativas.

Com base nisso, as principais características observadas neste problema são:

- Base de dados extensa, com muitas operações, tipos de ferramentas, e variáveis técnicas envolvidos na tomada de decisão;
- Complexas e custosas etapas de cálculo;
- Limitações computacionais para o processamento de todo esse volume de dados e das premissas para a correta alocação de ferramentas às operações;
- Cenários pré-determinados, com distribuições de probabilidades desconhecidas;
- Grande quantidade de cenários de incerteza que não necessariamente refletem alterações significativas em relação tanto ao cenário determinístico quanto aos outros cenários estocásticos.

Os quatro primeiros pontos envolvem questões técnicas de difícil solução. Não seria correto apenas simplificar o cálculo excluindo variáveis da análise, também seria complexo e demorado redefinir diretrizes da companhia sobre a divulgação das distribuições de probabilidades, que podem trazer informações sensíveis e confidenciais sobre as estratégias adotadas pela empresa em seu

planejamento. As limitações computacionais não serão resolvidas definitivamente apenas com a extensão da capacidade de armazenamento e processamento, ela poderia amenizar o problema até que se chegue novamente ao estado crítico, tempos depois.

No entanto, estes quatro pontos poderiam ser significativamente melhorados pela a solução do quinto ponto: fazer uma seleção inteligente que reduza drasticamente a quantidade de cenários, mas que ainda assim, todos os cenários de incerteza do conjunto universo estejam representados. Esta seleção deve atender aos requisitos complementares das características do problema citadas acima:

- Redução do volume de dados;
- Mínima perda de confiabilidade dos resultados;
- Redução do tempo e custo do processamento;
- Representação correta das incertezas;
- Cenários com alta variabilidade e representativos do conjunto universo.

Uma solução rápida e clássica para o problema seria a aplicação do método de clusterização K-medoids. Seriam selecionados cenários reais, cada um representando um agrupamento de cenários semelhantes de alguma maneira. Porém, durante a pesquisa sobre o tema (apresentada no capítulo 2) e estudo do exemplo trazido no presente trabalho, algumas questões foram levantadas sobre a viabilidade do uso do método:

- I. Um cenário, criado a partir de diferentes distribuições de probabilidade para os seus diferentes atributos, poderia ser fielmente representado por um único cluster?
- II. Um cenário que representa um objeto estático, que não varia com o tempo, com inúmeros atributos, e que não tem influência sobre outros cenários no tempo (são totalmente independentes), poderia ser tratado assim como um cenário gerado a partir de série temporal?
- III. Como fazer uma análise de sensibilidade para determinar o K, que vai mudar a cada atualização das incertezas, o que, por sua vez, implica na alteração dos dados, sendo as etapas de cálculo com este volume de dados tão custosas computacionalmente?

Pensou-se, com base nisso, em uma solução alternativa aos tradicionais métodos de clusterização particionais, que utilizasse uma visão não de árvore de cenários, onde há uma “hierarquia” e decisões influenciam decisões afrente, mas de grafo. Para tal, a busca na literatura descrita a seguir apresenta algumas possibilidades.

3.3 Trabalhos Relacionados

Uma busca foi realizada na base científicas Scopus incluindo algumas palavras-chaves relacionadas ao trabalho. As palavras-chave foram agrupadas por tema, como apresentado da Tabela 3, tendo no mesmo grupo sinônimos e variações dos nomes dos métodos, tanto em português como em inglês. No grupo A estão listadas as palavras chaves relacionadas ao objetivo técnico básico do trabalho, que é a seleção de cenários. No grupo B o contexto de incertezas em que está inserido, já que o termo seleção de cenários pode ser encontrada em estudos nada relacionados com o tema abordado aqui, como cenários econômicos, cenários otimistas/pessimistas de mercado ou de produção de bens, dentre outros. No grupo C, por fim, encontram-se as palavras chaves relacionadas do método da Cobertura de Conjuntos.

Tabela 3 - Palavras-chaves da busca por trabalhos correlatos

Grupo	Palavras-chave
A	"scenario reduction" OR "scenario selection" OR "scenario construction and reduction" OR “scenario generation and reduction”
B	"stochastic scenarios" OR "stochastic" OR "stochasticity" OR "under uncertainty scenarios"
C	"set-covering" OR "set covering"

A busca foi feita de forma incremental, adicionando os grupos seguintes, e os resultados obtidos foram contabilizados e se encontram na Tabela 4.

Tabela 4 - Resultados da busca por trabalhos correlatos no repositório de trabalhos científicos Scopus (realizada em 04/08/2021)

Grupo	Total de publicações
A	870
A+B	407
A+B+C	1

A única publicação encontrada na busca pelas palavras-chave dos três grupos em conjunto é a Wang et al. (2015), que utiliza um algoritmo de geração de cenários para a produção de energia solar e eólica baseado na distância de Wasserstein e aplica um algoritmo de K-Medoids para a redução de cenários. Ele não utiliza, de fato, nenhuma técnica relacionada ao Problema de Conjuntos (set covering), tendo sido encontrado na busca por coincidência na ordem de palavras. Avaliou-se, então, as 407 publicações encontradas na busca pelas palavras-chave dos grupos A e B, utilizando as ferramentas de filtro do site, e destacou-se a grande quantidade de publicações aplicando métricas probabilísticas para a mesma finalidade almejada neste estudo.

Diferentes estratégias de redução de cenários têm sido desenvolvidas com base em métricas probabilísticas. Um bom exemplo encontrado na busca é o estudo do brasileiro De Oliveira (2007) utilizou o algoritmo denominado Redução Ótima de Cenários (ROC), introduzido em Dupačová et al. (2003). Em linhas gerais, a redução de cenários proposta parte de uma distribuição de probabilidade discreta P , que pode ser representada por uma árvore de cenários, e possui o objetivo de encontrar uma outra distribuição de probabilidade discreta Q correspondente a uma árvore de cenários menor, que seja a mais próxima da original P . Em Dupačová et al. (2003) a métrica utilizada para a proximidade é do tipo Fortet-Mourier.

A ideia em Heitsch & Romisch (2003) deriva de resultados de estabilidade quantitativa em termos de uma dada probabilidade, resultando em duas abordagens para a decidir os cenários a serem removidos de P para que se obtenha Q (redução do número de cenários): a *forward selection* (seleção para frente) e a *backward*

reduction (redução para trás). Heitsch & Romisch (2003) estende a ideia original abordada em Dupačová et al. (2003), e melhora o procedimento iterativo, alcançando uma nova versão de seleção direta, a denominada *fast forward selection*, e simultânea redução para trás, denominada *simultaneous backward reduction*, com considerável melhoria de desempenho na solução do problema clássico de set-covering.

Estas técnicas de redução de cenário há muito vêm sendo usadas em aplicações de programação estocástica, como para gerenciamento de energia em Growe-Kuska et al., (2003). No entanto, elas pressupõem cenários que podem ser representados por uma distribuição de probabilidade P conhecida, uma árvore de cenários ou, minimamente, um *fan* (ventilador) que pode ser convertida em árvore, como ilustrado na Figura 6. Ou seja, é conhecida a distribuição de probabilidade P do objeto final de estudo, como a demanda final de energia ou, no caso em que este trabalho se aplica, a demanda por ferramentas e serviços para construção de poços de petróleo. Nisso pressupõe-se a principal diferença entre as abordagens encontradas na literatura e a aplicada no presente trabalho.

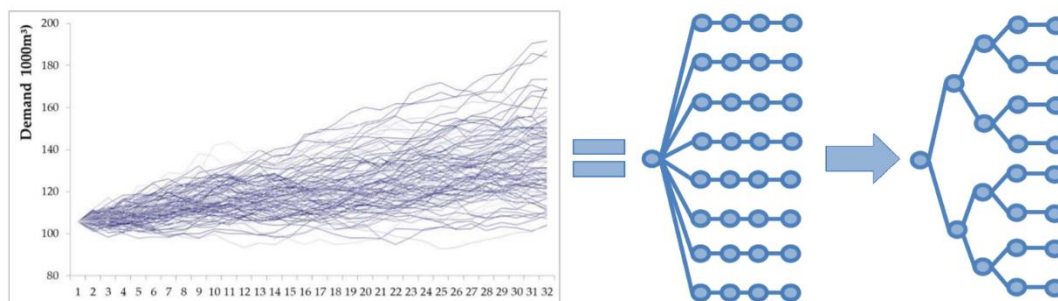


Figura 6 - Estrutura de *fan* convertida em árvore de cenários. Fonte: Hamacher & Oliveira, 2013.

O cenário neste trabalho consiste em cada cronograma simulado gerado a partir do sorteio. O cronograma é um objeto complexo com diversos atributos, gerado a partir de um processo único de amostragem, uma única perturbação do cronograma determinístico para cada atributo. Os cronogramas não são frutos de aplicações de incertezas sobre outros cenários estocásticos (decisões), como programação multi-estágios. Eles são independentes, estáticos e atemporais, e por isso poderiam ser considerados pontos espalhados no espaço n -dimensional e,

assim, o conjunto de cronogramas seria melhor representado por um grafo, não uma árvore de cenários. A demanda por ferramentas é o objeto final de estudo, porém, no ponto em que a metodologia é aplicada, a demanda ainda é desconhecida e o objetivo é reduzir o conjunto original para, justamente, não precisar calcular a demanda de todos os cronogramas simulados.

Com base nisso, a metodologia proposta, apresentada no próximo tópico, conta com uma forma mais genérica de solução do problema de set-covering introduzido em Heitsch & Romisch (2003), sem pressupor nada sobre o cenário. Ele é tratado como um objeto contendo diferentes atributos, e a proximidade entre os cenários (cronogramas) é medida segundo a semelhança entre cada atributo.

3.4

A metodologia aplicada

A seleção de cenários proposta neste trabalho não consiste apenas em um método único, mas sim em uma metodologia, composta pelas etapas da Figura 7 e descritas na sequência.

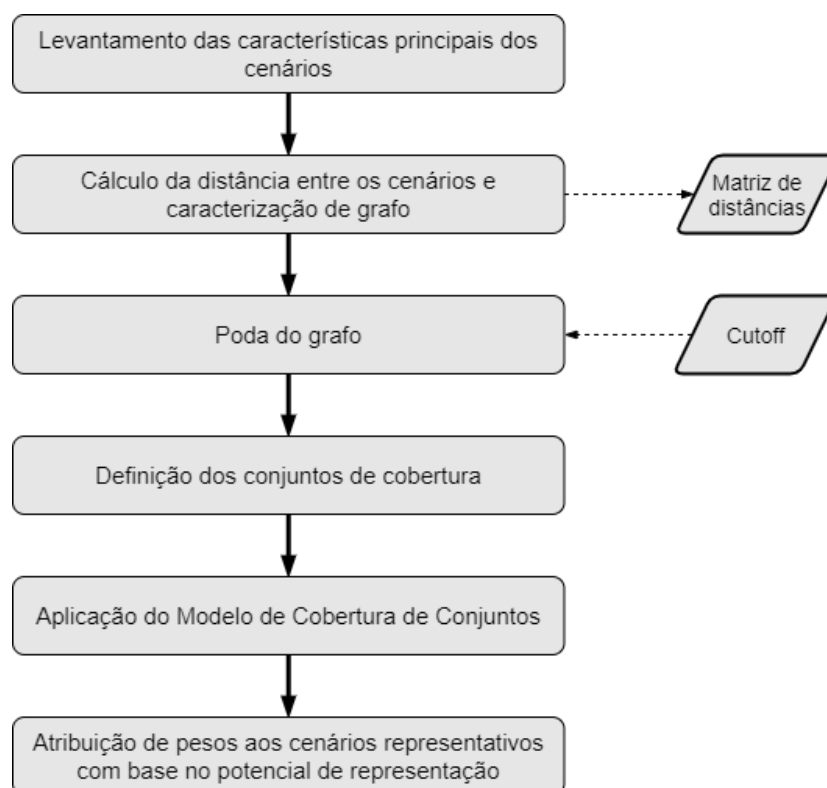


Figura 7 - Etapas da metodologia utilizada no estudo

Levantamento das características principais dos cenários – São selecionadas as características principais de cada cenário, a depender do contexto em que estão inseridos e das informações relevantes ao estudo em questão. No exemplo do cálculo da demanda estocástica para ferramentas e serviços de construção de poços marítimos de produção de petróleo, como introduzido no tópico anterior, as características definidas como mais importantes foram a duração das tarefas do cronograma, bem como a data de início e a probabilidade de cancelamento de cada tarefa. Estas características foram metrificadas da seguinte maneira:

- Quantidade de tarefas do cronograma estocástico que mantiveram a mesma data de início da tarefa no cronograma determinístico;
- Quantidade de tarefas do cronograma estocástico que tiveram a data de início da tarefa alterada em relação ao cronograma determinístico;
- Quantidade de tarefas do cronograma estocástico que tiveram a duração da tarefa mantida no cronograma determinístico;
- Quantidade de tarefas do cronograma estocástico que tiveram a duração da tarefa alterada (para mais ou para menos) em relação ao cronograma determinístico;
- Quantidade de tarefas que foram mantidas do cronograma estocástico (independente de alterações de data ou duração);
- Quantidade de tarefas que foram canceladas no cronograma estocástico;
- Média de variação (delta) da data de início das tarefas que tiveram data de início alterada no cronograma estocástico;
- Média de variação (delta) da duração das tarefas que tiveram a duração alterada no cronograma estocástico;
- Variância das datas de início das tarefas que tiveram data de início alterada no cronograma estocástico;
- Variância das durações das tarefas que tiveram duração alterada no cronograma estocástico.

No exemplo em questão, todas as características foram metrificadas e se tornaram variáveis numéricas, embora este não seja um requisito. A metodologia independe do formato do *dataset*, pois o *input* da seleção será a matriz de distâncias.

Cálculo da distância entre os cenários e caracterização de grafo – São calculadas as distâncias entre cada par de cenários. O método de cálculo das distâncias deve ser escolhido com base nos tipos de variáveis do *dataset*: se só possui variáveis numéricas, categóricas, ou se é híbrido. Os inúmeros métodos encontrados na literatura se adequam diferentemente para cada tipo de variável e a escolha errada pode imprimir uma falsa similaridade ou dissimilaridade entre os cenários. No exemplo apresentado, foi utilizada a distância de Gower, proposta em Gower (1971). O algoritmo permite a análise combinada de variáveis qualitativas e quantitativas, de forma que os valores da matriz de distância gerada fiquem entre 0 e 1. Segundo análise do autor, a técnica apresenta melhor uso de dados e traz maior robustez à análise de dissimilaridade. É possível definir em seus parâmetros o tipo de cada variável (coluna) do *dataset* e seus pesos. No exemplo utilizado neste trabalho, todas foram definidas como numéricas e receberam o mesmo peso.

Uma vez aplicado o método de Gower, obtém-se a matriz de distâncias $n \times n$, onde n é o tamanho do conjunto universo de cenários, U . Assim, obtém-se uma visão em formato de grafo não orientado completo de todos os cenários, onde as distâncias se tornam os pesos das arestas e pode-se, virtualmente, visualizar e determinar quais são mais parecidos e quais são mais diferentes, diante das características definidas como importantes.

Poda do grafo – Criado o grafo, são cortadas as arestas maiores, ou seja, cortadas as ligações entre cenários mais diferentes entre si. É definida uma distância máxima permitida, o *cutoff*, e todas os valores maiores que estes são apagados do grafo. Por isso, o *cutoff* é o ponto de maior sensibilidade em toda a metodologia. Ele será o responsável por limitar o espaço de soluções: se muito próximo de 0, restarão poucas arestas e muitos cenários serão definidos como únicos e sem cenários sequer semelhantes. Se muito próximo da maior distância entre dois vértices, haverá poucos cortes no grafo e poucos cenários serão definidos como distintos, em geral todos serão parecidos. A Figura 8 ilustra estes dois casos extremos e um intermediário, onde as arestas da diagonal do quadrado de lado 1 (que medem $\sqrt{2} \cong 1,41$) são podados pelo *cutoff*.

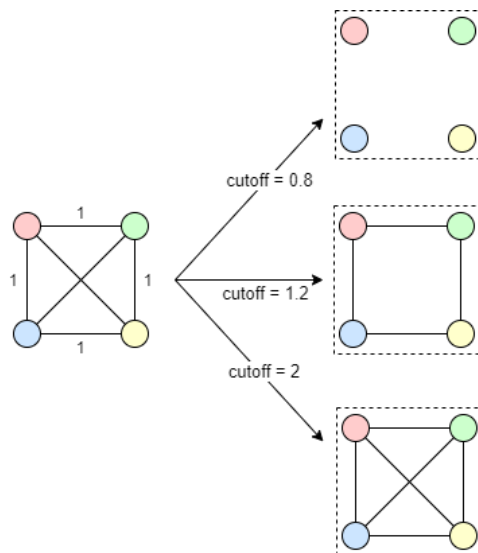


Figura 8 - Diagrama da intensidade da poda frente ao valor do cutoff

Definição dos conjuntos de cobertura – Após o corte no grafo, cada vértice será quantificado em relação a quantidade de arestas remanescentes congruentes a ele. Isto é, cada cenário terá uma lista de outros cenários que se encontram a uma distância menor que o *cutoff* e por isto ainda estão ligados por uma aresta. As n listas, chamadas de *subsets*, terão diferentes tamanhos e seus elementos compõem o espaço de cobertura de cada vértice: por exemplo, se o *subset* do elemento A contiver os elementos $\{A, B, C\}$, significa que os elementos B e C estão a uma distância suficientemente pequena (menor que o *cutoff*) do elemento A para que se possa considerar que A representa B e C . Diz-se, portanto, que A , além de cobrir a si mesmo, cobre também B e C .

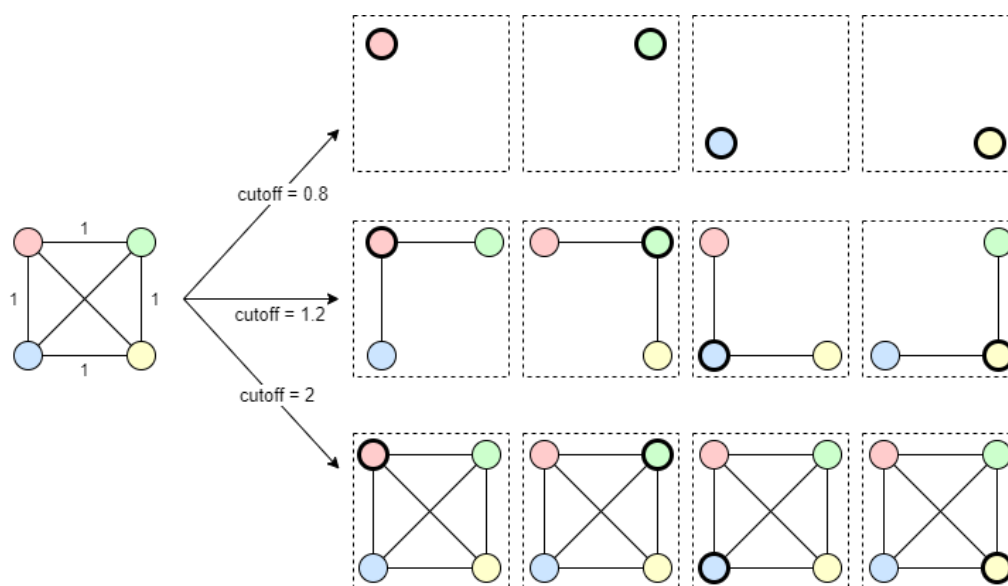
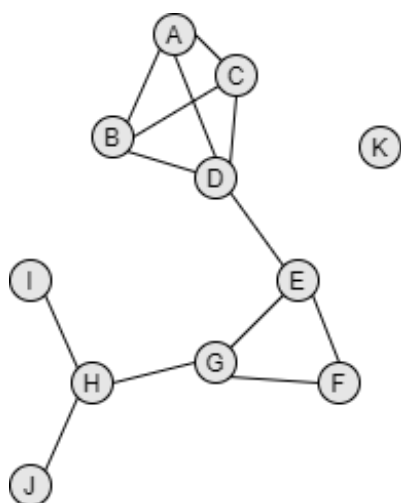


Figura 9 - *Subsets* de cada elemento para diferentes valores de *cutoff*

A Figura 9 ilustra os *subsets* de cada elemento destacado pela linha mais grossa, a depender do *cutoff*. O elemento vermelho, por exemplo, cobre apenas a si mesmo quando o *cutoff* é menor do que qualquer distância à outro elemento; cobre a si mesmo e aos elementos azul e verde quando o *cutoff* é maior do que o lado do quadrado e menor do que a diagonal do quadrado; e cobre todos os elementos quando o *cutoff* é maior do que a maior distância entre dois elementos do conjunto.

Note que o tamanho dos *subsets* é diretamente dependente do *cutoff* e da etapa de poda. Se o *cutoff* for muito próximo de 0, os *subsets* de muitos elementos conterão apenas eles mesmos, sendo obrigados a estar no conjunto solução para que se obtenha cobertura completa. Desta forma, a solução terá quase a totalidade de cenários. Se o *cutoff* for muito próximo da maior distância entre dois vértices, os *subsets* de muitos elementos conterão uma grande quantidade de outros elementos, podendo indicar uma falsa similaridade entre os pares e o conjunto solução terá 1 (ou próximo de 1) cenário.

Adotando um *cutoff* intermediário, o grafo completo será podado e as arestas remanescentes formarão os *subsets* de tamanhos variados e elementos dispersos, como mostra a Figura 10 para um gráfico ilustrativo genérico.



$$\begin{aligned}
 S_A = S_B = S_C &= \{A, B, C, D\} \\
 S_D &= \{D, A, B, C, H\} \\
 S_E &= \{E, D, F, G\} \\
 S_F &= \{F, E, G\} \\
 S_G &= \{G, E, F, H\} \\
 S_H &= \{H, G, I, J\} \\
 S_I &= \{I, H\} \\
 S_J &= \{J, H\} \\
 S_K &= \{K\}
 \end{aligned}$$

Figura 10 - *Subsets* dos elementos do grafo podado

Aplicação do Modelo de Cobertura de Conjuntos – Definidos os *subsets* dos vértices, é aplicado o algoritmo para solução do Problema de Cobertura de Conjuntos (PCC), em inglês *Set Covering Problem*, sendo necessária a cobertura completa para que todos os cenários sejam representados na solução final. Isto

implica que todos os cenários originais devem ser encontrados pelo menos uma vez dentro de algum dos *subsets* selecionados no modelo. A descrição matemática do tradicional PCC adaptado para a finalidade de redução de cenários possui os seguintes parâmetros:

$S = \{s_1, s_2, \dots, s_n\}, \text{ onde } s_d \in S$	Conjunto universo de cenários
$d: S \times S \rightarrow \mathbb{R}^+$	Função de distância entre cenários
$\rho \in \mathbb{R}^+$	Cutoff
$E = \{(s, s') s \in S \wedge s' \in S\}$	Arestas
$C_s = \{s\} \cup \{s' (s, s') \in E \wedge f(s, s') \leq \rho\}$	Critério de cobertura
$x_s \in [0,1]$	Variáveis de decisão: $\begin{cases} 1, \text{ se o cenário for selecionado} \\ 0, \text{ caso contrário} \end{cases}$

Função objetivo:

$$\text{Minimizar } z = \sum_{s \in S} x_s \quad (16)$$

Sujeito a

$$\sum_{s' \in C_s} x_{s'} \geq 1 \quad \forall s \in S \quad (17)$$

$$x_s \in \{0,1\} \quad \forall s \in S \quad (18)$$

onde (16) minimiza a quantidade de cenários (*subsets*) selecionados, (17) refere-se à obrigatoriedade de cobertura completa, ou seja, cada cenário deverá ser coberto por pelo menos um outro cenário, e (18) são as restrições de integralidade.

Neste estudo, foram utilizadas duas abordagens para solução do problema. O modelo exato do PCC, em sua forma clássica; e um algoritmo heurístico construtivo que realiza as seguintes etapas:

1. Cria uma lista de elementos ainda não cobertos pelos cenários que já estão no conjunto solução (na primeira iteração a lista terá todo o conjunto de cenários, e eles serão removidos gradativamente)
2. Ordena decrescentemente os *subsets* pela quantidade de elementos ainda não cobertos
3. Inclui o primeiro da lista no conjunto solução
4. Atualiza a lista de elementos não cobertos
5. Retorna ao passo (2). Critério de parada: lista de elementos não cobertos vazia.

O resultado esperado é um conjunto reduzido de cenários, onde todos os cenários do conjunto total sejam cobertos por pelo menos 1 cenário selecionado. A Figura 11 apresenta o conjunto solução para o grafo podado da Figura 10. Note que todos os elementos estão ligados a pelo menos um dos elementos da solução, destacados na cor laranja. As arestas pontilhadas são descartadas pois não são ligadas a nenhum elemento da solução.

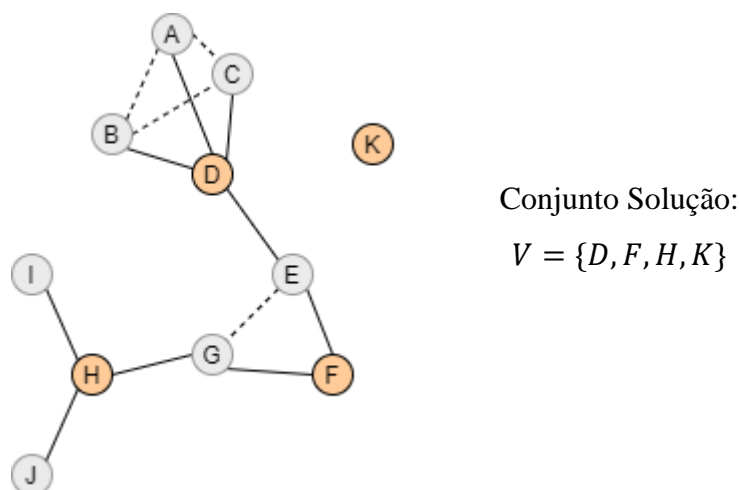


Figura 11 - Elementos representativos do conjunto total de vértices

Atribuição de pesos aos cenários representativos com base no potencial de representação – Por fim, a redução do número de cenários será balanceada já na aplicação do cálculo do objeto desejo do estudo em que a metodologia será aplicada. No caso trazido neste trabalho, o objeto desejo é a demanda estocástica de ferramentas e serviços para a construção de poços marítimos de exploração de petróleo, onde o cenário consiste em um cronograma estocástico, com uma demanda individual pré-calculada.

Os cenários selecionados serão ponderados pelo tamanho do seu *subset*. Logo, se o cálculo desejado originalmente for, por exemplo, uma média com o desvio padrão, ao utilizar a metodologia proposta será necessário calcular a média ponderada e o desvio padrão ponderado. Vale atentar que, como um cenário pode ser representado mais de uma vez no conjunto solução, o somatório dos pesos pode ser maior que a quantidade total de cenários.

Sob o mesmo fundamento, foi avaliada a possibilidade de utilização do Problema de Cobertura de Conjuntos Ponderado, ao invés de utilizar o problema

em sua forma tradicional e realizar a ponderação apenas no cálculo final. Esta alternativa foi desconsiderada pois, ela própria natureza do problema, em sua versão tradicional, todos os elementos serão representados pelo menor grupo possível de representantes, implicando que serão selecionados aqueles que dão ao conjunto o maior poder de cobertura, garantindo a cobertura completa. Por isso, a quantidade de elementos que um cenário cobre não precisaria ser considerada, e sua utilização poderia ainda restringir a área de solução do problema.

3.5 Limitações da Pesquisa

Como mencionado anteriormente, as distribuições de probabilidade das incertezas atribuídas às atividades do cronograma determinístico para o cálculo da demanda estocástica, objeto de estudo deste trabalho, não são conhecidas. Desta forma, os resultados obtidos através da aplicação da metodologia proposta serão comparados com os resultados de demanda calculados com todos os 3000 cenários, além de comparados entre si (método heurístico e exato). A validação da metodologia por completo demandaria, ainda, verificar se as distribuições de probabilidade foram honradas pelos cenários selecionados, etapa a ser realizada em trabalhos futuros.

Além disso, na etapa de atribuição de pesos aos cenários representativos com base no potencial de representação (última etapa da metodologia) exige o cálculo dos indicadores ponderados, incluindo os percentis. Pela limitação de tempo e por se tratar apenas da validação da metodologia proposta, optou-se por simplesmente replicar os cenários selecionados no PCC pelo número de elementos contidos nos seus respectivos *subsets*, garantindo o cálculo correto dos indicadores da demanda utilizando as funções padrão, não ponderadas.

4

Análise de Resultados

Este capítulo sintetiza os resultados obtidos a partir da implementação da metodologia. Foram utilizados dois métodos para solução do Problema de Cobertura de Conjuntos, o exato e o heurístico, com 7 diferentes valores de *cutoff*. Os resultados obtidos através dos dois métodos serão apresentados separadamente e, por fim, todos os resultados serão comparados entre si, e comparados com a demanda calculada sem seleção de cenários.

4.1

Construção computacional

Toda a metodologia foi implementada em etapas sequenciadas, no ambiente de desenvolvimento integrado Spyder em sua versão 4.1.5, na linguagem Python versão 3.8, com exceção da etapa de cálculo final, pois dependerá da finalidade do uso da metodologia.

A matriz de distâncias foi calculada com o pacote Gower e o modelo exato do *Set Covering* foi executado pelo Gurobi 9.1 através do Gurobipy, interface do solver para o Python.

As etapas seguintes ao cálculo da matriz de distâncias foram executadas para vários valores de *cutoff* a fim de se permitir uma análise de sensibilidade dos resultados, e após a definição dos conjuntos de cobertura, o código é bifurcado em um algoritmo para solução heurística e algoritmo para solução exata. Assim, utilizando 7 diferentes valores de *cutoff* em cada método, o código foi executado 14 vezes, como ilustrado na Tabela 5.

Os *cutoffs* foram determinados com base nos valores dos elementos na matriz de distância, onde a maior distância de Gower entre dois elementos é de 0,5. Para que a poda não fosse demasiadamente branda e houvesse a chance de manter conectados uma grande quantidade de vértices, foi determinado que a distância máxima estudada seria até 50% do valor da máxima distância, limitando os *cutoffs*

ao valor de 0.25. Os menores valores foram então espaçados de 0.025 até alcançar 0.1. De 0.1 a 0.25 foram espaçados de 0.05, chegando ao total de 7 *cutoffs*.

Tabela 5 - Execuções dos algoritmos com os valores de *cutoff*

Método Exato	Método Heurístico
1. <i>Cutoff</i> = 0.025	8. <i>Cutoff</i> = 0.025
2. <i>Cutoff</i> = 0.05	9. <i>Cutoff</i> = 0.05
3. <i>Cutoff</i> = 0.075	10. <i>Cutoff</i> = 0.075
4. <i>Cutoff</i> = 0.1	11. <i>Cutoff</i> = 0.1
5. <i>Cutoff</i> = 0.15	12. <i>Cutoff</i> = 0.15
6. <i>Cutoff</i> = 0.2	13. <i>Cutoff</i> = 0.2
7. <i>Cutoff</i> = 0.25	14. <i>Cutoff</i> = 0.25

Para o cálculo final das demandas estocásticas simuladas, com as diferentes quantidades de cenários, e a comparação com a demanda calculada com todos os cenários originais, foi escolhida a plataforma RStudio devido à necessidade de cálculos estatísticos robustos para validação dos resultados obtidos. Também no R foram realizadas as mesmas análises do estudo preliminar, selecionando as n primeiras linhas do conjunto e n cenários aleatórios, porém utilizando o mesmo número de cenários selecionados obtido no método heurístico, para comparação.

4.2 Método exato

A implementação feita em Python contou com um parâmetro de limite de tempo (LT) de 10 horas para o algoritmo exato. Esta medida foi adotada pois na fase de testes, utilizando os dados coletados no exemplo aplicado e o *cutoff* igual a 0.1, o tempo de execução foi superior a 2 dias. No caso em questão, não houve ganho significativo na função objetivo após ultrapassadas 8 horas que justificasse o longo período de execução, e os resultados dos testes com limite de tempo de 10 horas apresentaram a acurácia necessária para a proposta do trabalho e avaliar a metodologia.

Na Tabela 6 são apresentadas as quantidades de cenários selecionados pelo modelo exato para cada valor de *cutoff* e o tempo de execução do modelo. O tempo apresentado refere-se apenas à etapa de aplicação do Modelo de Cobertura de

Conjuntos. As etapas iniciais são comuns a todos os *cutoffs*, e as etapas poda do grafo e definição dos conjuntos de cobertura foram implementadas em conjunto. Esta parte do algoritmo foi executado em 3 minutos, que devem ser adicionados ao tempo individual dos *cutoffs* para a avaliação do tempo total.

Tabela 6 - Quantidades de cenários selecionados pelo algoritmo exato para cada *cutoff* e seus tempos de execução

Cutoff	Quantidade de cenários	Tempo de execução do algoritmo do PCC
0,025	2937	< 1seg
0,05	1100	< 1seg
0,075	215	> 10 horas (LT)
0,1	56	> 10 horas (LT)
0,15	9	6min 33seg
0,2	3	1min 38seg
0,25	2	1min 55seg

Foram salvos os índices dos cenários selecionados e os seus respectivos *subsets*. A demanda estocástica foi calculada, então, a partir do *dataset* contendo as demandas individuais de cada cenário (cronograma) selecionado pelo modelo do PCC. Logo, foram obtidas 7 demandas estocásticas para os 7 diferentes *cutoffs*, para cada ferramenta. As mesmas estatísticas utilizadas na análise preliminar foram aplicadas no resultado final e serão apresentadas a seguir. Um oitavo *cutoff* foi adicionado ao gráfico, com valor de 0.000000001. Ele representa o caso em que o *cutoff* é muito próximo de 0 e todo o conjunto universo, de 3000 cenários, é selecionado. Os gráficos de demanda média diária e desvio padrão sobre a média para cada cutoff testado mostram que a variabilidade das demandas com diferentes números de cenários aumentou se comparado aos métodos de seleção dos n primeiros cenários e n cenários aleatórios obtidos na análise preliminar. Os gráficos correspondentes à média das demandas diárias e ao desvio padrão diário sobre a média diária, respectivamente, encontram-se na Figura 33 e Figura 34 do Apêndice. Neles, nota-se que a variabilidade das demandas com diferentes números de cenários aumentou se comparado aos métodos de seleção dos n primeiros cenários e n cenários aleatórios obtidos na análise preliminar.

Em seguida, a Figura 12 apresenta o mesmo gráfico da média diária para os diferentes *cutoffs*, com um recorte da ferramenta A, para uma visualização mais detalhada da curva.

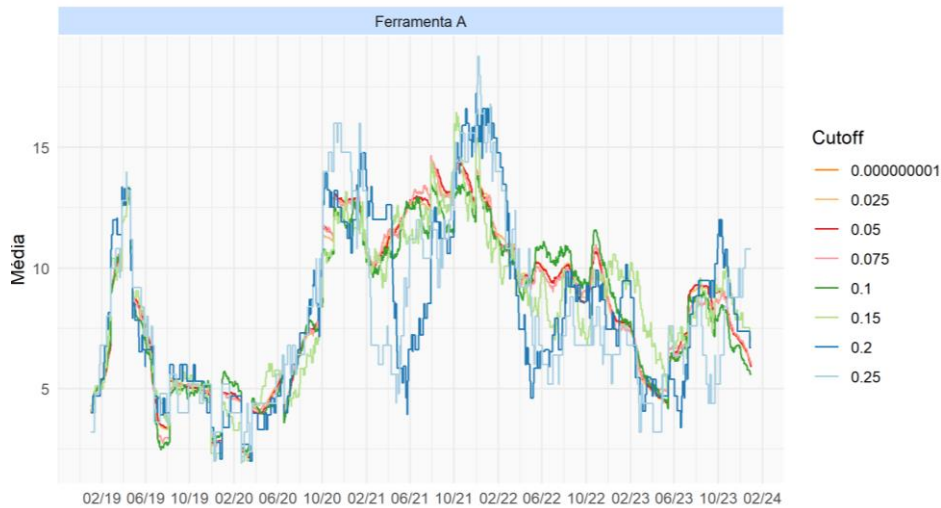


Figura 12 - Demandas médias diárias calculadas com os cenários selecionados por cada *cutoff*, para a ferramenta A

Como o esperado, observa-se que as linhas azuis e verde-claro, correspondentes aos *cutoffs* 0.25, 0.2 e 0.15, apresentam a maior distância da linha laranja, a curva de referência. Mesmo que em parte do tempo estas curvas estejam paralelas e próximas à referência, elas trazem picos e vales com maior amplitude, indicando a possibilidade de haver alguns pontos do horizonte de tempo onde há maior variabilidade no *dataset* original. Sendo selecionado um número muito reduzido de cenários selecionados (2, 3 e 9, respectivamente), seus *subsets* não se mostram capazes de representar o conjunto universo nestes intervalos críticos. O gráfico para o desvio padrão sobre a média diário, apresentado na Figura 13, reafirma este indício.

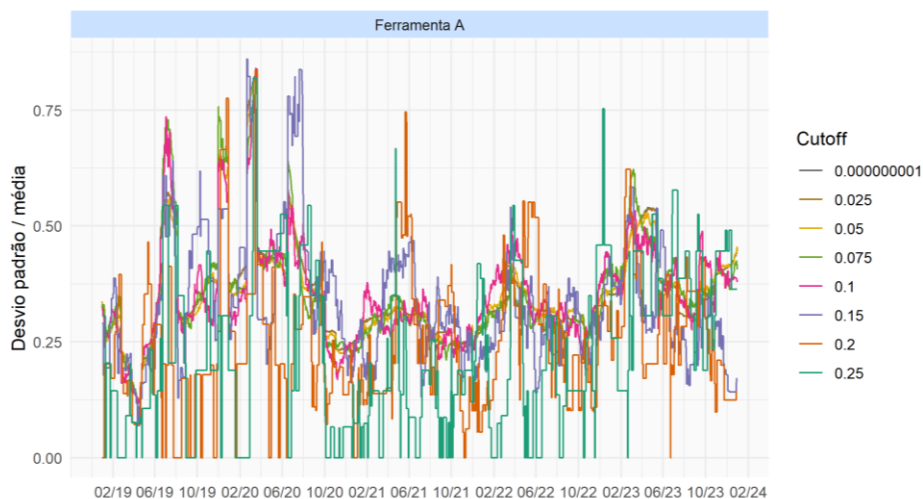


Figura 13 - Desvios padrões diários das demandas divididos pelas demandas médias diárias calculadas com os cenários selecionados por cada *cutoff*, para a ferramenta A

As curvas de maior valor de *cutoff* foram, então, removidas para análise das curvas de maior número de cenários. Variando de 56 a 2937 cenários, as curvas resultaram em um conjunto de médias muito semelhantes (Figura 12). A série referente ao *cutoff* 0.1 apresentou desvios um pouco mais acentuados que as demais nos picos e vales da média, bem como um leve descolamento do aglomerado de curvas formado nos gráficos de desvio padrão, desvio padrão sobre a média, percentis e variância, apresentados nas Figura 35 a Figura 41 do Apêndice.

No entanto, apesar da similaridade entre as curvas de *cutoff* restantes e a curva de referência e o pequeno desvio da curva de *cutoff* 0.1, a visualização gráfica não é capaz de orientar a análise pela exclusão desta ou escolha de uma das outras como o melhor *cutoff*. Isto se dá, inclusive, pois simplesmente selecionar a curva mais próxima da referência vai de encontro ao objetivo deste trabalho, uma vez que quando mais cenários mais próximas seriam.

Foi calculada, então, a distância entre as curvas de cada indicador para cada dia do horizonte de tempo, em relação à curva gerada pelo conjunto universo. Foi medida a distância absoluta, ponto a ponto, e calculada a média, conforme a fórmula (14) Na sequência, o gráfico da Figura 14 e a Tabela 7 de distâncias absolutas médias para cada indicador.

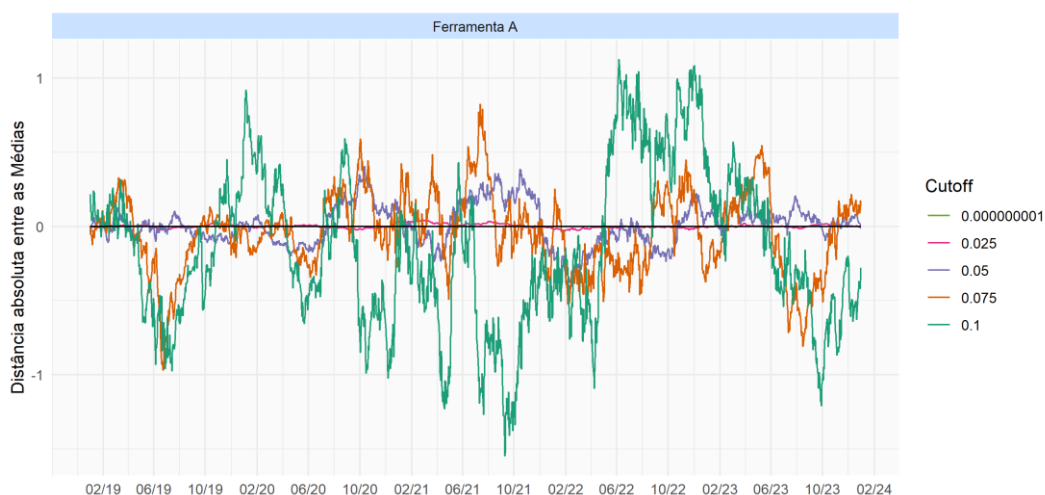


Figura 14 - Distâncias absolutas das demandas médias diárias calculadas com os cenários selecionados pelos *cutoffs* de valor até 0.1 e a demanda média diária calculada com todos os cenários, no método exato, para a ferramenta A

Tabela 7 - Distâncias absolutas médias das curvas referentes a cada indicador, por *cutoff* no método exato, para a ferramenta A

Cutoff (# Cenários)	Média	Desvio Padrão	DesvPad / Média	P10	P50	P90	Variância
0.025 (2937)	0,001	-0,002	0,000	0,004	0,008	-0,001	-0,014
0.05 (1100)	0,010	-0,009	-0,003	-0,009	-0,065	-0,044	-0,002
0.075 (215)	-0,043	-0,046	-0,001	-0,065	-0,083	-0,088	-0,253
0.1 (56)	-0,158	-0,018	0,001	-0,119	-0,237	-0,062	0,082
0.15 (9)	-0,079	-0,146	-0,020	-0,214	0,025	0,033	-0,166
0.2 (3)	-0,269	-0,880	-0,104	0,902	-0,062	-1,695	-3,493
0.25 (2)	-0,405	-1,086	-0,120	1,169	0,084	-2,338	-3,609

A distância absoluta nos ajuda a comparar resultados, mas não reflete, porém, o quanto os valores encontrados são representativos uma vez que a média não é considerada. Uma visão mais adequada seria obtida através da análise da distância relativa. Ela é a medida da distância absoluta, ponto a ponto, sobre a média da curva de referência neste mesmo ponto, seguindo a fórmula (15). A média das distâncias relativas de todo o horizonte de tempo é apresentada na Figura 15, onde a linha escura constante em zero representa a referência, e os valores para cada indicador estão consolidados na Tabela 8.

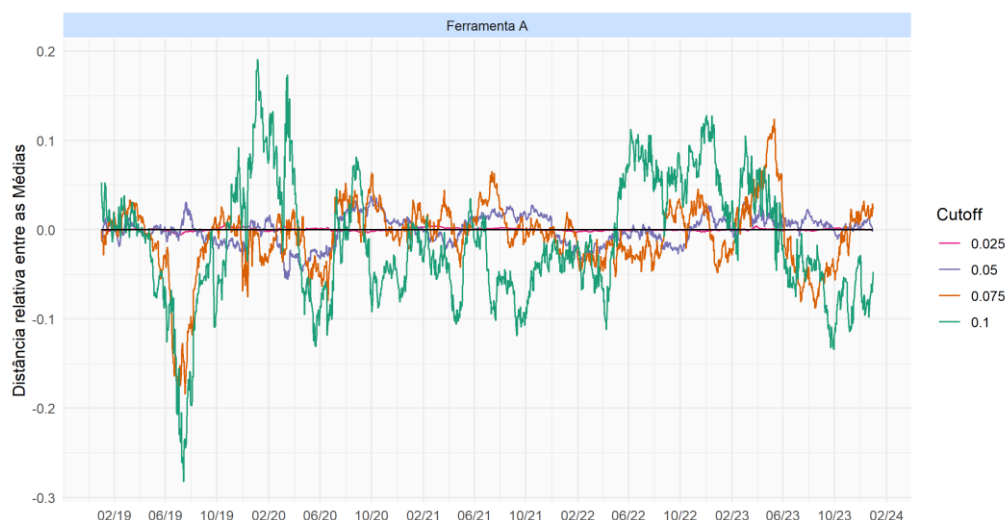


Figura 15 - Distâncias relativas das demandas médias diárias calculadas com os cenários selecionados pelos *cutoffs* de valor até 0.1 e a demanda média diária calculada com todos os cenários, no método exato, para a ferramenta A

Tabela 8 - Distâncias relativas médias das curvas referentes a cada indicador, por *cutoff* no método exato, para a ferramenta A

Cutoff (# Cenários)	Média	Desvio Padrão	DesvPad / Média	P10	P50	P90	Variância
0.025 (2937)	0,01%	-0,08%	-0,08%	0,18%	0,16%	-0,02%	-0,15%
0.05 (1100)	-0,03%	-0,64%	-0,59%	0,34%	-0,79%	-0,99%	-1,15%
0.075 (215)	-0,73%	-1,44%	-0,53%	-2,65%	-0,42%	-0,83%	-2,34%
0.1 (56)	-1,49%	-1,19%	0,55%	-0,76%	-0,78%	-0,51%	-1,00%
0.15 (9)	0,60%	-5,31%	-5,18%	1,11%	2,06%	1,38%	-1,72%
0.2 (3)	-2,26%	-33,33%	-27,74%	26,57%	1,41%	-14,42%	-45,46%
0.25 (2)	-3,38%	-38,52%	-35,87%	25,98%	6,27%	-18,55%	-43,03%

Foram marcados na tabela todos os valores acima de 5%. Este foi o valor limite escolhido para comparar o resultado obtido pelas curvas nos diferentes métodos testados neste trabalho. Foi escolhido um valor baixo considerando que as diferenças positivas e negativas podem se cancelar no cálculo da média de todos os dias do horizonte de tempo. Com base nestes valores, os *cutoffs* de valor 0.15, 0.2 e 0.25 podem ser desconsiderados, pois apresentaram variação maior que o limite estipulado em pelo menos um dos indicadores.

A análise individual feita para a ferramenta A também foi feita para as outras 13 ferramentas, que apresentaram comportamento similar, considerando as particularidades da demanda de cada ferramenta.

4.3 Método heurístico

Um procedimento análogo ao realizado para o método exato foi feito para o método heurístico. Sem limite de tempo, os resultados foram significativamente próximos do exato se considerada a simplicidade da heurística gulosa aplicada, e o tempo de execução mais homogêneo quando comparado entre os *cutoffs*. A Tabela 9 apresenta a quantidade de cenários selecionados para cada *cutoff*, e os tempos de execução do algoritmo do PCC e do algoritmo completo. O tempo anterior à execução do algoritmo do PCC foi o mesmo do método exato, de 3 minutos em média.

Tabela 9 - Quantidades de cenários selecionados pelo algoritmo heurístico para cada *cutoff* e seus tempos de execução

Cutoff	Quantidade de cenários	Tempo de execução do algoritmo do PCC	
		Método Exato	Método Heurístico
0,025	2937	< 1seg	1min 39seg
0,05	1167	< 1seg	1min 15seg
0,075	270	> 10 horas (LT)	23seg
0,1	74	> 10 horas (LT)	11seg
0,15	13	6min 33seg	4seg
0,2	4	1min 38seg	2seg
0,25	2	1min 55seg	1seg

Com os índices dos cenários (cronogramas) selecionados e os seus respectivos *subsets*, foi calculada a demanda estocástica para cada valor de *cutoff*. Foram obtidas as 7 demandas estocásticas para os 7 diferentes *cutoffs*, para cada ferramenta, utilizando os mesmos indicadores apresentados na análise preliminar. Assim como no método exato, um oitavo *cutoff*, um valor bem pequeno, foi adicionado ao gráfico como uma referência. Ele representa o caso em que o *cutoff* é muito próximo de 0 e todo o conjunto universo, de 3000 cenários, é selecionado.

As figuras Figura 42 e Figura 43 do Apêndice correspondem às demandas médias diárias e ao desvio padrão diário sobre a média diária, respectivamente. Observa-se que a variabilidade das demandas com diferentes números de cenários aumentou em comparação com os métodos de seleção dos n primeiros cenários e n

cenários aleatórios obtidos na análise preliminar, assim como observado no método exato.

As análises novamente indicaram que, em todas as ferramentas, as curvas com maior quantidade de cenários (*cutoff* 0.25, 0.2 e 0.15) apresentam uma distância relativamente grande da curva de referência. Fazendo, então, o mesmo recorte do *dataset* para a ferramenta A, com a quantidade de cenários variando entre 74 e 2937, foram gerados os gráficos dos indicadores demanda média diária e desvio padrão sobre a média para cada valor de *cutoff* (Figura 16 e Figura 17). Os demais indicadores, assim como no método exato, estão disponíveis nas Figura 44 a Figura 48 do Apêndice.

Novamente, as curvas restantes se mostraram muito semelhantes à curva de referência. Um pequeno desvio da curva de *cutoff* 0.1 também pode ser visualizado através da solução heurística, mas a visualização gráfica não é capaz de orientar a análise pela exclusão desta ou escolha de uma das outras como o melhor resultado. Seguindo a mesma premissa adotada anteriormente, foi calculada, então, a distância entre as curvas de cada indicador para cada dia do horizonte de tempo, em relação à curva gerada com todo o conjunto universo. Foi medida a distância absoluta, ponto a ponto, e calculada a média, conforme a mesma fórmula apresentada no método exato. A Figura 18 e a Tabela 10 mostra as distâncias absolutas médias para cada indicador.

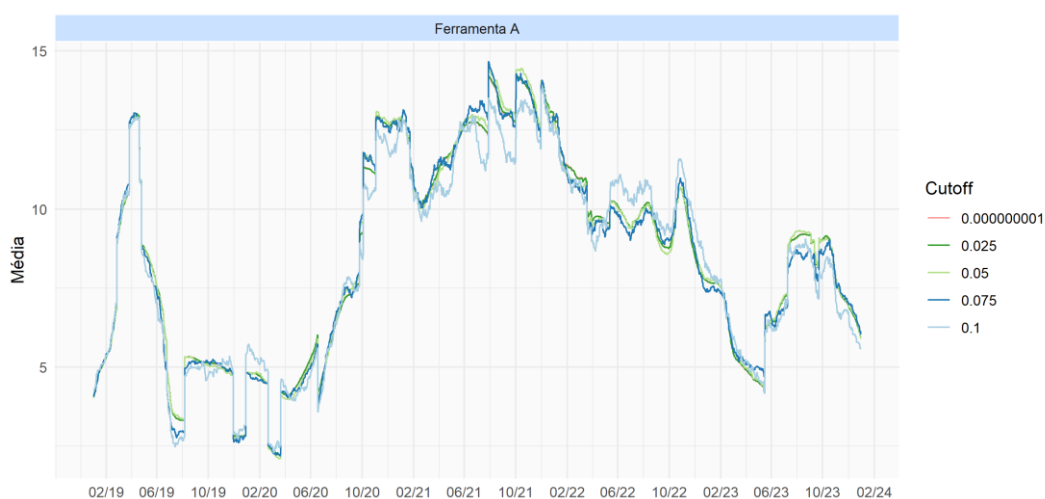


Figura 16 - Demandas médias diárias calculadas com os cenários selecionados pelos *cutoffs* até 0.1 no método heurístico, para a ferramenta A

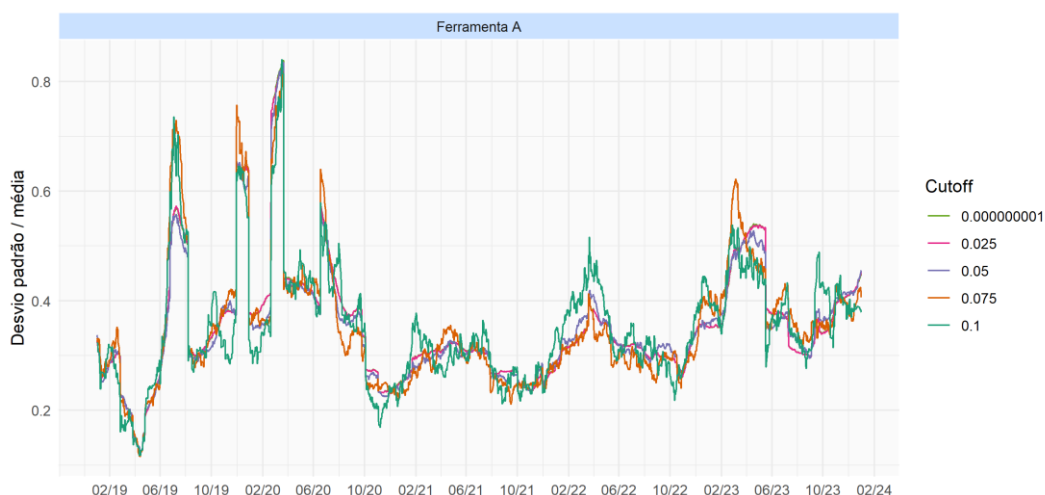


Figura 17 - Desvio padrão dividido pela média diário das demandas calculadas com os cenários selecionados pelos *cutoffs* até 0.1 no método heurístico, para a ferramenta A

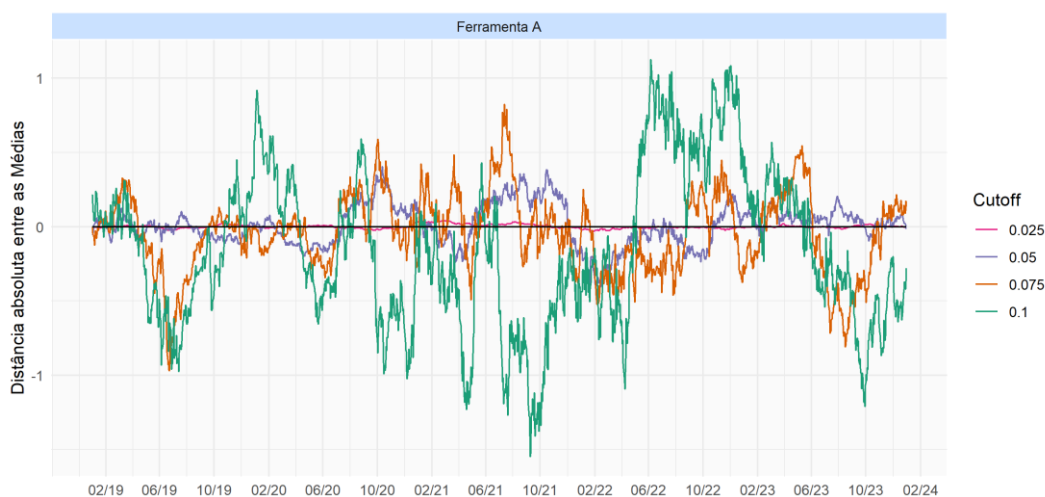


Figura 18 - Distâncias absolutas das demandas médias diárias calculadas com os cenários selecionados pelos *cutoffs* de valor até 0.1 e a demanda média diária calculada com todos os cenários, no método heurístico, para a ferramenta A

Tabela 10 - Distâncias absolutas médias das curvas referentes a cada indicador, por *cutoff* no método heurístico, para a ferramenta A

Cutoff (# Cenários)	Média	Desvio Padrão	DesvPad / Média	P10	P50	P90	Variância
0.025 (2937)	0,000	0,000	0,000	-0,001	0,002	-0,004	0,003
0.05 (1167)	0,019	-0,011	-0,002	0,095	0,005	0,025	-0,058
0.075 (270)	0,046	-0,044	-0,011	0,074	0,071	0,023	-0,184
0.1 (74)	-0,044	-0,082	-0,005	0,007	0,014	-0,133	-0,384
0.15 (13)	0,089	-0,186	-0,030	-0,190	0,350	0,102	-0,762
0.2 (4)	0,079	-0,539	-0,062	0,660	0,367	-0,694	-1,535
0.25 (2)	0,040	-1,571	-0,194	2,205	0,369	-2,337	-5,369

Foi realizada a análise da distância relativa, que consiste na distância absoluta dividida a média da curva de referência, calculada dia a dia. A média das distâncias relativas de todo o horizonte de tempo é apresentada na Figura 19, onde a linha escura constante em zero representa a referência, e os valores para cada indicador estão consolidados na

Tabela 11.

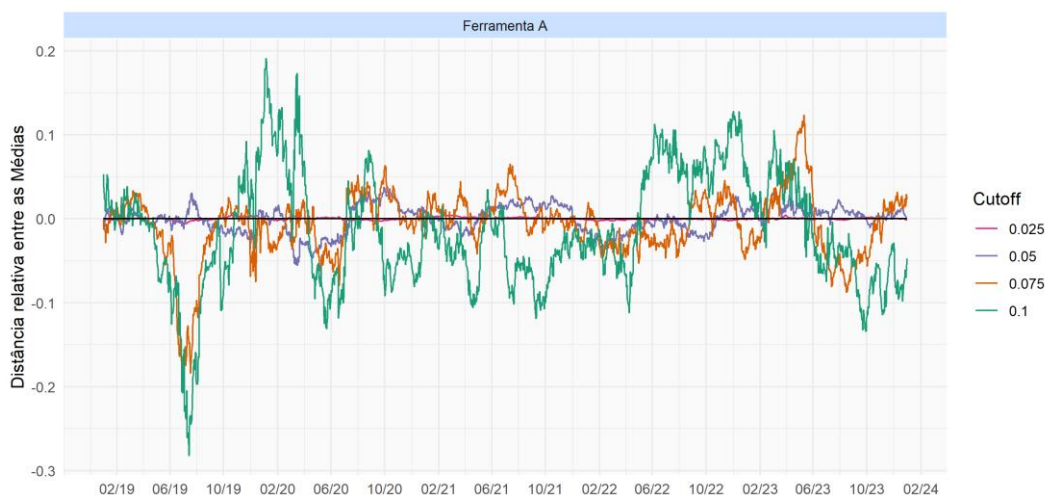


Figura 19 - Distâncias relativas das demandas médias diárias calculadas com os cenários selecionados pelos *cutoffs* de valor até 0.1 e a demanda média diária calculada com todos os cenários, no método heurístico, para a ferramenta A

Tabela 11 - Distâncias relativas médias das curvas referentes a cada indicador, por *cutoff* no método heurístico, para a ferramenta A

Cutoff (# Cenários)	Média	Desvio Padrão	DesvPad / Média	P10	P50	P90	Variância
0.025 (2937)	0,01%	0,01%	-0,01%	0,14%	0,12%	-0,02%	0,01%
0.05 (1167)	0,15%	-0,41%	-0,54%	1,99%	0,40%	0,08%	-0,75%
0.075 (270)	1,01%	-1,85%	-2,77%	3,46%	2,41%	0,20%	-3,27%
0.1 (74)	-1,19%	-3,07%	-1,65%	-0,91%	0,62%	-1,90%	-5,21%
0.15 (13)	1,51%	-7,17%	-7,91%	1,31%	6,53%	0,58%	-11,22%
0.2 (4)	-0,40%	-21,44%	-18,68%	15,17%	4,82%	-6,99%	-23,80%
0.25 (2)	3,22%	-58,82%	-56,76%	61,89%	6,73%	-19,34%	-70,29%

Foram marcados na tabela todos os valores acima de 5%, como no método exato. Diferentemente do método exato, o *cutoff* 0.1, que selecionou 74 cenários,

apresentou uma variação maior que 5% na média da variância. No método heurístico, o melhor candidato seria o *cutoff* de valor 0.075.

A análise individual feita para a ferramenta A também foi feita para as outras 13 ferramentas, que apresentaram comportamento similar, considerando as particularidades da demanda de cada ferramenta.

4.4 Comparação entre os métodos

Coletados os resultados individuais da aplicação da metodologia com o uso dos métodos exato e heurístico para a solução do PCC, esses podem ser não apenas comparados com o resultado calculado com o conjunto completo de cenários, mas também comparados entre si. A Tabela 9 mostrou que o tempo computacional para o método heurístico foi consideravelmente menor que o método exato, como esperado. Os gráficos das Figura 20 a Figura 26 apresentam as médias encontradas para a Ferramenta A em cada *cutoff* em cada método, dispostos lado a lado. A sequência de gráficos mostra, em ordem, o impacto da diminuição do número de cenários escolhido. Foi incluída em cada um deles a curva de referência, com a média diária da demanda com os 3000 cenários.

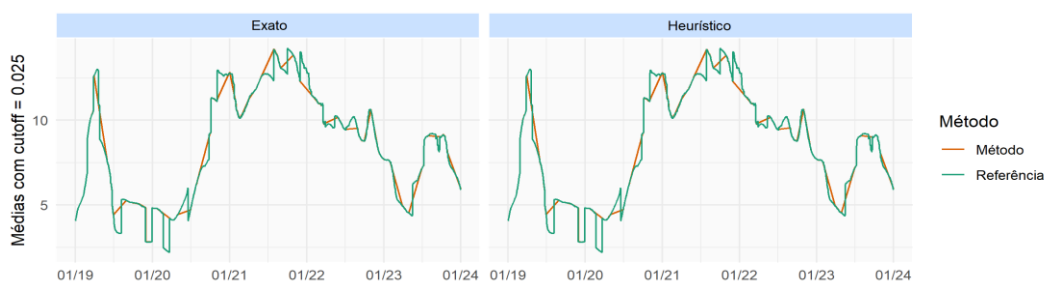


Figura 20 - Demanda média encontrada em cada método, para cutoff = 0.025.



Figura 21 - Demanda média encontrada em cada método, para cutoff = 0.05



Figura 22 - Demanda média encontrada em cada método, para cutoff = 0.075

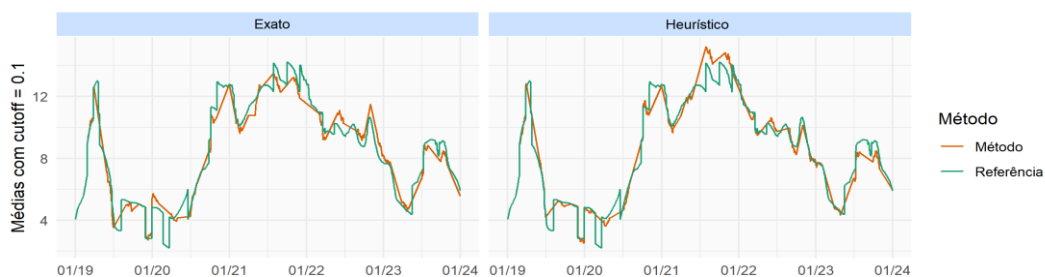


Figura 23 - Demanda média encontrada em cada método, para cutoff = 0.1

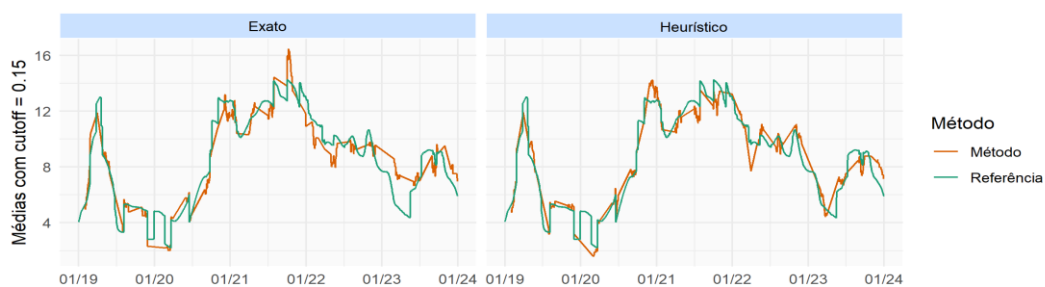


Figura 24 - Demanda média encontrada em cada método, para cutoff = 0.15



Figura 25 - Demanda média encontrada em cada método, para cutoff = 0.2

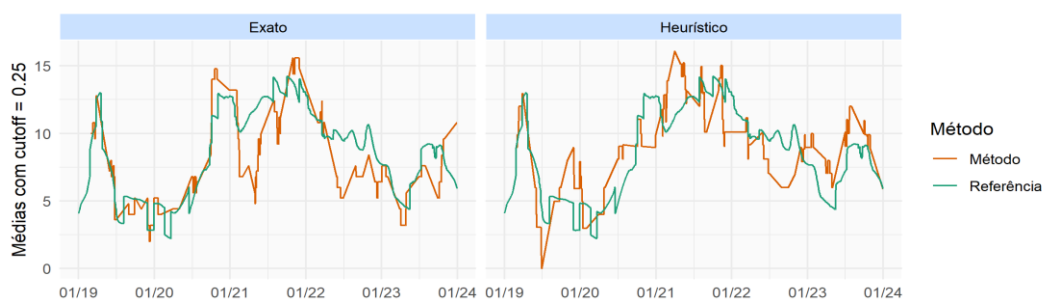


Figura 26 - Demanda média encontrada em cada método, para cutoff = 0.25

Na Tabela 12 estão dispostos os valores médios das curvas das Figura 20 a Figura 26 por *cutoff*, bem como os valores médios das curvas de cada indicador analisado neste estudo, separados por método, e da curva de Referência.

Tabela 12 - Médias totais dos resultados dos indicadores da Ferramenta A em cada método aplicado e demanda de referência, por valor de *cutoff*

Indicador	Exato	Heurístico	Referência
Cutoff = 0.025			
Media	8,525277052	8,524648045	8,524231
Desvio Padrão	2,740147508	2,742633424	2,742362
Desvio Padrão/Média	0,351084957	0,351355877	0,351366
P10	5,181161008	5,176341731	5,176999
P50	8,382256298	8,375684556	8,374042
P90	12,06276013	12,06024096	12,063855
Variância	7,978740274	7,994989368	7,992470
Cutoff = 0.05			
Media	8,534390084	8,543118452	8,524231
Desvio Padrão	2,733628295	2,731557777	2,742362
Desvio Padrão/Média	0,348489603	0,349419656	0,351366
P10	5,167579409	5,272070099	5,176999
P50	8,308871851	8,378970427	8,374042
P90	12,01971522	12,0890471	12,063855
Variância	7,990867969	7,934775167	7,992470
Cutoff = 0.075			
Media	8,481482467	8,570486236	8,524231
Desvio Padrão	2,696833413	2,698270216	2,742362
Desvio Padrão/Média	0,350467893	0,340421974	0,351366
P10	5,111719606	5,251369113	5,176999
P50	8,291347207	8,445235487	8,374042
P90	11,97634173	12,08652793	12,063855
Variância	7,739345783	7,80849541	7,992470
Cutoff = 0.1			
Media	8,366489735	8,480200051	8,524231
Desvio Padrão	2,724242942	2,660581719	2,742362
Desvio Padrão/Média	0,352188115	0,346557252	0,351366
P10	5,058050383	5,184008762	5,176999
P50	8,137458927	8,387732749	8,374042
P90	12,00219058	11,93099671	12,063855
Variância	8,074514368	7,608688855	7,992470

Observa-se que a demanda com os cenários selecionados pelo método heurístico se comportou de forma muito parecida com a gerada pelos cenários

selecionados pelo método exato no mesmo *cutoff*. É preciso lembrar que as quantidades de cenários selecionados para cada *cutoff* foram ligeiramente maiores para o método heurístico, o que certamente impacta na percepção de que o método heurístico chegou a um resultado próximo ao exato.

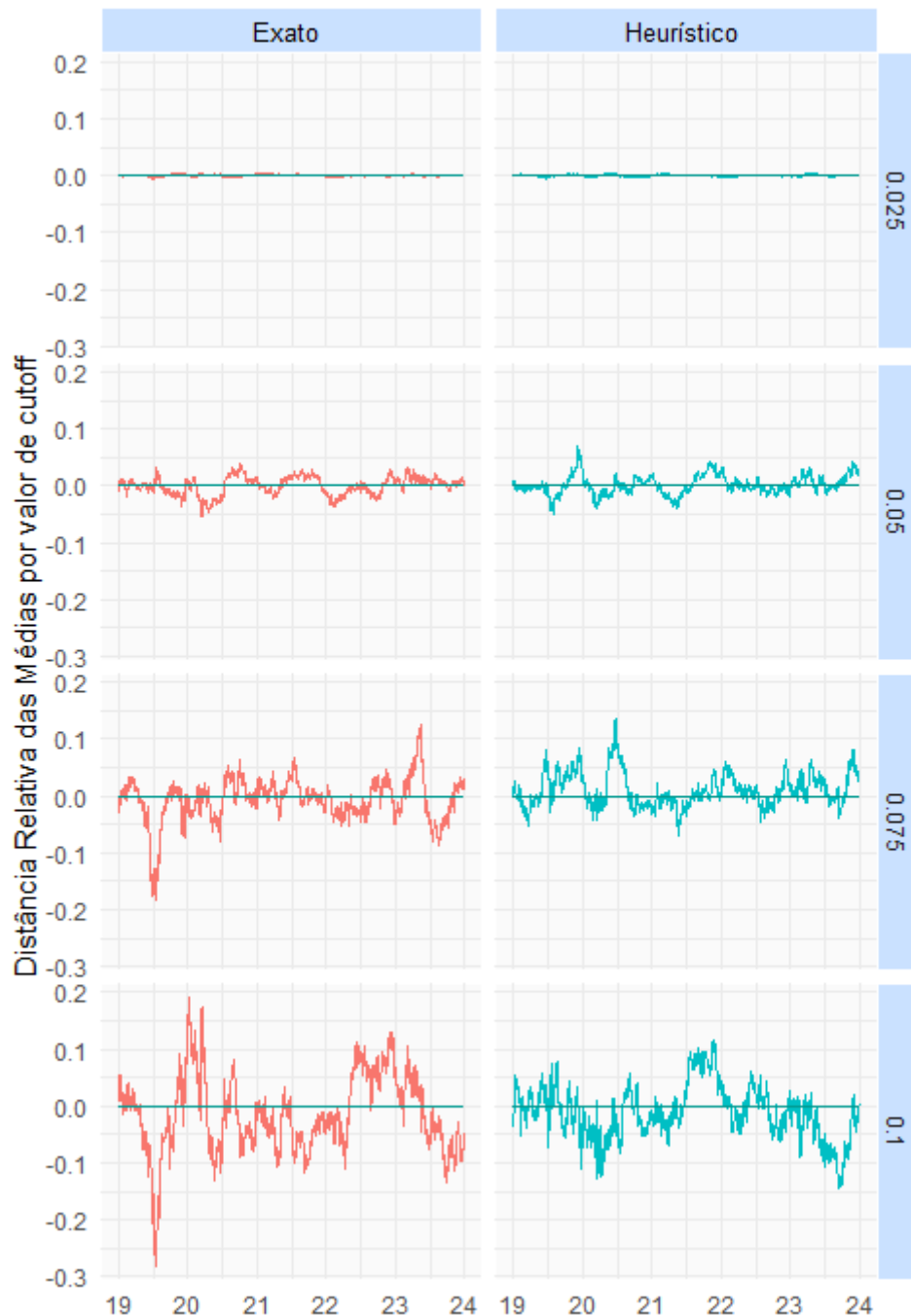


Figura 27 - Distâncias relativas entre as médias das demandas calculadas para cada *cutoff* em cada método e a média da demanda calculada com todos os cenários

A Tabela 13 reúne não apenas as médias das curvas de distância relativa da média apresentadas na Figura 27, mas também as médias das curvas de distância absoluta e relativa dos indicadores considerados mais representativos após análise: média, desvio padrão, desvio padrão dividido pela média e variância.

Tabela 13 - Médias totais das distâncias absolutas e relativas dos indicadores média, desvio padrão, desvio padrão dividido pela média e variância para cada *cutoff*, separados por método aplicado.

Indicador	Exato	Heurístico
0.025		
Distância Absoluta da Média	0,001046	0,000417
Distância Absoluta do Desvio Padrão	-0,002215	0,000271
Distância Absoluta do Desvio Padrão/Média	-0,000281	-0,000010
Distância Absoluta da Variância	-0,013730	0,002519
Distância Relativa da Média	0,000081	0,000120
Distância Relativa do Desvio Padrão	-0,000751	0,000054
Distância Relativa do Desvio Padrão/Média	-0,000830	-0,000064
Distância Relativa da Variância	-0,001495	0,000114
0.05		
Distância Absoluta da Média	0,010159	0,018887
Distância Absoluta do Desvio Padrão	-0,008734	-0,010804
Distância Absoluta do Desvio Padrão/Média	-0,002876	-0,001946
Distância Absoluta da Variância	-0,001602	-0,057695
Distância Relativa da Média	-0,000308	0,001487
Distância Relativa do Desvio Padrão	-0,006408	-0,004046
Distância Relativa do Desvio Padrão/Média	-0,005937	-0,005357
Distância Relativa da Variância	-0,011497	-0,007487
0.075		
Distância Absoluta da Média	-0,042749	0,046255
Distância Absoluta do Desvio Padrão	-0,045529	-0,044092
Distância Absoluta do Desvio Padrão/Média	-0,000898	-0,010944
Distância Absoluta da Variância	-0,253124	-0,183975
Distância Relativa da Média	-0,007281	0,010071
Distância Relativa do Desvio Padrão	-0,014407	-0,018472
Distância Relativa do Desvio Padrão/Média	-0,005299	-0,027649
Distância Relativa da Variância	-0,023429	-0,032691
0.1		
Distância Absoluta da Média	-0,157742	-0,044031
Distância Absoluta do Desvio Padrão	-0,018119	-0,081780
Distância Absoluta do Desvio Padrão/Média	0,000822	-0,004809
Distância Absoluta da Variância	0,082044	-0,383781
Distância Relativa da Média	-0,014904	-0,011927
Distância Relativa do Desvio Padrão	-0,011852	-0,030655
Distância Relativa do Desvio Padrão/Média	0,005450	-0,016488
Distância Relativa da Variância	-0,009978	-0,052059

Na Figura 27, com auxílio da Tabela 13 e considerando a curva de referência como o ideal, é possível ver a regressão da qualidade dos resultados ao diminuir a quantidade de cenários em ambos os métodos analisados, conforme o esperado. De maneira geral,

A partir deste ponto do cálculo, a questão chave se torna a definição de qual percentual de diferença do resultado obtido com todos os cenários seria aceitável frente aos ganhos em tempo de execução e quantidade de dados no processamento do cálculo. Se mantido o valor de 5% utilizado até este ponto da análise, podemos concluir que, dentre os *cutoffs* apresentados, o valor 0.1 foi o mais adequado para a aplicação da metodologia utilizando o método exato, pois foi o resultado com uma distância relativa média abaixo de 5% com o menor número de cenários. Sob o mesmo critério, o valor de 0.075 foi o melhor *cutoff* para a aplicação da metodologia com o método exato. Os valores de *cutoff* acima dos mencionados foram desconsiderados pois apresentaram variações além dos 5% em pelo menos um dos indicadores.

5 Considerações Finais

Até o momento da finalização deste trabalho, com base nas pesquisas por trabalhos correlatos na literatura, não foram localizadas publicações que abordassem um objeto complexo e com múltiplos atributos que é um cronograma, como um cenário, tão obstante um conjunto de cronogramas gerados a partir de incertezas em suas atividades a ser reduzido. É o caso da abordagem trazida neste trabalho, que se mostrou ser não apenas coerente, mas também promissora.

Na revisão bibliográfica foi possível rever alguns conceitos e características da Programação Inteira e do Problema de Cobertura de Conjuntos aprendidas ao longo do curso e aprofundar o conhecimento específico de modelos de seleção de cenários e tratamento da estocasticidade, possibilitando a definição da metodologia mais adequada para ser implementada computacionalmente.

Os resultados da aplicação, gerados na linguagem Python e validados na linguagem R, mostram a relação do número de cenários e a qualidade da demanda gerada. A demanda gerada com 56 cenários (*cutoff* igual a 0.1) escolhidos pelo método exato apresentou, em média, menos de 5% de diferença da demanda gerada com os 3000 cenários utilizados atualmente no cálculo. A demanda calculada com os cenários selecionados pelo método heurístico ficou pouco atrás em termos de qualidade. Para os valores de *cutoff* testados, a demanda que atendeu os mesmos requisitos de percentual de distância da curva referência selecionou 270 cenários (*cutoff* igual a 0.075), 91% menos que o praticado atualmente. Este valor significa uma redução de 91% na quantidade de dados processados no cálculo da demanda e no tempo de processamento desta etapa, que é linear, reduzindo apenas 5% na assertividade. Devido à simplicidade da heurística utilizada, é possível melhorar este tempo utilizando métodos mais robustos na etapa de solução do Problema de Cobertura de Conjuntos.

Como pode ser observado, neste ponto do cálculo a questão chave se torna a definição de qual percentual de diferença do resultado obtido com todos os cenários seria aceitável frente aos ganhos em tempo de execução e quantidade de

dados no processamento do cálculo. Esta questão só pode ser respondida com clareza pelo tomador de decisão, quem está aplicando a metodologia com uma finalidade, pois depende intimamente do contexto em que está inserida.

Diante dos fatos, mesmo que com pontos de melhoria, mostra-se vantajosa a utilização da metodologia proposta para a otimização do cálculo da demanda estocástica e qualquer outra aplicação com as mesmas características de cenários de incerteza. Novamente, vale frisar que o algoritmo desenvolvido jamais substituirá o trabalho do gestor que utiliza a metodologia, quem, de fato, conhece as particularidades do negócio. Pelo contrário, a metodologia melhora sua capacidade de análise e tomada de decisões.

Para a validação da metodologia como um método de seleção de cenários, no entanto, validações adicionais precisam ser feitas. Sugestões iniciais, tanto para o aprimoramento quanto para a validação da metodologia proposta, são levantadas no próximo tópico.

5.1

Sugestões para desenvolvimentos futuros

Para o aprimoramento da metodologia proposta:

- Análise de sensibilidade com uma escala menor de valores de *cutoff*. Para mesmo *dataset* utilizado neste trabalho, os valores podem estar compreendidos entre 0.075 e 0.15;
- Desenvolvimento de funções para cálculos ponderados, como o de percentis;
- Métodos metaheurísticos na etapa de solução do Problema de Cobertura de Conjuntos, como a implementação de algoritmo genético, com busca local.

Para validação da metodologia:

- Comparação com métodos tradicionais de clusterização, como o K-medoids;
- Criação de um problema de teste, onde o cenário determinístico, que deu origem aos cenários de incerteza, e as distribuições de probabilidade, são conhecidos. Aplicar a metodologia partindo de cenários já criados e comparar os resultados finais acerca da

representatividade dos cenários selecionados para o conjunto universo e também para as distribuições;

- Aplicação da metodologia em outras instâncias e base de dados, de outros setores da indústria que trabalham com cenários estocásticos, como o planejamento de matriz energética.

Bibliografia

ARENALES, M.; ARMENTANO, V.; MORABITO, R.; YANASSE, H. *Pesquisa Operacional*. Rio de Janeiro: Elsevier, 2015

BALAS, E.; HO, A. *Set covering algorithms using cutting planes, heuristics and subgradient optimization: a computational study*. Mathematical Programming Study, 1980. cap. 12, p. 37-60.

BEALE, E. *On Minizing a Convex Function Subject to Linear Inequalities*, Journal of the Royal Statistical Society, Series B (Methodological), 1955. cap. 17, 2, p. 173-184.

BERKHIN, P. *A Survey of Clustering Data Mining Techniques*. Technical Report, Accrue Software, 2002.

BIRGE, J.; LOUVEAUX, F. *Introduction to Stochastic Dynamic Programming*. 1997.

BUSSAB, W. de O.; MIAZAKI, E. S.; ANDRADE, D. F. de. *Introdução à análise de agrupamentos*. São Paulo: USP-IME, 1990.

CHARNES, A.; COOPER, W. *Chance-constrained programming*. Management Science, 1959. p. 73-79.

CHEN, Y.; LAI, Y. *Sparse dynamical Boltzmann machine for reconstructing complex networks with binary dynamics*. Physical Review E, 2018.

CHIRALAKSANAKUL, A. *Monte Carlo Methods for Multistage Stochastic Programs*. Dissertation of Philosophy Doctor. The University of Texas at Austin, 2003.

CHRYSOCHOOU, E.; ZILIASKOPOULOS, A.; LOIS, A. *An exact algorithm for the stochastic inventory routing problem with transshipment*, 2015.

CHVÁTAL, V. *A greedy heuristic for the set covering problem*. Mathematics of Operations Research, 4: 233-235, 1979.

DANTZIG, G., *Linear Programming Under Uncertainty*. Management Science, 1955. p. 1764-1769.

DE OLIVEIRA, W. L. *Redução Ótima de Cenários em Programação Estocástica. Aplicação às Vazões Afluentes aos Aproveitamentos Hidroelétricos*. Dissertação de Mestrado em Matemática. Instituto Nacional de Matemática Pura e Aplicada - IMPA, 2007.

DIÁRIO DO PORTO. *Petrobras recontrata sonda da Ocyan por mais 3 anos*. 12 de julho de 2021. Disponível em: <<https://diariodoporto.com.br/petrobras-recontrata-sonda-da-ocyan-por-mais-3-anos/>>. Acessado em 30/07/2021.

DUPAČOVÁ, J.; GRÖWE-KUSKA, N.; RÖMISCH, W. *Scenario Reduction in Stochastic Programming: An Approach Using Probability Metrics*. Springer-Verlag, 2003.

GAREY, M.R.; JOHNSON, D.S. *Computer and intractability - a guide to the theory of NP-completeness*. San Francisco, W.H. Freeman and Company, 1979.

GOWER, J. C. *A general coefficient of similarity and some of its properties*. Biometrics, 1971. v. 27, p. 857-874.

GRAY, C. F.; LARSON, E. W. *Project Management – The Managerial Process*: McGraw-Hill. International Editions, 2000.

GROWE-KUSKA, N.; HEITSCH, H.; ROMISCH, W. *Scenario reduction and scenario tree construction for power management problem*. IEEE Power Tech Conference Proceedings, 2003. V. 3, 7 pp.

HAIR, J. *Multivariate. Data Analysis*. New Jersey: Prentice Hall, 1998.

HAMACHER, S.; OLIVEIRA, F. *Introdução a Otimização Estocástica*, 2013. 90 slides. Disponível em : <<http://www.din.uem.br/sbpo/sbpo2013/pdf/arq0002.pdf>>. Acesso em 20/06/2021.

HEITSCH, H.; ROMISCH, W. *Scenario reduction algorithms in stochastic programming*. Computational Optimization and Applications, 2003. v. 24(2), p. 187-206.

HE Y, ZHAO X, ZHAO L, HE J. *Coordinating a supply chain with effort and price dependent stochastic demand*. Appl Math Model 2009. p. 2777-2790.

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. Prentice-Hall Inc., Upper Saddle River, 1988.

JAIN, A.K.; MURTY, M.N.; FLYNN, P.J. *Data Clustering: A Review*, ACM Computing Surveys, 1999. v. 31, p. 264-323.

JAIN, A.K. *Data clustering: 50 years beyond K-means*. Pattern Recognition Letters. 2009.

KARP, R. M. *Reducibility Among Combinatorial Problems*. University of California at Berkeley, 1972. Disponível em <<https://people.eecs.berkeley.edu/~luca/cs172/karp.pdf>> Acesso em: 21 fev. 2021.

KAUFMAN, L.; ROUSSEEUW, P. J. *Clustering Large Data Sets*. Pattern Recognition in Practice II, 1986.

KISIALIOU, Y.; GRIBKOVSKAIA, I.; LAPORTE, G. *Supply vessel planning with uncertain demand and weather conditions*. Workshop of the EURO Working Group on Vehicle Routing and Logistics optimization (VeRoLog), Seville (Spain), 2019.

KNIGHT, F. H. *Risk, uncertainty and profit*. Houghton Mifflin Company, New York, 1921.

KOGAN, J. *Introduction to Clustering Large and High-Dimensional Data*. New York: Cambridge, 2007.

KRANEN, M. E. *Aplicação de Técnicas de Agrupamento Sobre Séries de Potência Para Avaliação de Confiabilidade em Sistemas Hidrotérmicos*, 2004.

LABABIDI, H. M. S.; et al. *Optimizing the supply chain of a petrochemical company under uncertain operating and economic conditions*. Industrial & Engineering Chemistry Research, 2004. p. 63-73.

LANZ, L. Q.; TOMEI, P. A. *Confiança nas organizações: Como gerenciar a confiança interpessoal, organizacional e interorganizacional*, 1 ed., Rio de Janeiro: Elsevier: PUC-Rio, 2014.

Laporte, G.; Louveaux, F. V.; Van Hamme, L. *An integer L-shaped algorithm for the capacitated vehicle routing problem with stochastic demands*, 2002. p. 415-423.

LUSTOSA, L.; et al. *Planejamento e Controle da Produção*. Editora Elsevier, 2008.

MACEIRA, M. E. P.; PENNA, D. D. J.; DAMÁZIO, J. M. *Geração de Cenários Sintéticos de Energia e Vazão para o Planejamento da Operação Energética*, In: XVI Simpósio Brasileiro de Recursos Hídricos, João Pessoa, Paraíba, Brasil, 2005. p. 3-12.

MALHOTRA, N. K. *Pesquisa de marketing: uma orientação aplicada*. Porto Alegre, 2001.

MARTINS, M. M. *Comparação de estratégias de construção de poços marítimos incorporando incertezas*. Dissertação de Mestrado, PUC-Rio, 2014.

MACQUEEN, J. *Some methods for classification and analysis of multivariate observations*. Fifth Berkeley Symposium on Mathematics, Statistics and Probability. University of California Press, 1967. p. 281-297.

MCKAY, M. D.; BECKMAN, R. J.; CONOVER, W. J. *A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code*. Technometrics, 1979. v. 21, n. 2, p. 239-245.

NIEDERREITER, H. *Random Number Generation and quasi-Monte Carlo Methods*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1992.

PENNA, D. D. J.; MACEIRA, M. E. P.; DAMÁZIO, J. M. *Selective Sampling Applied to Long Term Hydrothermal Generation Planning*. In: 17TH POWER SYSTEMS COMPUTATION CONFERENCE. Sweden, 2011.

PFLUG, G. C. *Version-independence and nested distribution in multistage stochastic optimization*. SIAM Journal on Optimization, 2009. v. 20, p. 1406-1420.

PROJECT MANAGEMENT INSTITUTE (PMI). *A guide to the Project management body of knowledge (PMBOK Guide)*. 5a edição, 2013.

PROUD, J. F. *Master scheduling: practical guide to competitive manufacturing*. 2nd ed. New York, 1999.

RASMUSSEN, K. M.; CLAUSEN, J. *Mortgage loan portfolio optimization using multi-stage stochastic programming*. Journal of Economic Dynamics and Control, 2007. v. 31, p. 742-766.

RIBAS, G.; HAMACHER, S; STREET, A. *Optimization under uncertainty of the integrated oil supply chain using stochastic and robust programming*, 2010. p. 777-796.

RIBAS, G. *Modelo de programação estocástica para o planejamento estratégico da cadeia integrada de petróleo*. Thesis for Electrical Engineering Master Program – PUC-Rio, 2008.

SALAVATI-KHOSHGHALB, M.; et al. *A hybrid recourse policy for the vehicle routing problem with stochastic demands*. EURO Journal on Transportation and Logistics, 2018.

SEN, S.; HIGLE, J. *An Introductory Tutorial on Stochastic Linear Programming Models, Interfaces*, 1999. p. 33-61.

SHANMUGAM, G.; GANESAN, P.; VANATHI, P. T. *Meta heuristic algorithms for vehicle routing problem with stochastic demands*. J Comput Sci, 2011. p. 533-542.

SILVEIRA, C. M. D. *Análise de Métodos Heurísticos de Características Gulosa*. Dissertação de Mestrado, Universidade Federal do Rio Grande do Sul, Instituto de Informática, Porto Alegre, 1999.

SLACK, N.; et al. *Administração da produção*. São Paulo: Atlas, 1999.

SORENSEN T. *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and Its Application to Analyses of the Vegetation on Danish Commons*, 1948. p. 1-34

STEINHAUS, H. *Sur la division des corp materiels en parties*. Bulletin of Acad. Polon. Sci., IV (C1. III), 1956. p. 801-804.

TENEMBAUM, Y. *International Conflicts: What is the difference between Risk and Uncertainty?* OXPOL – The Oxford University Politics Blog. 2012. Disponível em <<https://blog.politics.ox.ac.uk/international-conflicts-what-is-the-difference-between-risk-and-uncertainty/>> Acesso em: 29 jan. 2021.

TUBINO, D. F. *Planejamento e Controle da Produção: Teoria e Prática*. São Paulo: Atlas, 2007.

VALLE, A. B.; et al. *Fundamentos do Gerenciamento de Projetos*. Série Gerenciamento de Projetos. 3. Ed. São Paulo: FGV, 2014.

VASKO, F. J.; WILSON, G. R. *An efficient heuristic for large set covering problems*. Naval Research Logistics Quarterly, 1984. p. 163-171.

VIANNA, D. S. *Heurística Híbridas para o Problema da Filogenia*. Tese de Doutorado, PUC-Rio, Departamento de Informática, Rio de Janeiro, 2004.

XAVIER, V. L. *Resolução do Problema de Agrupamento Segundo o Critério de Minimização da Soma de Distâncias*. Rio de Janeiro: UFRJ/COPPE, 2012.

WANG, Q.; DONG, W. YANG, L. *A wind power/photovoltaic typical scenario set generation algorithm based on Wasserstein distance metric and revised K-medoids cluster*. Proceedings of the Chinese Society of Electrical Engineering, 2015. v. 35, issue 11, p. 2654 – 2661.

WOLSEY, L. *Integer Programming*. Wiley Series in Discrete Mathematics and Optimization: Wiley, 1998.

Apêndice

Os gráficos a seguir correspondem aos resultados de demanda obtidos para cada indicador, a cada valor de *cutoff*.



Figura 28 - Demanda média diária entre os cronogramas, por ferramenta, por quantidade de cronogramas

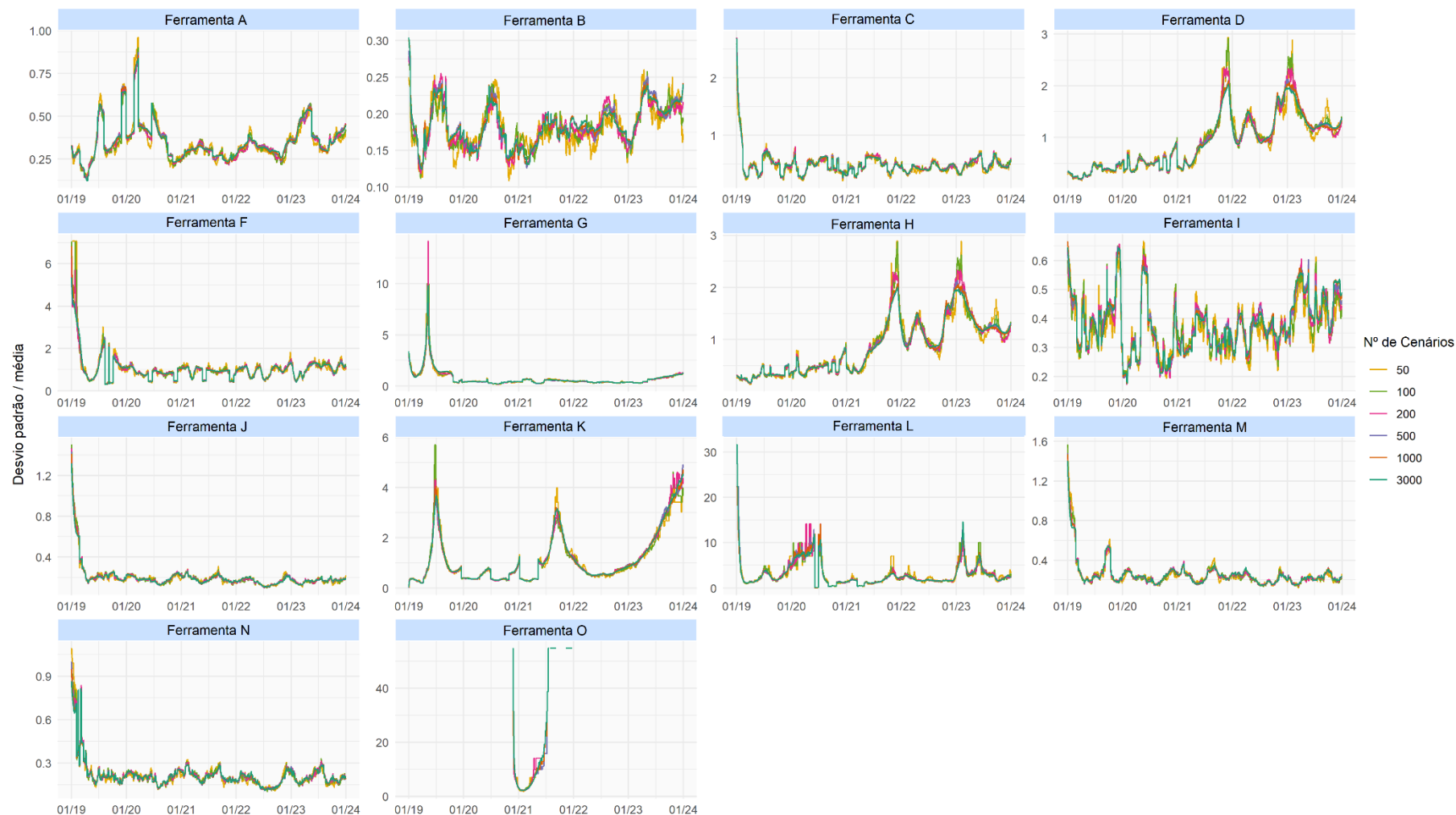


Figura 29 - Desvio padrão médio diário dividido pela média diária entre os cronogramas por ferramenta, por quantidade de cronogramas



Figura 30 - Distância absoluta entre as demandas diárias médias de todas as quantidades de cronogramas candidatas e a demanda diária média de 3000 cronogramas



Figura 31 - Distância relativa entre as demandas diárias médias de todas as quantidades de cronogramas candidatas e a demanda diária média de 3000 cronogramas

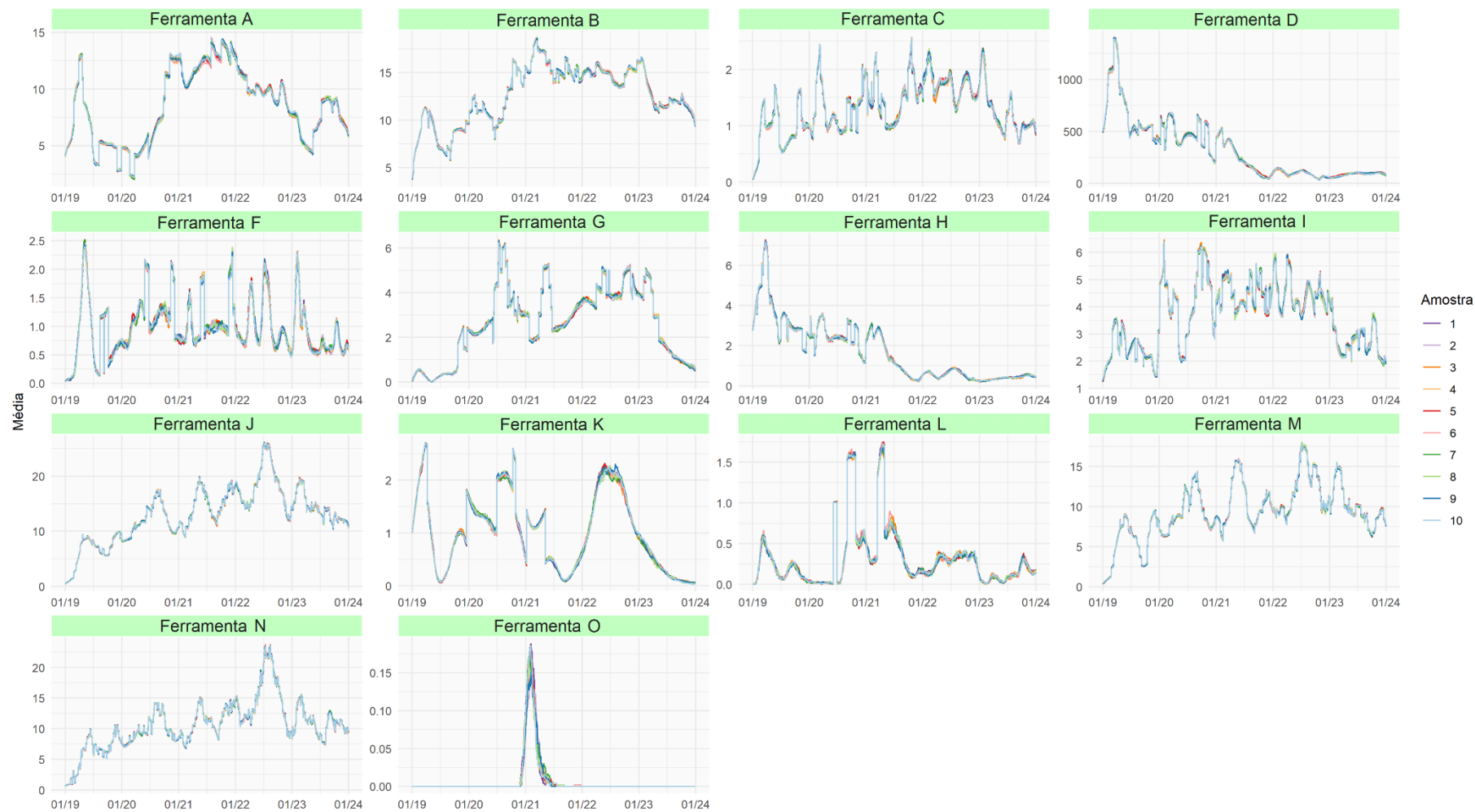


Figura 32 – Média diária das 10 demandas gerados com amostras aleatórias de 500 cronogramas

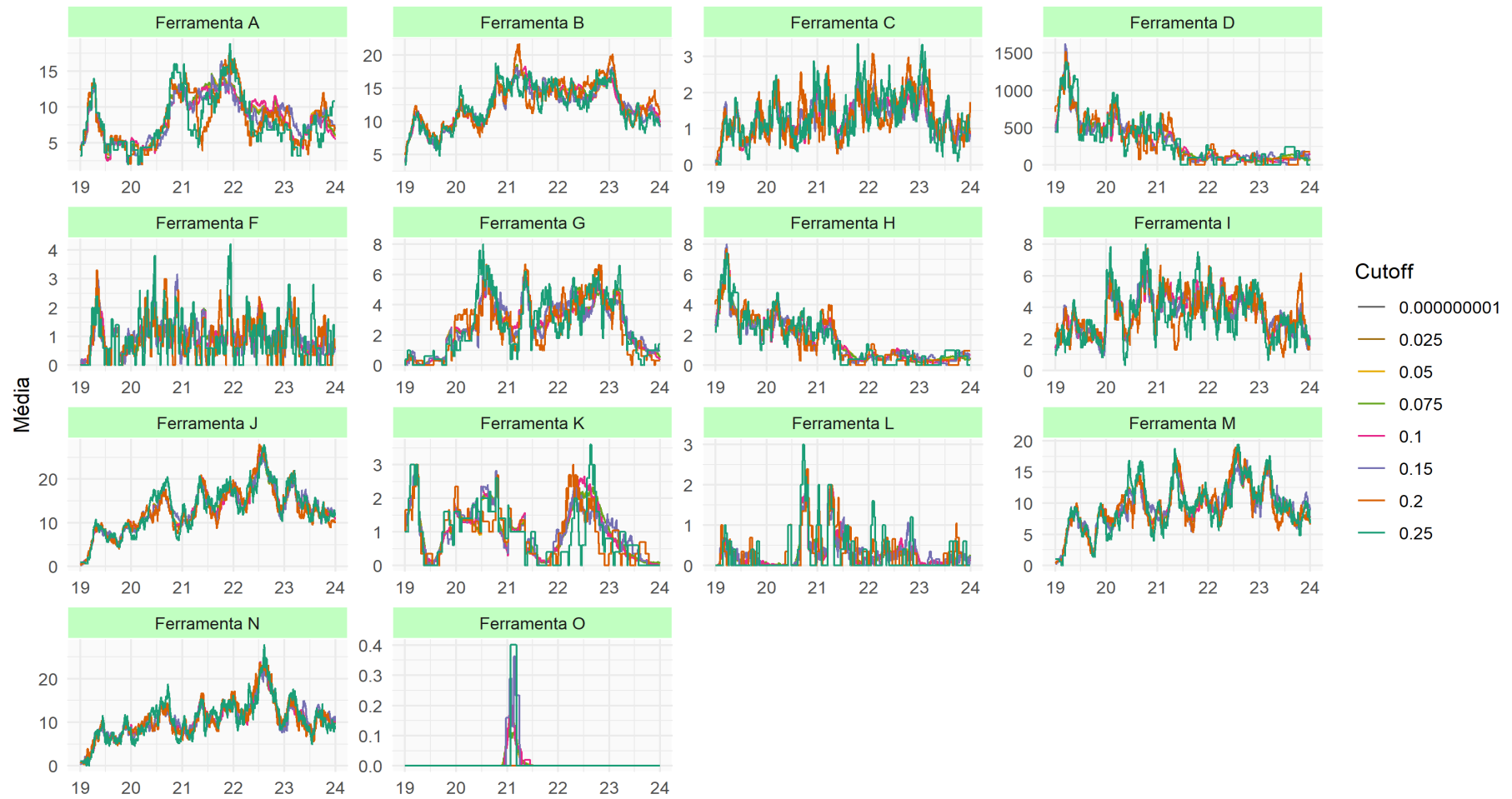


Figura 33 - Gráficos das demandas médias diárias calculadas com os cenários selecionados por cada *cutoff* no método exato, por ferramenta.

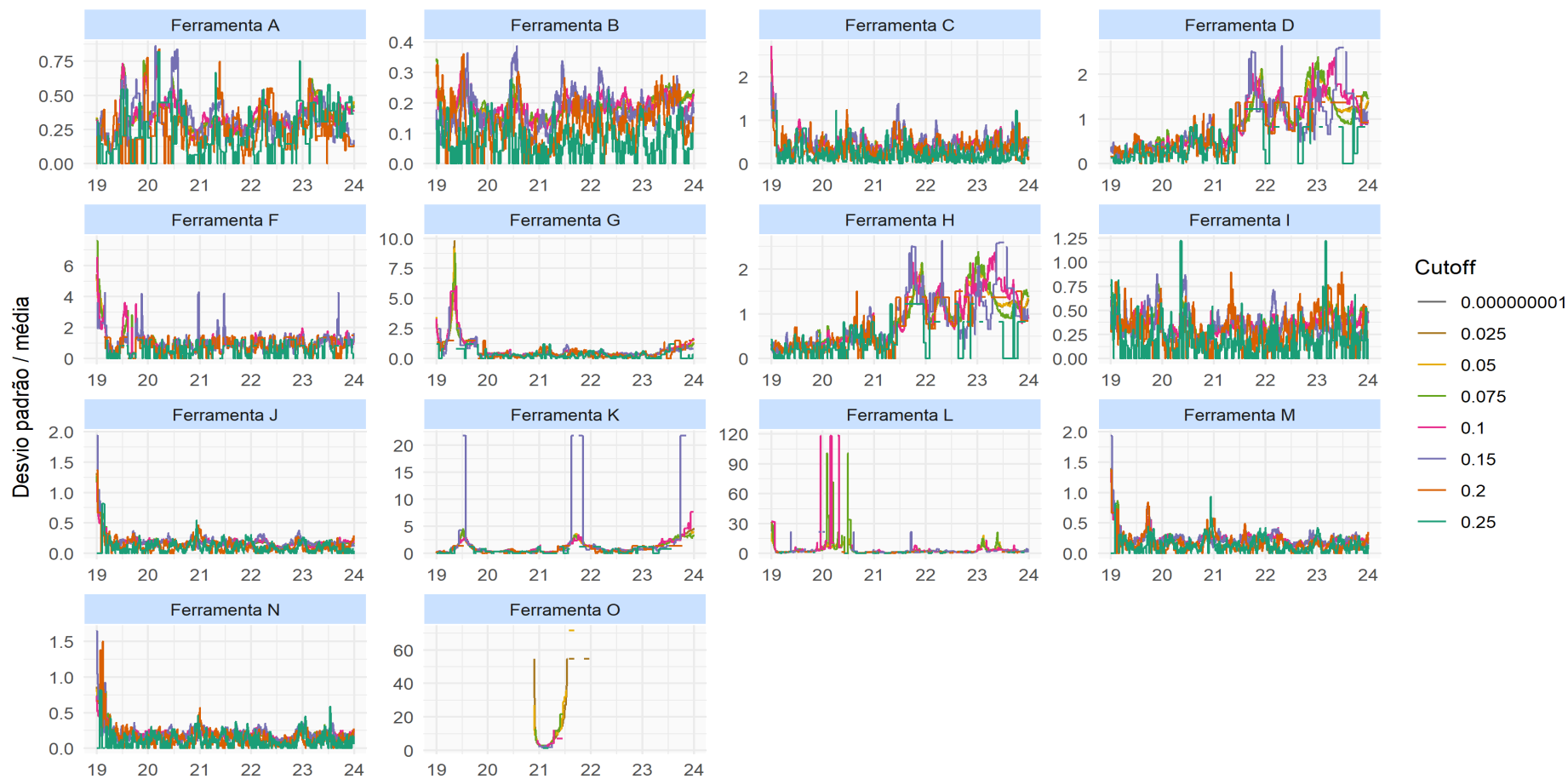


Figura 34 - Gráficos dos desvios padrões divididos pelas médias diárias das demandas calculadas com os cenários selecionados por cada *cutoff* no método exato, por ferramenta.

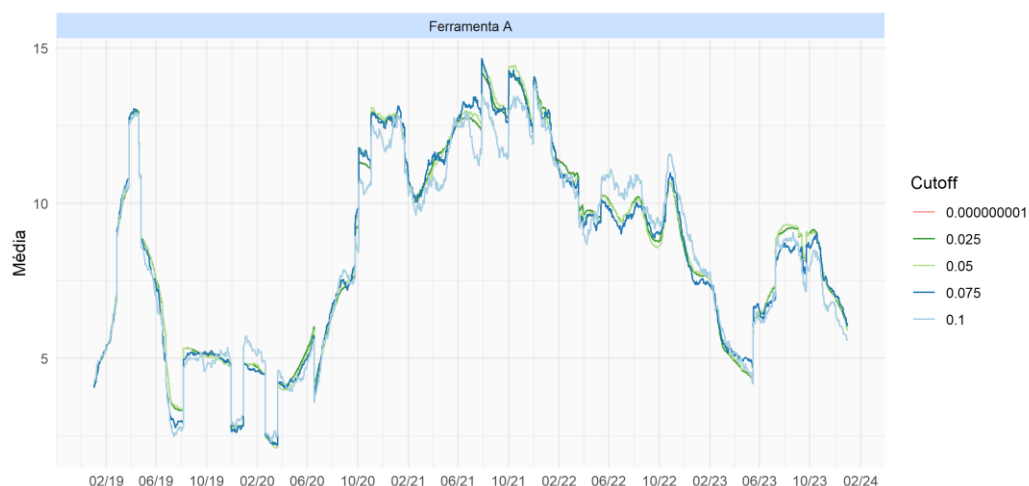


Figura 35 - Demandas médias diárias calculadas com os cenários selecionados pelos *cutoffs* de valor até 0.1 no método exato, para a ferramenta A

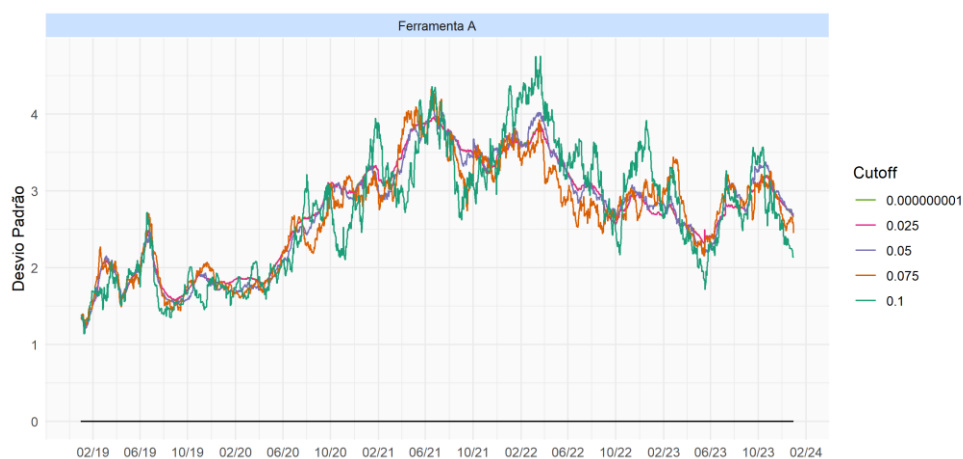


Figura 36 - Desvio padrão médio das demandas calculadas com os cenários selecionados pelos *cutoffs* de valor até 0.1 no método exato, para a ferramenta A

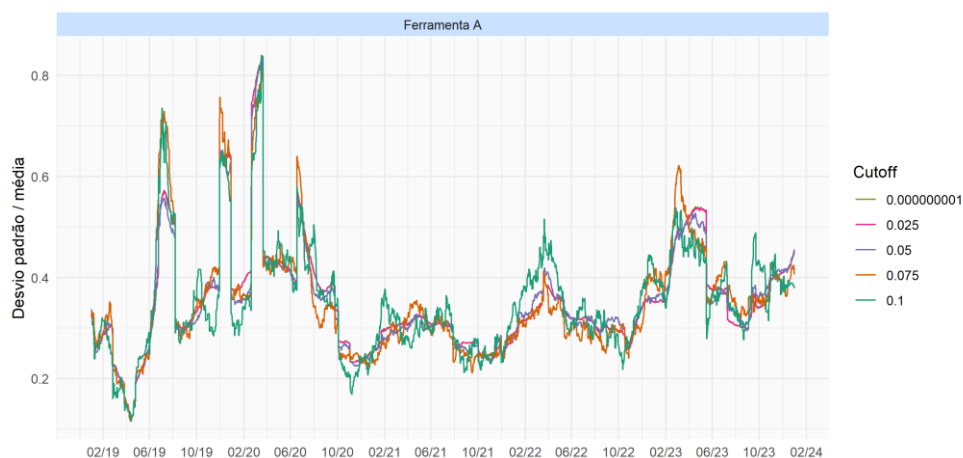


Figura 37 - Desvio padrão dividido pela média diário das demandas calculadas com os cenários selecionados pelos *cutoffs* de valor até 0.1 no método exato, para ferramenta A

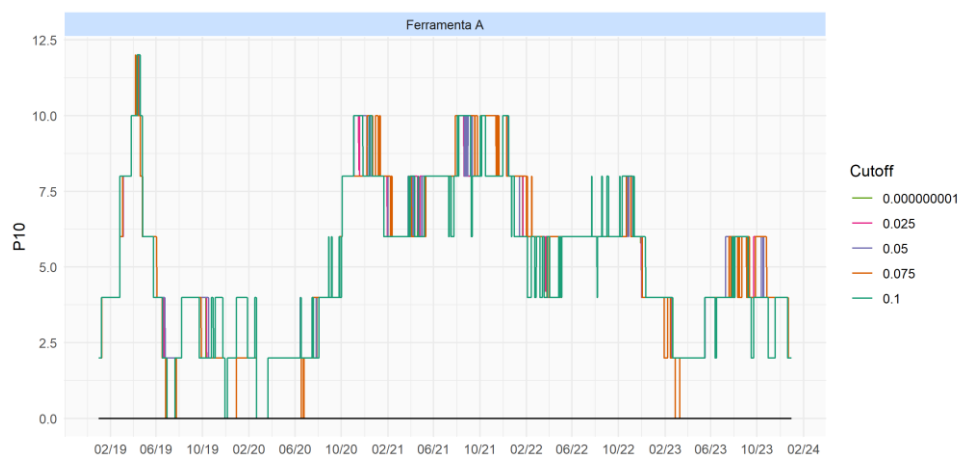


Figura 38 - Percentil 10 das demandas diárias calculadas com os cenários selecionados pelos *cutoffs* de valor até 0.1 no método exato, para a ferramenta A

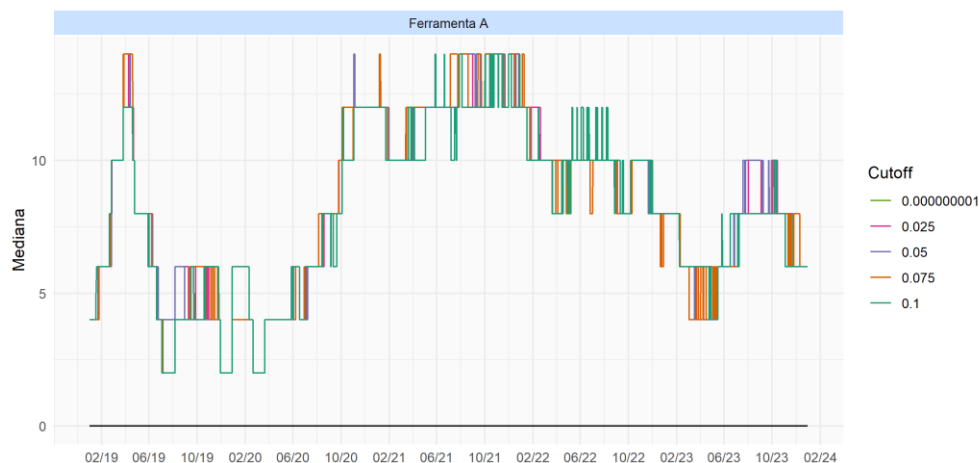


Figura 39 - Mediana (percentil 50) das demandas diárias calculadas com os cenários selecionados pelos *cutoffs* de valor até 0.1 no método exato, para a ferramenta A

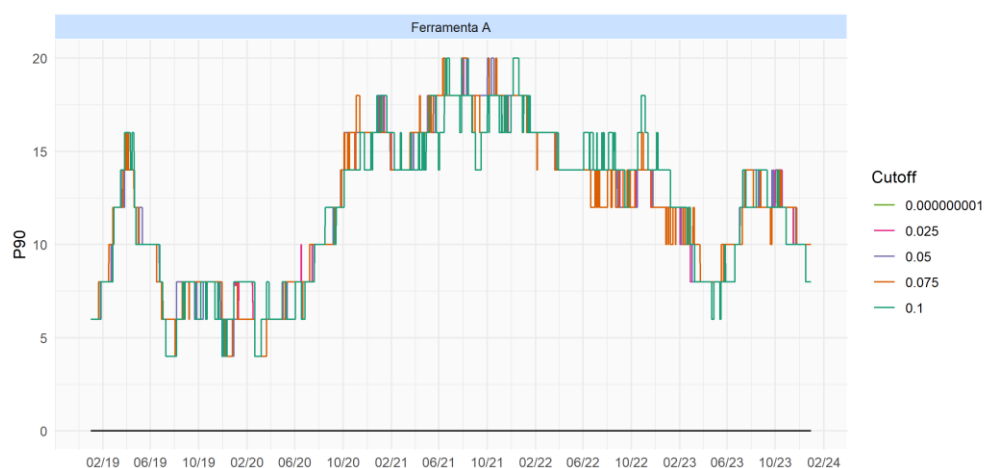


Figura 40 - Percentil 90 das demandas diárias calculadas com os cenários selecionados pelos *cutoffs* de valor até 0.1 no método exato, para a ferramenta A

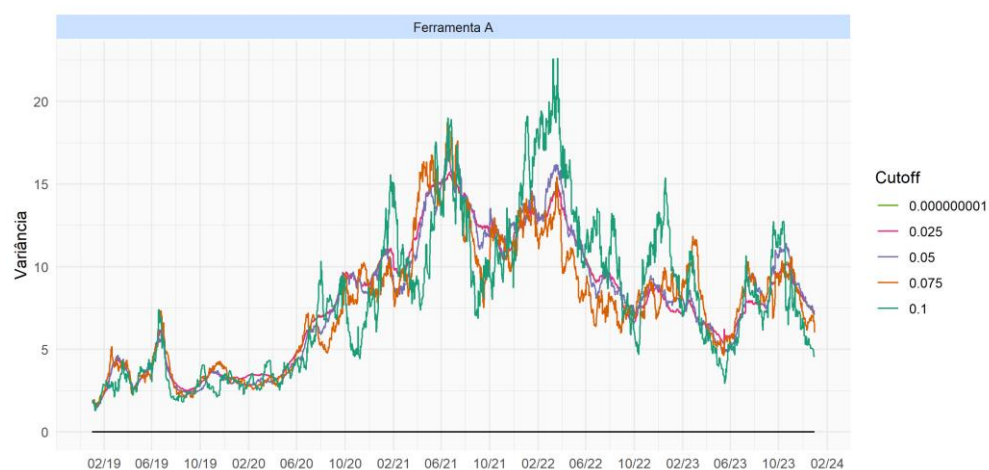


Figura 41 - Variâncias das demandas diárias calculadas com os cenários selecionados pelos *cutoffs* de valor até 0.1, para a ferramenta A

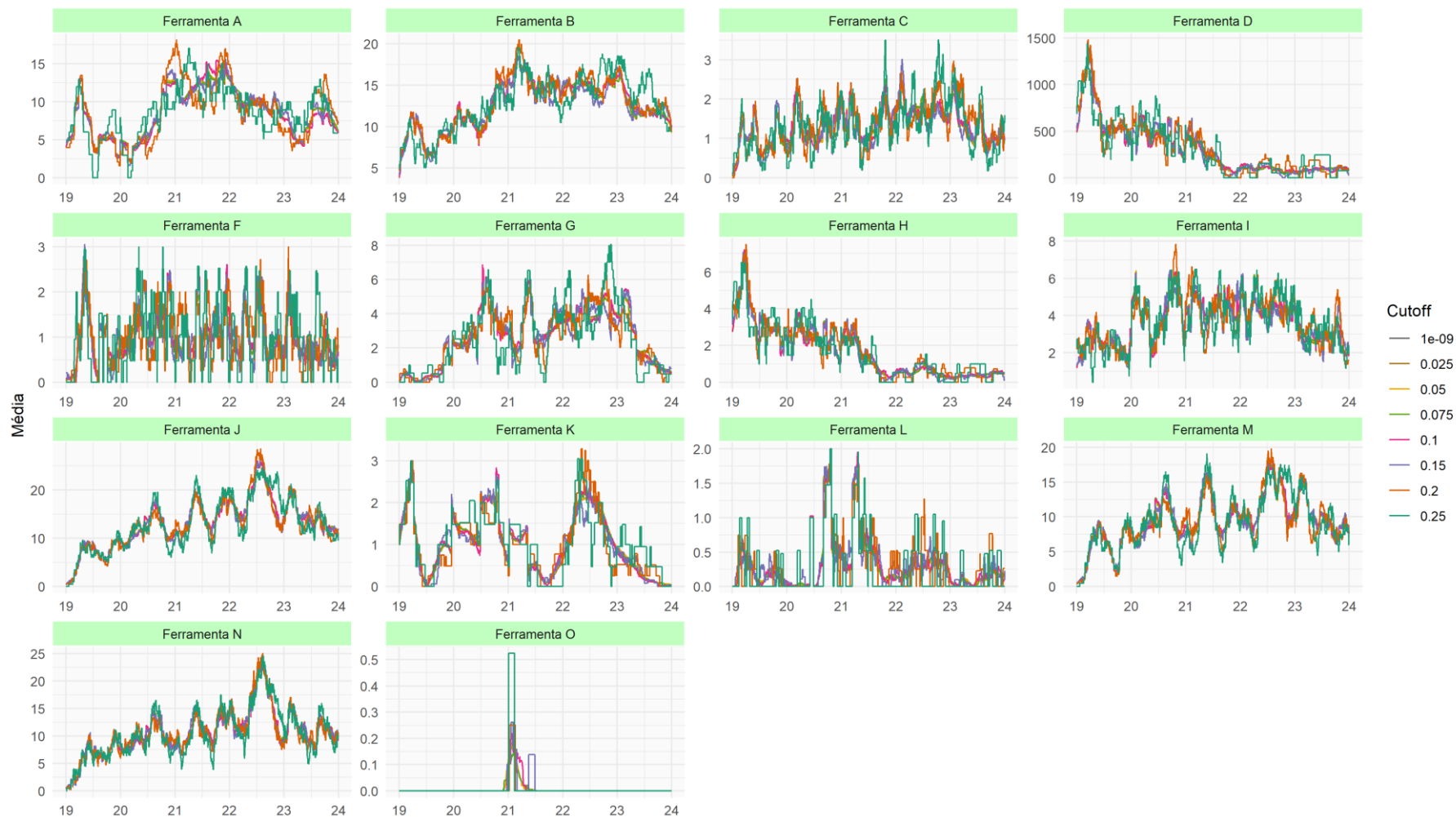


Figura 42 - Gráficos das demandas médias diárias calculadas com os cenários selecionados por cada *cutoff* no método heurístico, por ferramenta

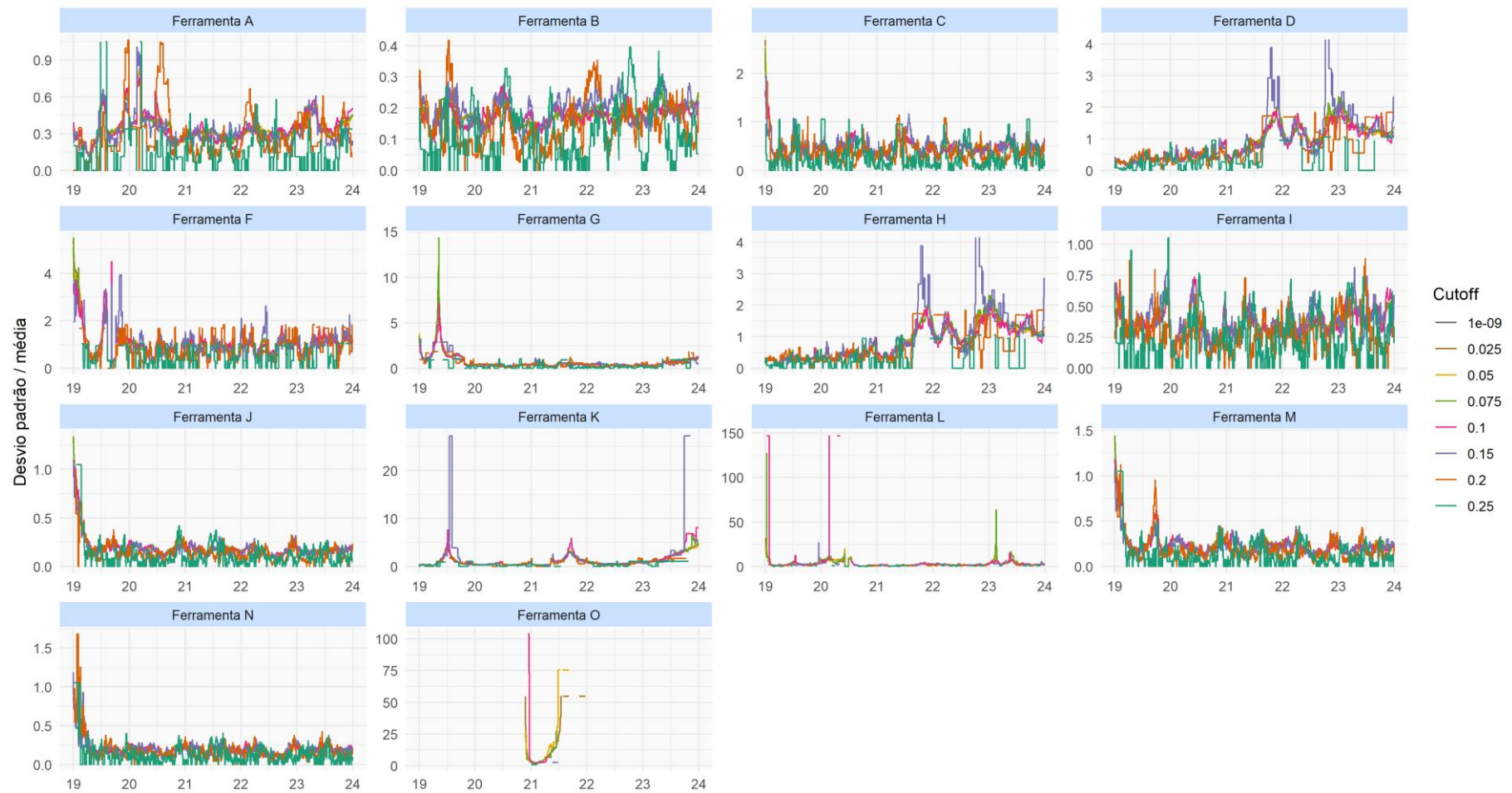


Figura 43 - Gráficos dos desvios padrões diários divididos pelas médias diárias das demandas calculadas com os cenários selecionados por cada *cutoff* no método heurístico, por ferramenta.

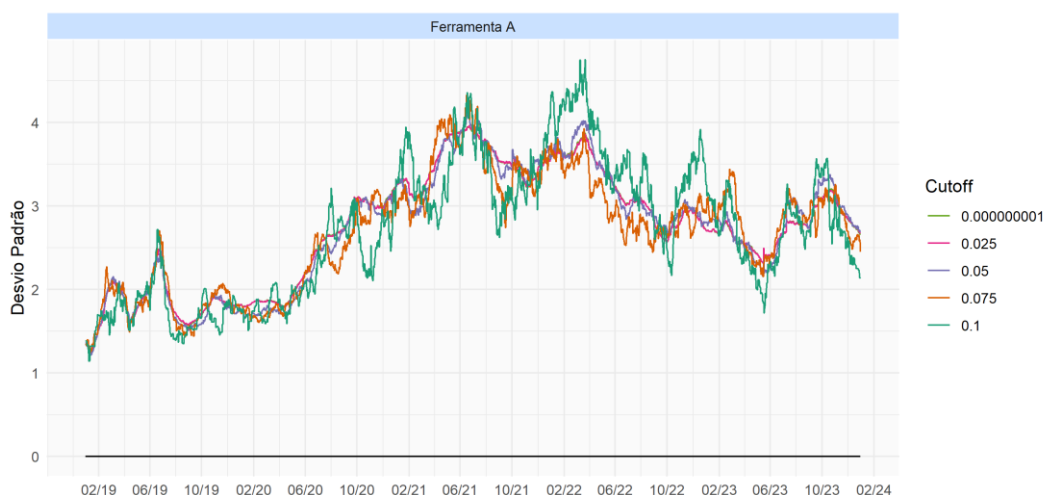


Figura 44 - Desvio padrão médio das demandas calculadas com os cenários selecionados pelos *cutoffs* até 0.1 no método heurístico, para a ferramenta A

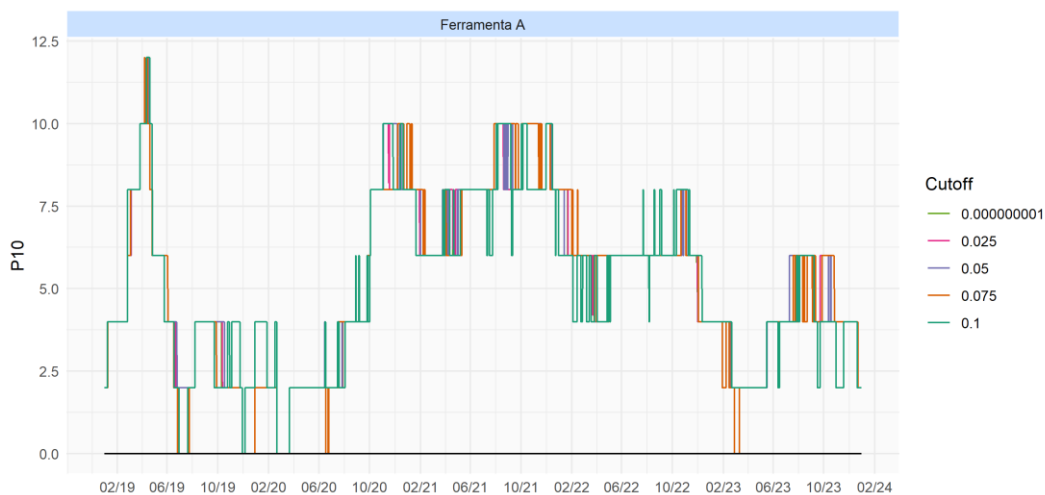


Figura 45 - Percentil 10 das demandas diárias calculadas com os cenários selecionados pelos *cutoffs* até 0.1 no método heurístico, para a ferramenta A

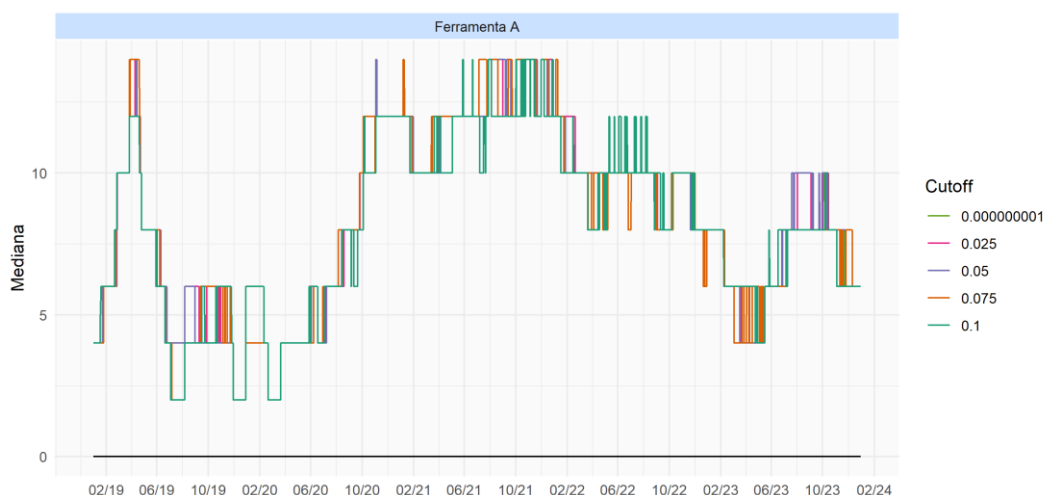


Figura 46 - Mediana (percentil 50) das demandas diárias calculadas com os cenários selecionados pelos *cutoffs* até 0.1 no método heurístico, para a ferramenta A

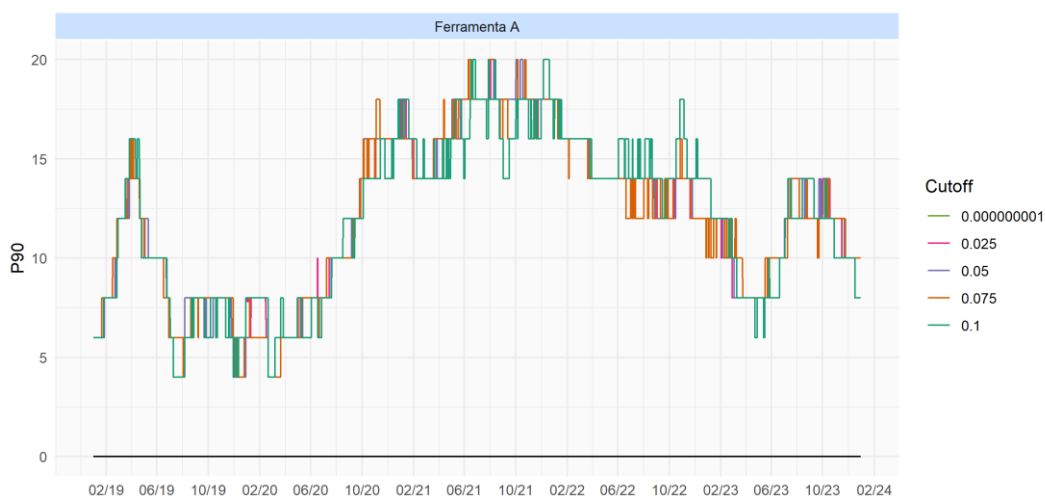


Figura 47 - Percentil 90 das demandas diárias calculadas com os cenários selecionados pelos *cutoffs* até 0.1 no método heurístico, para a ferramenta A

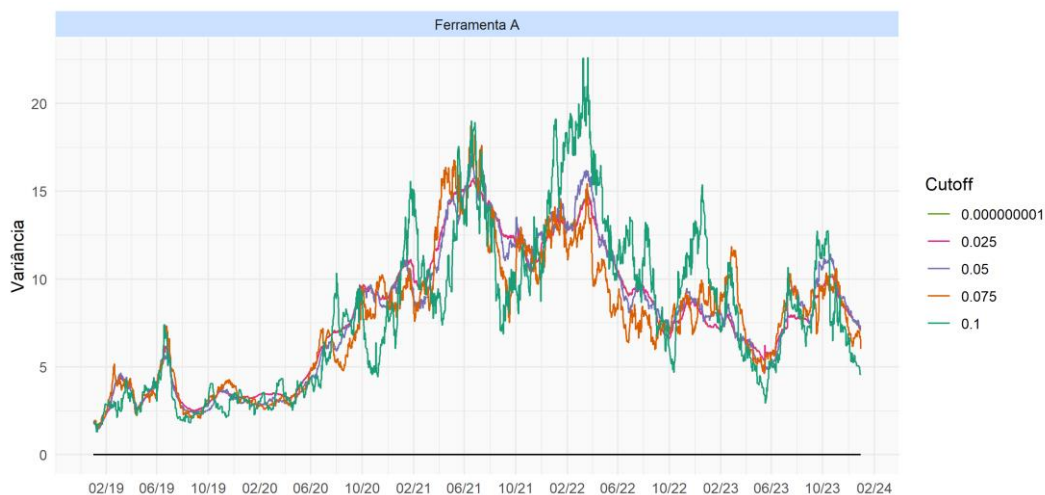


Figura 48 - Variâncias das demandas diárias calculadas com os cenários selecionados pelos *cutoffs* até 0.1 no método heurístico, para a ferramenta A

