Pontifícia Universidade Católica
do Rio de Janeiro

**Rachel Martins Ventriglia**

# Uncertainty and scenario reduction in material resources allocation of offshore rigs: a machine learning approach

**Dissertação de Mestrado**

Dissertation presented to the Programa de Pós-Graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia de Produção

Advisor: Prof. Leonardo dos Santos Lourenço Bastos
Co-advisor: Prof. Silvio Hamacher

Rio de Janeiro
February, 2024

**Rachel Martins Ventriglia**

# Uncertainty and scenario reduction in material resources allocation of offshore rigs: a machine learning approach

Dissertation presented to the Programa de Pós-Graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia de Produção. Approved by the Examination Committee:

**Prof. Leonardo dos Santos Lourenço Bastos**
Advisor
Departamento de Engenharia Industrial - PUC-Rio

**Prof. Silvio Hamacher**
Co-Advisor
Departamento de Engenharia Industrial - PUC-Rio

**Prof. Rafael Martinelli Pinto**
Departamento de Engenharia Industrial - PUC-Rio

**Prof. Paulo Cesar Ribas**
Molde University College

Rio de Janeiro, February 29th, 2024

**Rachel Martins Ventriglia**

Graduated in Production/Industrial Engineering at Pontifícia Universidade Católica do Rio de Janeiro in 2021, the same university where joined the master's degree program the following year (2022), also in Industrial Engineering, with emphasis in Operations Research. Works at Instituto Tecgraf/PUC-Rio, and has experience with Data Science, using machine learning and statistical tools for data analysis and process improvement and Project Management, using agile methodology tools.

Bibliographic data

Ventriglia, Rachel Martins

Uncertainty and scenario reduction in material resources allocation of offshore rigs : a machine learning approach / Rachel Martins Ventriglia ; advisor: Leonardo dos Santos Lourenço Bastos ; co-advisor: Silvio Hamacher. – 2024.
96 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2024.
Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Tarefas de construção de poços. 3. Sondas marítimas. 4. Redução de cenários. 5. Clusterização. I. Bastos, Leonardo dos Santos Lourenço. II. Hamacher, Silvio. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. IV. Título.

CDD: 658.5

To my parents, Ana Elisabeth and Roberto and my brother
Rafael, for their unconditional support.

## Acknowledgments

To my parents, Ana Elisabeth and Roberto, my brother Rafael, for the love and support always throughout my whole academic journey.

To my advisor, Leonardo dos Santos Lourenço Bastos, for all the guidance, advice, learning, brainstorms, and trust in me during the development of this project, always helping me when I needed it.

To my co-advisor, Silvio Hamacher, for the support, insights, and collaboration to carry out this work.

To my friends at Tecgraf, Luana Carrilho, João Gelli, Raphael Bittencourt, Davi Mecler and Gabriela Ribas, for the support, assisting in the development and understanding of this study and providing grateful insights.

To CNPq, PUC-Rio, and Tecgraf, for providing adequate resources, without which this work would not have been accomplished.

To all of those who contributed somehow for the fulfillment of this work.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

This study was also financed by the Brazilian Agency Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) under grant E-26/201.500/2023.

## Abstract

Ventriglia, Rachel Martins; Bastos, Leonardo dos Santos Lourenço (Advisor); Hamacher, Silvio (Co-Advisor). **Uncertainty and scenario reduction in material resources allocation of offshore rigs: a machine learning approach**. Rio de Janeiro, 2023. 96p. Dissertação de Mestrado - Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Material resource planning is an integral part of supply chain management. The tasks in the supply chain need materials and resources to be executed, thus, allocating resources correctly is an important part of task scheduling. Specifically, construction tasks for subsea wells require the use of resources, such as rigs, and planning the schedule of these operations involves the sizing of various materials and services necessary for their execution. This study is motivated by real-life scheduling planning from a large Oil and Gas company that estimates the demand for materials and services stochastically due to the uncertainties associated with the tasks in their start dates and durations. The calculation of the demand is subject to the current schedule that the company has and a set of rules that indicate allocation conditions, logistics parameters, disembarking conditions, and dependencies to allocate the tools and services needed for each task and estimate their quantity and how many days they will be used. These sets of tools and rules can change depending on the user and their operation knowledge. Additionally, the company uses a large number of scenarios, which results in extremely high computational times and impacts operational decision-making. In this context, scenario reduction could assist the company in its decision-making process. The methodology proposed in this work evaluates and identifies representative scenarios of uncertainty in strategic planning schedules of offshore rigs in order to reduce the number of scenarios used in the calculation of the demand for tools and services. With the use of unsupervised techniques, such as k-means and hierarchical clustering, we identified a subset with the most representative scenarios for the scenario reduction. The Wasserstein Distance and graphical visualization were used to measure the representativeness of the selected scenarios and find the best subset. Moreover, the scenario reduction subset was also used to analyze the impact of the reduction in the demand calculation. The Agglomerative Clustering with Ward Linkage (hierarchical clustering) obtained the best clustering evaluation and representativeness metrics, resulting in a selected subset of 782 scenarios. To find a minimal representative set of scenarios, the best clustering method and the Wasserstein Distance were used, resulting in a number of 343 scenarios. This presents a reduction of 84% in the execution time of the demand calculation, with the highest error of 11% in the demand calculation.



## Keywords

Well Construction Tasks, Offshore Rigs, Scenario Reduction, Clustering.

# Resumo

Ventriglia, Rachel Martins; Bastos, Leonardo dos Santos Lourenço (Advisor); Hamacher, Silvio (Co-Advisor). **Análise de incertezas e redução de cenários em alocação de recursos de tarefas de sondas marítimas: uma abordagem de machine learning**. Rio de Janeiro, 2023. 96p. Dissertação de Mestrado - Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

O planejamento de recursos materiais é uma parte importante do gerenciamento da cadeia de suprimentos. As tarefas na cadeia de suprimentos precisam de materiais e recursos para serem executadas e, portanto, alocar os recursos corretamente é uma parte importante do planejamento de tarefas. Especificamente, as tarefas de construção de poços submarinos requerem a utilização de recursos, como sondas, e o planejamento do cronograma dessas operações envolve o dimensionamento de diversos materiais e serviços necessários para sua execução. Este estudo é motivado pelo planejamento de programação real de uma grande empresa de Óleo e Gás que estima estocasticamente a demanda por materiais e serviços devido às incertezas associadas às tarefas em suas datas de início e durações. O cálculo da demanda varia de acordo com o cronograma atual que a empresa possui e a um conjunto de regras que indicam condições de alocação, parâmetros logísticos, condições de desembarque e dependências para alocar as ferramentas e serviços necessários para cada tarefa e estimar sua quantidade e quantos dias em que serão usados. Este conjunto de ferramentas e regras pode mudar dependendo do usuário e de seu conhecimento operacional. Além disso, a empresa utiliza um grande número de cenários, o que resulta em tempos computacionais extremamente altos e impacta a tomada de decisões operacionais. Nesse contexto, a redução de cenários poderia auxiliar a empresa no seu processo de tomada de decisão. A metodologia proposta neste trabalho avalia e identifica cenários representativos de incerteza nos cronogramas de planejamento estratégico de sondas offshore, a fim de reduzir o número de cenários utilizados no cálculo da demanda por ferramentas e serviços. Com a utilização de técnicas não supervisionadas, como k-means e agrupamento hierárquico, foi identificado um subconjunto com os cenários mais representativos para a redução de cenários. A Distância de Wasserstein e a visualizações gráficas foram utilizadas para calcular a representatividade dos cenários selecionados e encontrar o melhor subconjunto. Além disso, o subconjunto de cenários proveniente da redução também foi utilizado para analisar o impacto da redução no cálculo da demanda. O Clustering Aglomerativo com Ward Linkage obteve os melhores resultados de clusterização e representatividade, resultando em um subconjunto de redução de 782 cenários. Para encontrar um conjunto mínimo representativo de cenários, foi utilizado o melhor método de agrupamento, junto com a Distância de Wasserstein, e por fim obtido um número de 343 cenários. Isto apresenta uma redução de 84% no tempo de execução do cálculo da demanda, com o erro maior de 11% no cálculo da demanda.

## Palavras-chave

Tarefas de Construção de Poços, Sondas marítimas, Redução de Cenários, Clusterização.

## Table of Contents

## List of Figures

## List of Tables

# 1
# Introduction

Oil Exploration and Production (E&P) is a fundamental part of the supply chain in the Oil and Gas industry. It involves complex technical operations that can take long periods to be completed and require significant investments (SUSLICK; SCHIOZER; RODRIGUEZ, 2009, DEVOLD, 2013). Due to the high costs and complexity of the operations, it is necessary that the planning is done correctly, considering the tools required to support decision-making involving project planning systems and resource scheduling (LI; MCMAHON, 2007).

One of the most critical phases of Oil E&P is the construction and maintenance of wells, which depend mainly on oil platforms. These platforms are typically expensive and scarce of resources, with daily rates ranging between US$50,000 and US$500,000, depending on the platform, market, and operational specifications (KAISER; SNYDER, 2013, OSMUNDSEN; ROLL; TVETERÅS, 2010). Companies hire platforms to perform important well operations, such as drilling, assessment, completion, and workover, which are organized based on a schedule.

These tasks require materials and services to be executed, which must be allocated in the best possible way to carry out the operation. Therefore, planning rigs' schedules and operations involves robust resource planning for both materials and services of tasks. Moreover, specialized contracts of materials and services for well construction tasks can go from US$2M to US$1B.

To support the planning of these contracts, it’s important to correctly estimate the demand for materials and services, considering the uncertainty of the operation, which can involve the task durations, availability of materials, task start dates, and such. Tasks are subject to two main uncertainties: the start dates, which can begin earlier or later than initially planned, and their duration, which can take longer or shorter than what was originally scheduled.

This work is motivated by a problem of sizing the contracts of tools and services for well construction tasks performed by offshore rigs of a large Oil and

Gas company in Brazil. Currently, these tasks are subject to uncertainties, which can impact their start dates and duration, affecting their planning. To deal with the risks and uncertainties of the rig schedule, the company calculates the demand for tools and services stochastically, considering different scenarios of the rig schedule. From the risk assessment made by the company, considering uncertainties in the execution of operations (in this case, tasks), which can affect their durations and start dates, many scenarios are generated, which results in long computational times for the adequate calculation of tools to guarantee the desired service level. In addition, the company has little knowledge about the differences between the considered scenarios and how they impact the estimated demand for each tool in the operation.

In this context, scenario reduction methods could assist the company in its decision-making process and in better understanding the instances used. They are used to approximate the characteristics of the original scenario set by using a subset. Applications such as the dispatch of electric vehicles and online reconfiguration of networks involve a large number of random variables and can be favored by scenario reduction (LI et al., 2022a). In the context of oil and gas, Meira et al. (2016) used an optimization-based method to find the most representative models in oil fields.

Despite the extensive literature on resource allocation problems and scenario reduction methods, little is known about applications of this nature in the context of oil and gas, where resources have high added value, more specifically with the allocation of materials and services to tasks construction of marine wells.

## 1.1 Objectives

This study aims to reduce the number of scenarios used to calculate the demand for materials and services and to find the most representative ones using clustering-based methods and statistical analysis.

As complementary objectives, the list the following:

i. Survey the literature on the main methods for reducing scenarios.
ii. Analyze the generated scenarios and their respective decisions in terms of schedule and demand.

iii. Develop a methodology based on machine learning for reducing scenarios to be incorporated into the decision-making process.

# 2
# Theoretical Foundation

This section provides important concepts and works in the literature about the problem. First, we offer an overview of works that approach uncertainty in Oil Exploration and Production and Scenario Reduction methods. Then, works that use these methods in the oil E&P context.

## 2.1
## Uncertainty in Oil Exploration & Production

Due to the high costs and complexity of offshore rig operations in well construction tasks, it is necessary that the planning of these activities is done correctly, providing the necessary materials and services so that the tasks can be performed within the planned schedule. In addition to being sized, the materials must be available at the right time for the execution of the task. Thus, uncertainties in carrying out the task can result in high variability in the demand for tools, and the unavailability of the resource implies delays and operating costs for the rig responsible.

In the Oil E&P context, there are uncertainties to be considered in different aspects, such as geological concepts, the structure, the reservoir seal, or the hydrocarbon charge; the economic evaluations, in costs, oil price, technology, and probability of finding and producing economically viable reservoirs; as well as development and production, which take into consideration infrastructure, production schedule, operational costs, reservoir characteristics, and so on. In addition to the incorporation of uncertainties, there are also important decisions concerning the allocation of scarce resources and long horizons (SUSLICK; SCHIOZER; RODRIGUEZ, 2009).

According to Santos, Hamacher, and Oliveira (2021), uncertainty can be presented in optimization methods, such as simulation and optimization models, when uncertain parameters are simulated and then used in the optimization model or in the modeling approach, such as optimization under uncertainty. Bassi, Ferreira

Filho, and Bahiense (2012) used the first approach to minimize opportunity costs within certain operating constraints. They approach the problem of planning and scheduling a fleet of offshore oil rigs, considering the uncertainty in the service time.

Regarding resource planning and materials requirements for the rig scheduling problem, Marchesi et al. (2019) proposed a mixed-integer linear programming model for the construction of wells to minimize task tardiness and earliness, taking into consideration rigs and equipment. Drouven and Grossmann (2016) also propose a mixed-integer linear programming model, in this case for the shale gas development, maximizing the net presented value, defining which rigs, crews, and equipment will perform the drilling.

Resource planning is related to the rig scheduling problems that consider other resources when planning well operations. There has been a slight growth in studies considering resources such as offshore support vessels, lighter vessels, crews, and equipment (SANTOS; HAMACHER; OLIVEIRA, 2021).

## 2.2 Scenario Reduction Methods

Scenario analysis usually involves scenario generation and reduction. The first one creates scenarios from an original dataset, and the second one aims to reduce the scale of the scenario dataset, preserving its original characteristics (LI et al., 2022a). Li et al. (2022a) categorize the scenario reduction methods into four categories: distance-based methods, scenario tree-based methods, optimization model-based methods, and clustering-based methods. In this work, we focus on clustering-based methods. Table 1 shows the characteristics, advantages, and disadvantages of each method.

Table 1: Scenario Reduction methods. Source: (LI et al., 2022a)

| Category | Approach | Application | Advantage | Disadvantage |
|---|---|---|---|---|
| Scenario distance-based methods | Distances among scenarios | Measure the dissimilarity between scenario sets | Measurement of the dissimilarity is clear. Efficient to the scenario set whose scale is not too large. | Only quantify the mathematical properties and ignore the feature of variables in the real problem. The number of reserved scenarios should be determined in advance. |
| | Heuristic algorithms | The scenario set is measured by a distance. | | |
| | Mathematical programming algorithms | The scenario set is measured by a criterion. | | |
| Scenario tree-based methods | Scenario tree-based method | Spatiotemporally correlated scenario set | The relationship between scenarios is reserved after the reduction | Reduction of large-scale scenario sets is time-consuming |
| Optimization model-based methods | Single objective models | Only one criterion is considered in the reduction | More precise and optimal reduction strategies can be obtained | Computational complexity is hardly avoided when the scale of scenarios is large |
| | Multiple objective models | More than one criterion is considered in the reduction. | | |
| Clustering-based methods | Partitioning clustering | Select the "representative scenarios" from the scenario set | Efficiently reduce the scale of the scenario set | The quality of reserved scenarios is sensitive to clustering criteria and the number of clustering centers. |
| | Hierarchical clustering | | | |

As we can see in Table 1, each method is used for different applications. Scenario tree-based methods are used for spatiotemporally correlated scenario sets because they preserve the relationship between the scenarios. For this reason, Growe-Kuska, Heitsch, and Romisch (2003) used it in power management problems to reduce the number of nodes in individual scenarios by modifying the tree structure and bundling similar scenarios.

Optimization model-based methods involve single or multiple objective models and result in an optimal reduction. However, they can be computationally complex when the scale of the scenarios is large. Gil, Aravena, and Cardenas (2015) propose a stochastic mixed-integer programming formulation for deciding future generation investments considering uncertainty on the hydrological resource and use an optimization model to reduce the yearly hydropower output scenarios to make the optimization problem tractable.

Distance-based methods involve distance, heuristic, and mathematical programming algorithms and are applied to measure the distance between scenario sets, being efficient to scenarios whose scale is not too large.

Finally, clustering-based methods are used to select the representative scenarios of the set, being able to efficiently reduce the scale of the scenario set. Sumaili et al. (2011) applied clustering techniques for wind power scenario reduction, specifically the mean shift clustering, and were able to find a reduced set of representative scenarios associated with their probability of occurrence.

According to Li et al. (2022a), even though, for clustering-based methods, the quality of selected scenarios is sensitive to clustering criteria and the number of clustering centers, our focus is selecting the most representative scenarios, which is the subset that preserves the statistical characteristics of the original dataset, and for this reason, we chose this method for this study.

### 2.2.1 Clustering-based methods

Clustering techniques are unsupervised machine learning algorithms that group data according to intrinsic characteristics. In scenario reduction analysis, grouping is used to obtain representative scenarios from the data set. Commonly used clustering algorithms include partitioning clustering and hierarchical clustering.

Partitional Clustering aims to split the data and group the centroids of each group as the representative scenarios, and this approach includes methods such as K-Means (MACQUEEN, 1967) and K-Medoids (KAUFMAN; ROUSSEEUW, 2005).

The K-Means is a method to partition a dataset into a specific number of K clusters that are distinct and non-overlapping. The algorithm assigns each observation of the data set to exactly one of the clusters, considering that the within-cluster variation - how much observations within a cluster differ from each other - should be as small as possible. This is measured by a distance metric, most commonly the Euclidean distance (JAMES et al., 2023).

The optimization problem that defines the K-Means clustering is to minimize the within-cluster variation. When this metric is the Euclidean distance, it can be

written as Equation (**1**, where $C_1, \ldots, C_K$ denote the sets containing the indices of the observations in each cluster and $|C_k|$ is the number of observations in the k-th cluster and $\left(x_{ij} - x_{i'j}\right)^2$ is the squared Euclidean distances between the observations in the the k-th cluster (JAMES et al., 2023).

$$\min_{C_1,\ldots,C_k} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 \right\} \quad (1)$$

In other words, this means that the K-Means aims to partition the observations into K clusters in a way that the total within-cluster variation, summed over all the clusters, is as small as possible. James et al. (2023) describe the algorithm to solve this problem as the following:

1. Randomly assign a number from 1 to K to each of the observations as an initial solution.
2. Iterate until the cluster assignments stop changing.

In step number 2, the centroid is computed for each of the K clusters, given a vector of the feature means for the observations in the cluster. Then, each observation is assigned to the cluster whose centroid is the closest, according to the distance metric (JAMES et al., 2023).

One of the limitations of this method is the initialization since the first step is a random assignment, and the algorithm finds a local optimum. Therefore, it is sensitive to the initial centroid locations. Moreover, it is also sensitive to the presence of outliers in the data since the mean used to calculate the centroids is not a robust statistic (WU et al., 2008).

Another partitional clustering method is the K-Medoid. It is similar to the K-Means, but the centroids ("medoids") belong to the data being clustered. The medoid is located in the center of the cluster and also at the smallest sum of the distance to the other points. It uses the partitioning around medoids algorithm (PAM), and it minimizes the sum of the dissimilarities between the object and their closest object in the cluster, also known as the absolute error function (SUREJA; CHAWDA; VASANT, 2022).

The authors summarized the K-Medoids algorithm as follows:

1. Find the initial representative K centroids of the data randomly.
2. Assign each data point to its closest medoid.
3. Update:
    a. Select a non-medoid object randomly.
    b. Swap medoid with a data point.
    c. Compute the total cost.
    d. Select the medoid with the lowest cost for the next step.
4. Stop if the termination criteria are satisfied or go back to step 2 and repeat.

Hierarchical clustering can be divided into agglomerative and divisive. Agglomerative clustering is the most common type of hierarchical clustering, and it results in a dendrogram with the grouping of patterns and similarity levels at which the clusters change. This type of graph is typically described as an upside-down tree and is built from the leaves up, combining clusters to the trunk. Each leaf of the dendrogram corresponds to an observation in the data, and higher up the tree, some leaves fuse into branches, which means that the observations are similar to each other. The height of the fusion indicates the similarity and determines the number of clusters in the data (JAMES et al., 2023).

Figure 1 shows an example of a dendrogram, where (A) shows the graph with all of the observations, and (B) and (C) shows the results of the clustering method based on a cut in the height. The higher cut in (B) results in 2 clusters, while the cut in (C) results in 3 clusters (JAMES et al., 2023).

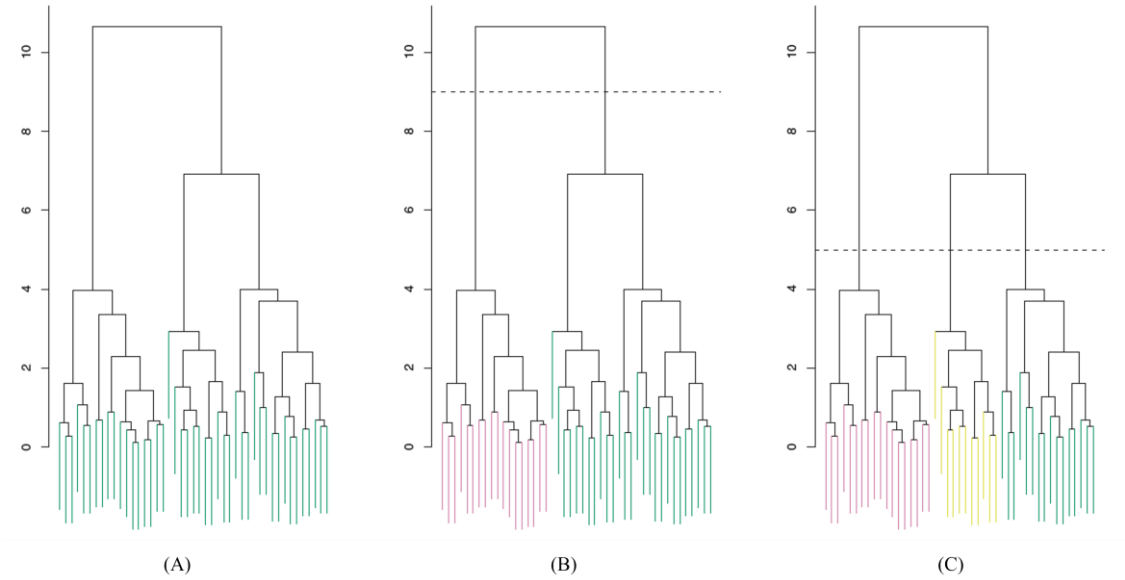



Figure 1: Example of dendrogram with (A) all its observations, (B) the clustering results with a higher cut in similarity, and (C) the clustering results with a lower cut in similarity. Adapted from: (JAMES et al., 2023)

The agglomerative clustering algorithm begins computing the similarity matrix with the distance between each pair of clusters and treats each pattern as a cluster. The Euclidean distance could be used as a distance measure for this. Then, it finds the most similar pair of clusters using the similarity matrix, merges them into one, and updates the matrix afterward. It stops after all the patterns are in one cluster. The divisive approach, on the other hand, begins with a single cluster of all the objects and splits the clusters at each step (JAIN; MURTY; FLYNN, 1999).

Hierarchical methods have a disadvantage in that they cannot repair what was done in previous steps, which means that the agglomerative method is not able to join two objects once they are separated, and the divisive method can't split two objects that were united. This results in shorter computational times but less flexibility for the algorithm since it is unable to correct wrong decisions (KAUFMAN; ROUSSEEUW, 2005).

For agglomerative clustering, there are different ways to define the dissimilarity between clusters, given by the linkage, which resulted in different algorithms. Common types of linkage are Single, Complete, Average, Centroid, and Ward Linkage. The Single linkage uses the minimal inter-cluster dissimilarity, computing the minimum distance between all pairs of patterns from the two clusters. The Complete linkage is the opposite; it uses the maximal inter-cluster

dissimilarity and computes the maximum distance. According to Jain, Murty, and Flynn (1999), these are the most popular algorithms for agglomerative clustering.

For the Average linkage, the algorithm uses the mean inter-cluster dissimilarity, computing the mean distance between the observations in a cluster. Centroid linkage uses the dissimilarity between the centroid of the clusters, but it has a disadvantage: an inversion can occur with this linkage, where two clusters are fused below the individual clusters of the dendrogram, leading to difficulties in visualization and interpretation of the dendrogram (JAMES et al., 2023).

Finally, the Ward Linkage, also known as the minimum-variance linkage, calculates the dissimilarity between two clusters based on the Euclidean distance between their centroids multiplied by a factor (KAUFMAN; ROUSSEEUW, 2005).

#### 2.2.1.1 Clustering evaluation metrics

Cluster validation is an important part of this analysis, and metrics are used to evaluate the performance of the cluster methods described in the last section. The cluster validation techniques can be classified as internal and external validation. The external focus is on validating a partition by comparing it with the correct partition, and the internal focus is on validating a partition by examining just the partitioned data (ARBELAITZ et al., 2013).

Since clustering is an unsupervised machine-learning approach, the correct partition is often not available for comparison. In this case, some metrics are most used, such as the Davies–Bouldin score (DAVIES; BOULDIN, 1979) or the Calinski–Harabasz score (CALINSKI; HARABASZ, 1974).

The Davies-Bouldin (DB) score is the average similarity between each cluster and its most similar one. It is used to estimate the cohesion based on the distance from the points in a cluster to its centroid and the separation based on the distance between centroids. Since it is preferable that clusters have the minimum possible similarity to each other, the goal is to minimize this score (HALKIDI, 2001).

This metric is defined as:

$$\boldsymbol{DB(C)} = \frac{\mathbf{1}}{\boldsymbol{K}} \sum_{\boldsymbol{c_k \epsilon C}} \max_{\boldsymbol{c_l \epsilon C \setminus c_k}} \left\{ \frac{\boldsymbol{S(c_k)} + \boldsymbol{S(c_l)}}{\boldsymbol{d_e(\overline{c_k}, \overline{c_l})}} \right\} \tag{2}$$

where $x_i$ is an object of a dataset X with N features ($X = \{x_1, x_2, \dots, x_N\}$), $c_k$ and $c_l$ are clusters from a partition in X into K groups ($C = \{c_1, c_2, \dots, c_K\}$, $\overline{c_k}$ and $\overline{c_l}$ are its centroids and $S(c_k) = 1/|c_k| \sum_{x_i \epsilon\, c_k} d_e(x_i, \overline{c_k})$, where $d_e(x_i, \overline{c_k})$ is the Euclidean Distance between objects $x_i$ and $\overline{c_k}$ (ARBELAITZ et al., 2013).

The Calinski-Harabasz (CH) score estimates the cohesion of the clustering based on the distances from the points in the cluster to its centroid. The separation is measured by the distance from the centroids to the global centroid (ARBELAITZ et al., 2013). It is defined as:

$$\boldsymbol{CH(C) = \frac{N-K}{K-1} \frac{\sum_{c_k \epsilon C} |c_k| d_e(\overline{c_k}, \overline{X})}{\sum_{c_k \in C} \sum_{x_i \epsilon c_k} d_e(x_i, \overline{c_k})}} \quad (3)$$

Where $\bar{X}$ is the centroid of dataset X and $d_e$ is the Euclidean Distance of the objects, following the same notation as Equation (2).

The Silhouette Coefficient, proposed by Rousseeuw (1987), also used as a validity metric, was a graphical display for partition techniques. Each cluster was represented by a silhouette based on the comparison of its tightness and separation, and it showed which objects were within their cluster and which were in between clusters.

It is a normalized summation-type index, where the clustering cohesion is measured based on the distance between all the points in the same cluster, and the separation is based on the nearest neighbor distance (ARBELAITZ et al., 2013). It is defined as:

$$\boldsymbol{Sil(C) = 1/N \sum_{c_k \in C} \sum_{x_i \epsilon c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{\max\{a(x_i, c_k), b(x_i, c_k)\}}} \quad (4)$$

where $a(x_i, c_k) = 1/|c_k| \sum_{x_j \in c_k} d_e(x_i, x_j)$ and $b(x_i, c_k) = \min_{c_l \in C \setminus c_k} \{1/|c_l| \sum_{x_j \in c_l} d_e(x_i, x_j)\}$, following the same notations as Equation (2).

#### 2.2.1.2 Scenario Representativeness metrics

Some clustering techniques use distance measures in their method to calculate the distances between clusters centroids. The K-Means, as mentioned before, uses the Euclidean Distance in the minimization of the within-cluster variation. Moreover, in the context of Scenario Reduction, distance measures can be used alongside clustering methods as an auxiliary tool, such as the Euclidean distance, as mentioned before. This section defines the scenario representativeness metric used in this work.

According to Panaretos and Zemel (2019), Wasserstein (W) distances are metrics between probability distributions that are inspired by the problem of optimal transportation. These distances are used in various problems, from fluid mechanics to optimization and statistics. The p-Wasserstein distance, also known as the Kantorovich or "earth moving" distance, between probability measures $\mu$ and $\nu$ is defined as:

$$\boldsymbol{W_p(\mu,\nu) = \inf_{\substack{X\sim\mu \\ Y\sim\nu}} (\mathbb{E}\|X-Y\|^p)^{\frac{1}{p}}, \qquad p \geq 1} \tag{5}$$

Besides the Wasserstein distance, bootstrap (EFRON, 1979) was used to obtain a confidence interval for the results obtained in this work. This resampling method is mostly used to provide a measure of the accuracy of a parameter estimate or of a given selection statistical learning method. The procedure involves randomly selecting *n* observations of a dataset to produce a bootstrap dataset, and the sampling is performed with replacement and repeated *B* times (where *B* is a very large value). This generates *B* different bootstrap (resampled) datasets with corresponding estimates (JAMES et al., 2023).

In this study, we used the percentile bootstrap to obtain the confidence interval of the mean absolute percentage error in the calculation of the demand for tools and services. According to Efron and Tibshirani (1986), defining $\theta$ as an unknown parameter, this method uses the parametric bootstrap cumulative distribution function (cdf) of $\hat{\theta}^*$, defined by Equation (6):

$$\widehat{\boldsymbol{G}}(\boldsymbol{s}) = \mathbf{Prob}_{*}\{\widehat{\boldsymbol{\theta}}^{*} < \boldsymbol{s}\} \quad (6)$$

Where $\mathrm{Prob}_{*}$ is the probability computed according to the bootstrap distribution of $\hat{\theta}^{*}$. The percentile method interval takes $\theta \in [\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)]$ as am approximate $1-2\alpha$ central interval for $\theta$. In other words, this method is the interval between the $100 \cdot \alpha$, and $100 \cdot (1\text{-}\alpha)$ percentiles of the bootstrap distribution of $\theta$. The percentile interval endpoints are described in Equation (7) (EFRON; TIBSHIRANI, 1986).

$$\boldsymbol{\theta_p}[\boldsymbol{\alpha}] \equiv \widehat{\boldsymbol{G}}^{-\mathbf{1}}(\boldsymbol{\alpha}) \quad (7)$$

### 2.2.2 Applications

Scenario reduction methods play an important role in problems that use a high number of scenarios, impacting the computational execution time and, consequently, the speed of decision-making. Chapaloglou et al. (2022), for example, consider the generation of uncertainty scenarios for energy storage sizing problems in isolated electrical systems. The study cites scenario reduction as an important step due to the constraints that make the problem computationally intractable and proposes a scenario generation methodology that selects minimal subsets of scenarios, using the Kantorovich distance to rank the scenarios and a K-Means to select them. The statistical quality of the scenarios is guaranteed by performing bootstrap to select the number of clusters and tests to monitor the statistical properties of the subsets. This is incorporated into the optimization problem, allowing them to explore an optimized combinatorial space of different uncertainty results.

Li et al. (2022b) use scenario reduction in the multiyear planning problem for the integration framework that combines distributed energy systems and electric vehicle charging in a neighborhood business center in Beijing. They propose a new data-driven method using real meteorological data to generate loading scenarios and use elbow-criterion clustering methods to perform the scenario reduction.

Abouelrous, Gabor, and Zhang (2022) propose a clustering-based reduction of scenarios applied to the inventory optimization problem for a retailer facing

online and in-store stochastic demand in a fixed-length sales season. The method proposed uses a non-pre-established number of clusters, giving greater flexibility than traditional clustering methods.

Hu and Li (2019) present an optimal scenario reduction method based on a new optimization framework to eliminate redundant initial scenarios using the concept of loss of correlation. Okada et al. (2019) used machine learning techniques such as multidimensional scaling and hierarchical clustering analysis to reduce the number of scenarios based on the Euclidean distance between simulation grids.

## 2.3 Scenario Reduction Methods in Uncertainty in Oil Exploration & Production

In the context of oil and gas, Mahjour et al. (2021) studied a reservoir development problem and argued that the scenario reduction technique with distance-based clustering with a simple correspondence coefficient can be used with other models, preserving representativeness. The authors applied unsupervised machine learning, considering different adjacency matrix constructions, dimensionality reductions, and clustering and sampling algorithms to generate several sets of representative geological realizations. The best algorithms for the UNISIM-I-D benchmark case under flooding were *Hausdorff*, *IsoMap*, and hierarchical clustering with *Ward Linkage* (MAHJOUR et al., 2022).

More specifically, in the context of oil exploration and production, Meira et al. (2016) proposed a methodology to identify representative scenarios in oil fields. For this, they used a mathematical function that modeled representativeness and an optimization tool to identify them, called *RMFinder*. Complementing this work, Meira et al. (2020) proposed an extension of the *RMFinder* technique to improve the reduction of the number of scenarios used in oil field decision-making. There are several uncertainties associated with this process, and therefore, many scenarios needed to be analyzed, which was time-consuming. Scenario reduction was modeled as a multicriteria optimization problem, and the number of representative models ranged from 1 to 25 in the performed experiments.

In the context of allocating resources for well-construction tasks performed by offshore rigs and calculating the stochastic demand for these tools and services,

Vieira (2021) proposed a methodology for selecting scenarios using the classic Set Covering model inspired by the forward selection method. It consisted of obtaining the main characteristics of the scenarios, calculating the distances between each other and the graph characterization, and pruning the graph to find a scenario subset. The results of the demand calculation of the scenarios selected by the Set Covering model in comparison to the original set show that there was only a 5% assertiveness loss using the scenario reduction subset, with a reduction of 91% in terms of data processed. The model was solved using an exact algorithm and a heuristic algorithm.

Even though there are studies that use machine learning for scenario reduction in the context of oil production and exploration, to the best of our knowledge, few of them focused on applying machine learning techniques in this context, specifically to the problem of allocating materials and services of well construction tasks performed by rig schedules.

# 3
# Methodology

This chapter presents the problem understanding, data preparation, and methods for this work based on the life cycle of data science. The methods and steps followed are presented in Figure 2.

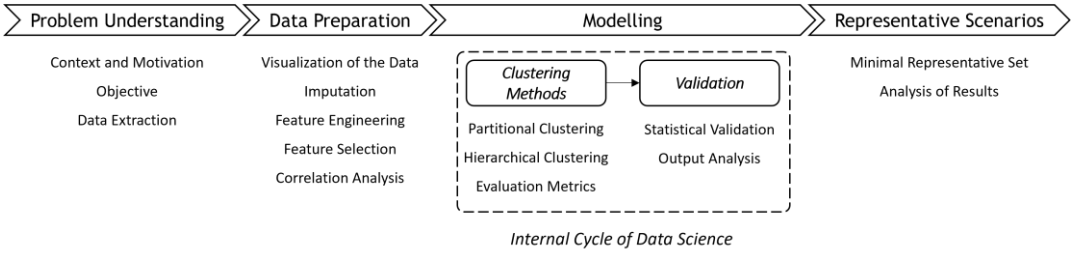



Figure 2: Methods for the Scenario Reduction Analysis

The process begins with problem understanding, which involves comprehension of the context, motivation, and objective of the analysis. This is done through meetings and interviews with managers and company operators. Then, the data is extracted from the company's database, followed by the visualization of the data. Imputation is performed if necessary. Feature engineering is used to create relevant features for our analysis, and then feature selection is applied. A correlation analysis was performed, which contributed to the selection of the features for the final database. In the data modeling step, we performed the scenario reduction and validation of the results. In the clustering methods step, we used partitional and hierarchical clustering methods to perform the scenario reduction and used evaluation metrics to analyze the results. Then, a validation was performed. The statistical validation used the Wasserstein Distance to measure the goodness of fit of each feature and then the visualization of the results to confirm whether the reduction found a representative subset of scenarios. Lastly, we performed an output analysis. The modeling part represents the internal cycle of data science and can be repeated if necessary. The last step consists of finding the

minimal representative subset of scenarios, using the best method from the step before, and analyzing the results. The description of each step is detailed below.

## 3.1 Problem Understanding

The problem addressed in this work is the sizing of tools and service contracts for marine well construction tasks performed by offshore rigs of a large oil and gas company. This company needs to execute the planning of the tasks carried out by offshore rigs, which includes the resource planning for both tools and services.

To support the planning, a decision-making support tool is used to estimate the demand for tools and services, considering the uncertainty of the operation. These tasks are subject to uncertainties, which can impact their start date or duration. This means that the tasks can begin at an earlier or later date and have a larger or smaller duration than what was originally planned.

Moreover, the calculation of the demand is subject to the current schedule that the company has and a set of rules that indicate allocation conditions, logistics parameters, disembarking conditions, and dependencies to allocate the tools and services needed for each task and estimate their quantity and how many days they will be used. These groups of tools and sets of rules can change depending on the user and their operation knowledge. Figure 3 shows the workflow of the demand calculation in the decision-making support tool.

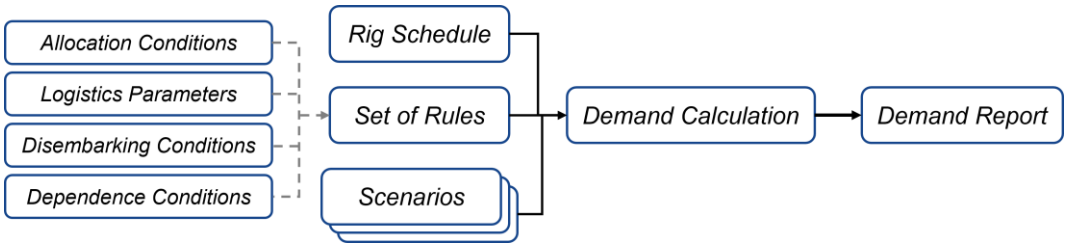



Figure 3: Workflow of the demand calculation

These sets of rules are important and mandatory for calculating the demand for tools and services, as they provide information necessary for this calculation, such as the number of tools needed in each task, embark and disembarking durations of tools, and dependency relationships between them.

The allocation conditions are the rules that indicate the quantity of the tool to be allocated per task and the probability of needing the tool for the execution of the task. This allocation is defined by the type of activity, but there can be exceptions where it can be defined by other factors, such as the location where the task will be executed or the type of rig. Figure 4 shows an example of how this rule translates to the calculation of the demand.

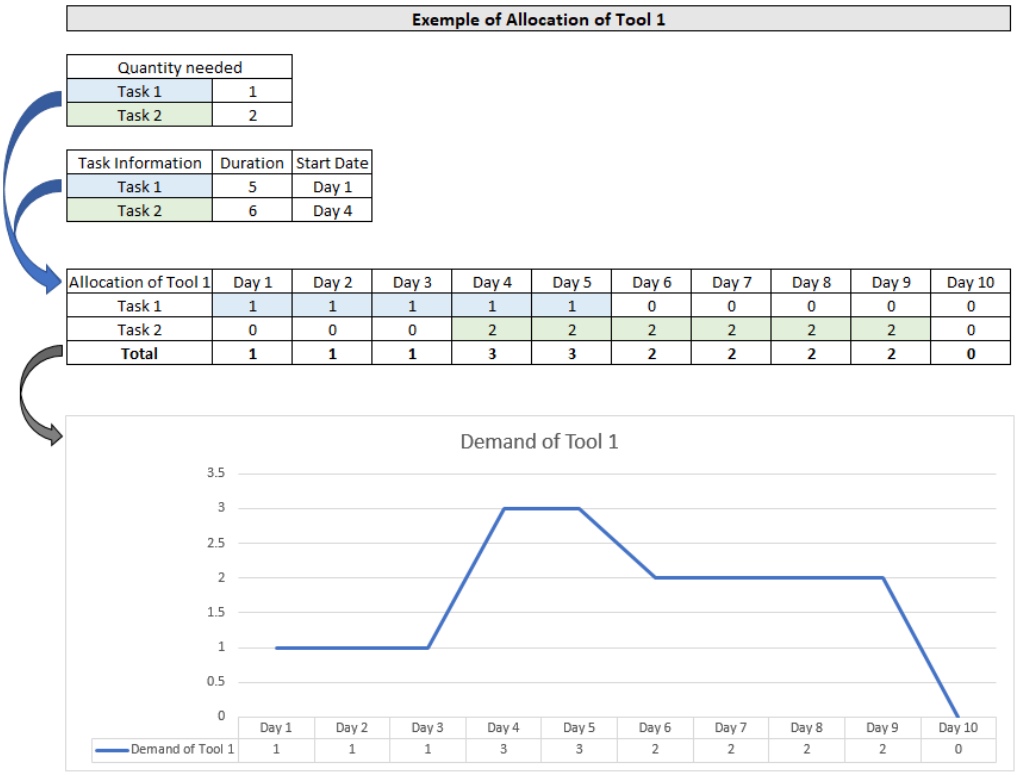



Figure 4: Example of allocation condition and demand calculation

The logistics parameters include the definition of the duration of embarking and disembarking the tools and the duration of technical services. These services include assembly, disassembly, and maintenance time, and also how the first two are defined. The assembly time can happen before or after loading, and the disassembly time can happen before or after unloading. The duration of embarking and disembarking the tools is defined by the oil basin.

The disembarking conditions define the criteria for staying at the location and maintenance, disembarking, and duration of use of the tool in an activity or installation and removal at the location. The first criterion defines whether the resource remains installed at the location after the rig is demobilized (for example, a temporary abandonment plug). If this is not applicable, it is necessary to inform the disembarking criteria.

Lastly, the dependence conditions are the set of rules that establish the relationship between tools. Sometimes, more than one tool is needed to execute a task, and these dependencies need to be informed. With the information provided by all the rules, it is possible to proceed to the demand calculation.

The demand calculation generates a report that shows the quantity needed for each tool and service for each day of the execution of the tasks in the rig schedule for each scenario. Based on these results, the demand for each scenario is aggregated, and statistical metrics are calculated, such as the percentiles 10% (P10), 50% (P50), and 90% (P90), and mean, minimum, and maximum for each tool and service and each day of rig schedule. This helps decision-makers plan the contracts for these tools and services. Figure 5 shows an example of a graphical demand report for a particular tool, which is one of the outputs of the decision-making support tool.

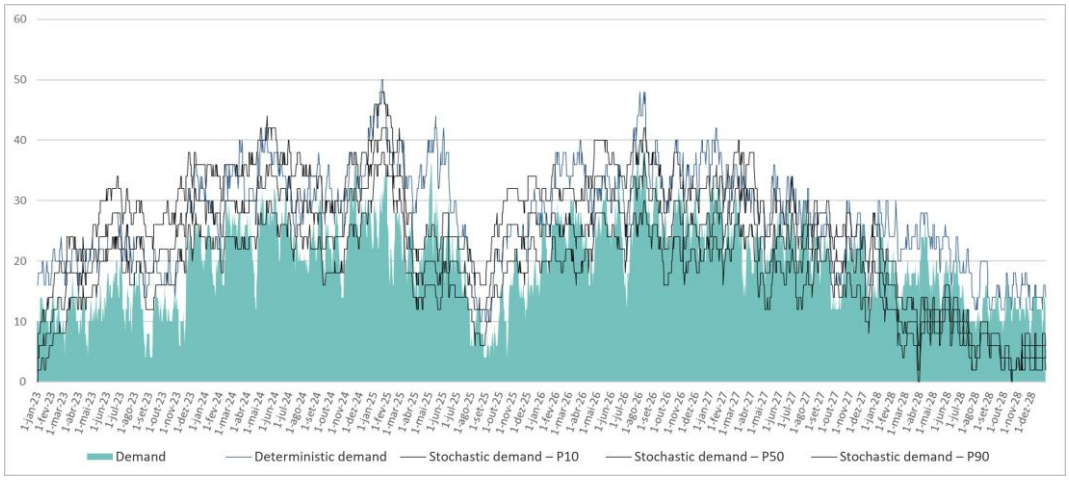



Figure 5: Graphical demand report example showing the calculation of the deterministic and stochastic demand of tools and services and its statistics, such as P10 and P90.

To deal with the uncertainty that the tasks are subject to, the demand calculation is done stochastically, considering different rig schedule scenarios with different start dates and durations of tasks. Currently, the company uses a considerably large number of scenarios, 2000 to be specific, to do this estimation, which increases the complexity of the problem and leads to high computational times (approximately 8 hours) for the adequate calculation of demand.

### 3.1.1
### Data Extraction

The data was obtained by the company's system, with data from 2023 to 2028. Two main databases were collected:

- The Deterministic Rig Schedule, with approximately 1800 tasks with start dates from 2023 to 2028, which represents the strategic planning period.
- Scenarios: originated from the deterministic rig schedule, where task uncertainties were applied, generating different start dates and durations for the tasks.

The Deterministic Rig Schedule contains the task identification number and their start date and duration. The Scenarios also contain this information for each of the 2000 scenarios.

## 3.2
## Data Preparation

After the data extraction, the data is prepared. This was an important step because the data extracted was not in the best format for the analysis proposed. The Deterministic Rig Schedule and the Scenarios data had information about the tasks, and this study needs information and statistics regarding the scenarios.

We begin this process with the visualization of the data, more specifically, the visualization of the uncertainties presented in the data (start dates and duration), to better understand it.

The Deterministic Rig Schedule was used as auxiliary data for the Scenarios. Firstly, information from the deterministic start dates and duration of the tasks was included in the Scenarios; then feature engineering was used to include two new features in the data: Difference in Start Date (between the deterministic schedule and the scenario) and Relative Change in Duration. These features are useful to better understand the uncertainty presented in the data regarding the start dates and durations. Table 2 shows the features used from each data and their description.

Table 2: Description of features

| Features | Origin | Description | |
|---|---|---|---|
| Task ID | Scenarios | Identification number of the task | |
| Scenario number | Scenarios | Identification number of the scenario | |
| Task Duration in Scenario | Scenarios | Duration of task in the scenario | |
| Task Start Date in Scenario | Scenarios | Start date of task in the scenario | |
| Deterministic Task Duration | Deterministic Rig Schedule | Deterministic Duration of task | |
| Deterministic Task Start Date | Deterministic Rig Schedule | Deterministic Start date of task | |
| Difference in Start Date | Feature Engineering | $\lvert startdate_{js} - startdate_{j} \rvert$ | (8) |
| Relative change in Duration | Feature Engineering | $\frac{p_{js}}{p_j}$ | (9) |

In Equation (8), $startdate_{js}$ is the start date of the j task in scenario s, and $startdate_j$ is the deterministic start date of task j. In Equation (9), $p_{js}$ is the duration of task j in scenario s and $p_j$ is the deterministic duration of task j.

Then, to obtain information about the scenarios, the data was grouped so that each row represented a scenario, and their statistics were calculated. The following statistics were obtained for each scenario:

- Mean and Standard Deviation (Std) of the Tasks Duration.
- Number of Tasks.
- Minimum, Mean, Standard Deviation, Maximum, and 90% Percentile (p90) of the Tasks Relative change in Duration.
- Mean Absolute Difference in Start Date (between the deterministic schedule and the scenario).
- Number of Tasks that were brought forward and pushed back.
- Number of Tasks that had duration decreased and increased.

Moreover, to understand the changes in the scenario from the original deterministic schedule, two other features were included: the number of tasks that were brought forward and pushed back and the number of tasks that had duration decreased or increased. The first two indicate how many tasks had changes in their start date, whether they started earlier or later than initially scheduled, and the last two indicate how many tasks had their duration changed. This indicates how much the scenario changed from the deterministic schedule.

In addition, a correlation analysis was performed using the Pearson Correlation to see if any of the features were highly correlated. A high correlation

was considered to be above 0.6 or below -0.6. With the results of the correlation analysis, redundant features were removed, and the final dataset was composed.

## 3.3 Modeling

In the modeling step, we performed the scenario reduction using clustering methods, validated the results using metrics to evaluate the distance between distributions, and analyzed the impact of the reduction on our problem. This cycle was repeated as many times as needed to identify the best features and methods.

### 3.3.1 Clustering Methods

Partitional and Hierarchical methods were applied to the database for scenario reduction: Two partitional methods, the K-Means and the K-Medoids, were selected. For the Hierarchical Methods, Agglomerative Clustering was used, varying their Linkage (Ward, Average, Single, and Complete).

For the evaluation of the results, three evaluation metrics were used: the Silhouette Coefficient (ROUSSEEUW, 1987), the Davies-Bouldin score (DAVIES; BOULDIN, 1979), and the Calinski-Harabasz score (CALINSKI; HARABASZ, 1974). Table 3 summarizes the description of each evaluation metric.

Table 3: Summarization of the evaluation metrics

| Metric | Description | Interpretation |
|---|---|---|
| Silhouette Coefficient | A normalized summation type of index. Measures the cohesion of the clusters based on the distance between all the points in the same cluster and the separation based on the nearest neighbor distance. (ARBELAITZ et al., 2013) | The higher the Silhouette Coefficient, the better the clustering. |
| Davies-Bouldin score | Calculates the cohesion of the clustering based on the distance from the points in a cluster to its centroid, and the separation is measured by the distance between the centroids. (ARBELAITZ et al., 2013) | The lower the metric, ideally near zero, the better the clustering is considered. |
| Calinski-Harabasz score | Estimates the cohesion of the clustering based on the distances from the points in the cluster to its centroid. The separation is measured by the distance from the centroids to the global centroid. (ARBELAITZ et al., 2013) | A clustering is considered good if this metric has a high value. |

For the scenario reduction, the number of clusters for each method varied from 2 to 1999. In the first and second data science cycles, the best number of clusters will be selected by the Silhouette Coefficient. The other metrics were calculated accordingly.

To obtain the subset of scenarios based on the results of each clustering method, the scenario with the lowest Euclidean distance from the centroid was selected to compose the subset of the representative scenarios. For the hierarchical clustering, which does not have the centroids as an attribute of the method, the centroids were calculated based on the nearest centroid.

### 3.3.2 Validation

#### 3.3.2.1 Statistical Validation

The Wasserstein Distance was used to evaluate the goodness of fit of every feature's distribution after the reduction in comparison to the original set of scenarios. The main use for this metric has been as a tool for statistical inference,

and it is a good measure for carrying out goodness-of-fit tests (PANARETOS; ZEMEL, 2019).

This metric has been used as a discrepancy measure for the hierarchical clustering method to improve time aggregation performance (CONDEIXA; OLIVEIRA; SIDDIQUI, 2020), as a tool to select representative scenarios for stochastic programming (HEITSCH; MISCH, 2003), among other applications.

The Wasserstein Distance was calculated for each feature, comparing it to the corresponding feature in the original set, and a Mean Wasserstein Distance was calculated for each clustering method. The lower the Mean Wasserstein Distance, the better the clustering method, as it means that the scenario reduction is statistically close to the original set.

Moreover, a visualization of the features of the reduced dataset was made to compare to the 2000 scenarios using a density plot. We compared the distribution of each feature to the original dataset to evaluate the results of the scenario reduction further.

#### 3.3.2.2 Output Analysis

To analyze the impact of the scenario reduction in our problem, this validation consists of applying the reduction to the calculation of the demands for tools and services and comparing it to the original results using the 2000 scenarios. The clustering method with the lowest Mean Wasserstein Distance, based on the statistical validation, was selected for this analysis. We chose two groups of tools and services with a particular set of rules to perform this analysis. The first group refers to tools and services that are needed in fluid operations. The second one uses a group of tools and services for tasks that acquire geological data from the marine wells.

Based on the demand calculation results from the scenario reduction and the original set, we calculated the daily demand average for each tool and then the absolute percentage error from the reduction to the original set results. Lastly, we calculated the Mean Absolute Percentage Error (MAPE) for the following statistics: P10, P50, P90, mean, minimum, and maximum. Bootstrap was applied to calculate the confidence interval of the MAPE, using 500 resamples.

## 3.4 Representative Scenarios

After defining the best clustering method for scenario reduction, our goal is to find the minimal representative set of scenarios for our problem. In other words, we want to obtain the minimal number of clusters where the distance between the set and the original scenarios is small enough to consider them a representative set for scenario reduction.

We used the features and clustering method from the data modeling step and varied the number of clusters from 2 to 1999 clusters. For each number of clusters, we applied the clustering method and obtained the scenario reduction set, as described in section 3.3.1. For each feature of the subset, we normalized the data based on the original dataset, using the Min-Max Normalization, and calculated the Wasserstein distance. After analyzing all the features, we calculated the Mean Wasserstein Distance.

As the number of clusters increases and approaches 2000, the Mean Wasserstein Distance becomes smaller, so the minimal representative subset would be selected when the change in the decreasing rate of the Wasserstein Distance is considerably small. The decreasing rate was calculated using the difference between the Mean Wasserstein Distance of $n$ number of clusters and $n-1$, and the change in the decreasing rate was calculated using the same difference in the decreasing rate.

When the absolute change in the decreasing rate of the Mean Wasserstein Distance was smaller than a certain value (α), this means that the change is small enough, and this number of clusters is selected as the minimal representative subset. We selected an α of $10^{-4}$ and the flowchart of this algorithm is described in Figure 6.

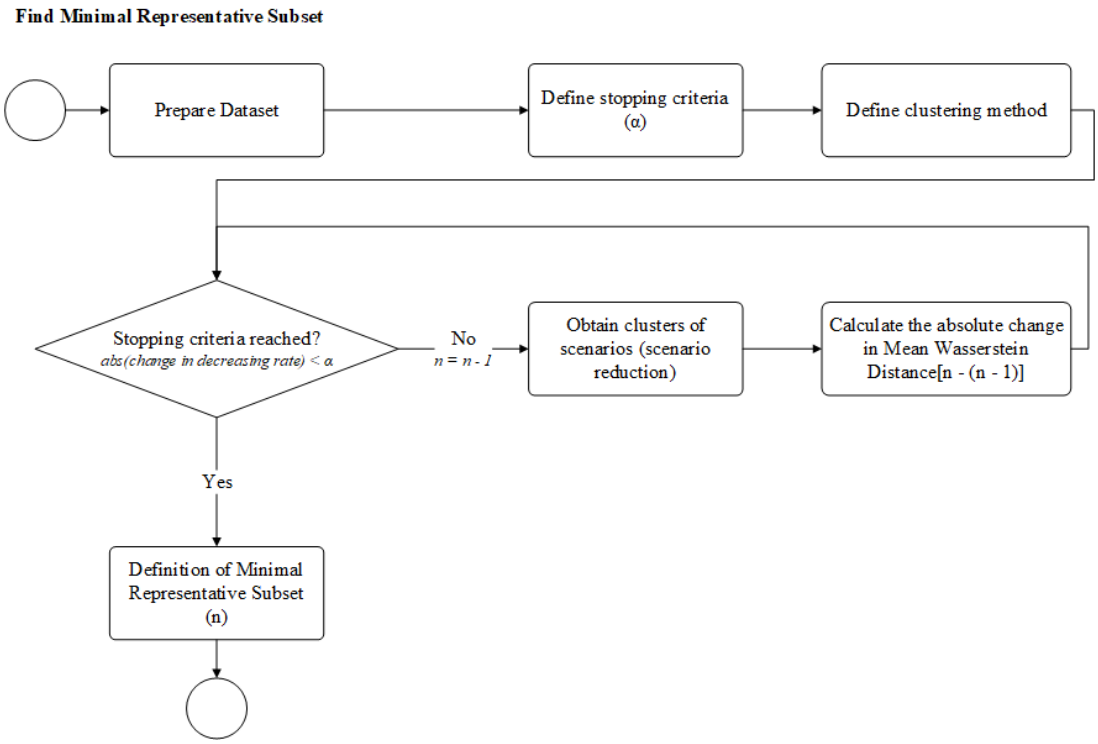



Figure 6: Find Minimal Representative Scenarios Algorithm

After selecting the minimal representative set of scenarios, the same validation process described in Section 3.3.2 was performed, and the results were analyzed as part of the output analysis.

The experiments detailed on this Chapter were performed on a computer with an Intel Core i7 3.6 GHz, 64 GB of RAM, Windows 10, and Python 3.11.3. Among the existing packages in Python, the main ones used in this study were *scikit-learn*, *scipy*, *statsmodels*, *pandas*, *matplotlib*, and *seaborn*.

# 4
# Results

This chapter presents the results of the scenario reduction. First, we performed the steps described in the data preparation. Then, we applied clustering methods and evaluated the results. Lastly, the statistical validation was performed.

## 4.1
## Data Preparation

### 4.1.1
### Visualization of the data

We visualized the uncertainty of the start dates and duration of tasks in Figure 7 and Figure 8, to better understand them.

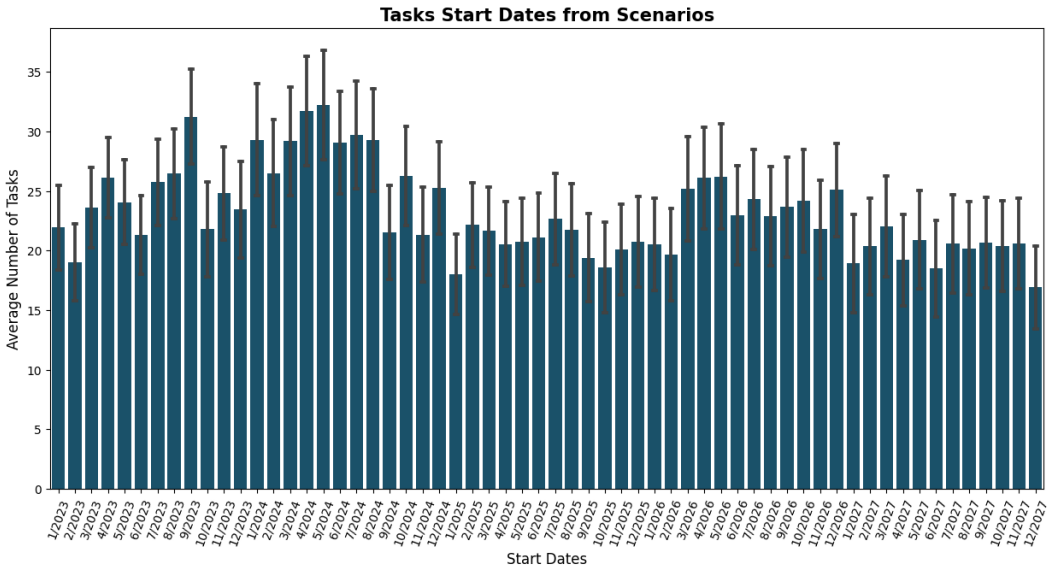



Figure 7: Average number of Tasks for the Start Dates from Scenarios

Figure 7 shows the average number of tasks for each month and year of the data. It is possible to see that between the end of 2023 and 2024, there is a higher number of tasks starting in comparison to the years that follow.

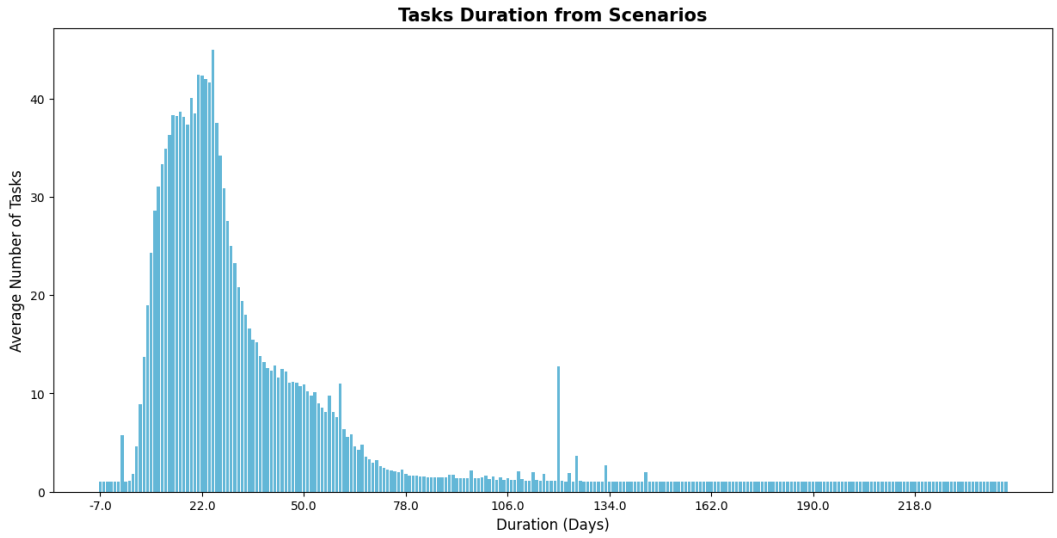



Figure 8: Average number of tasks for Duration from Scenarios

Figure 8 shows the average number of tasks for each duration from the Scenarios. Most of the tasks present a duration from 10 to 40 days, while there are some tasks that present a higher duration, above 80 days. It is possible to see that some of the tasks present negative or zero duration, which are values not possible for this variable. In order to maintain these tasks in the database, those values were processed afterward.

### 4.1.2 Imputation

Zero or negative values for task durations are not acceptable values since a task's duration should always have a positive value. To avoid removing these registers from the database, we decided to input the median of the tasks that had positive values for the ones that had zero or negative durations. Tasks with zero or negative duration in all the scenarios were removed. Table 4 shows the total number of tasks and the median of their duration and start date for the dataset before and after the imputation.

Table 4: Dataset before and after the imputation process

| Dataset | Pre-Imputation | Post Imputation |
|---|---|---|
| Total Number of Tasks | 2718212 | 2718212 |
| Duration | 25 [16 - 40] | 25 [16 - 40] |
| Difference in Start Dates | 5 [-13 - 73] | 5 [-13 - 73] |
| Number of tasks with negative duration | 14 | 0 |

### 4.1.3 Feature Engineering

To obtain information about the changes from the scenarios to the Deterministic Rig Schedule, two features were included in the data: The Difference in Start Date and the Relative Change in Duration, defined in Section 3.2. In addition, the data was restructured to gather relevant information about each scenario. This was made by grouping it by scenario and calculating their statistics. Table 5 describes the features in the database after the grouping was done.

Table 5: Statistics of the features grouped by Scenarios.

| Statistics by Scenario | Median [Q1 – Q3] |
|---|---|
| Task Duration (Mean) [days] | 32.03 [31.83 - 32.23] |
| Task Duration (Std) [days] | 24.75 [24.47 - 25.03] |
| Relative change in Duration (Min) | 0.2 [0.13 - 0.2] |
| Relative change in Duration (Mean) | 0.98 [0.97 - 0.99] |
| Relative change in Duration (Std) | 0.33 [0.32 - 0.34] |
| Relative change in Duration (Max) | 3.67 [3.3 - 4.27] |
| Relative change in Duration (p90) | 1.28 [1.27 - 1.29] |
| Number of Tasks | 1362 [1350 - 1371] |
| Absolute Difference in Start Date (Mean) | 79.39 [77.54 - 81.52] |
| Number of Tasks that were brought forward | 390 [369 - 410] |
| Number of Tasks that were pushed back | 710 [687 - 732] |
| Number of Tasks that had duration decreased | 710 [697 - 723] |
| Number of Tasks that had duration increased | 479 [468 - 491] |

Q1: First quartile; Q3: Third Quartile

From Table 5, we can see that most of the scenarios presented approximately 1360 tasks, their average duration is 32 days, and the Absolute Difference in Start Date (Mean) for the scenarios is approximately 79 days. There are a higher number of tasks that were pushed back and had their duration decreased in comparison to tasks that were brought forward or had their duration increased. Regarding the Relative change in Duration, the mean, standard deviation, and 90% Percentile (p90) of the scenarios have a small interquartile range, while the minimum and maximum present a higher range of values.

### 4.1.4
### Feature Selection

The Pearson correlation was applied to the database to identify highly correlated features. Figure 9 shows the results of the correlation analysis.

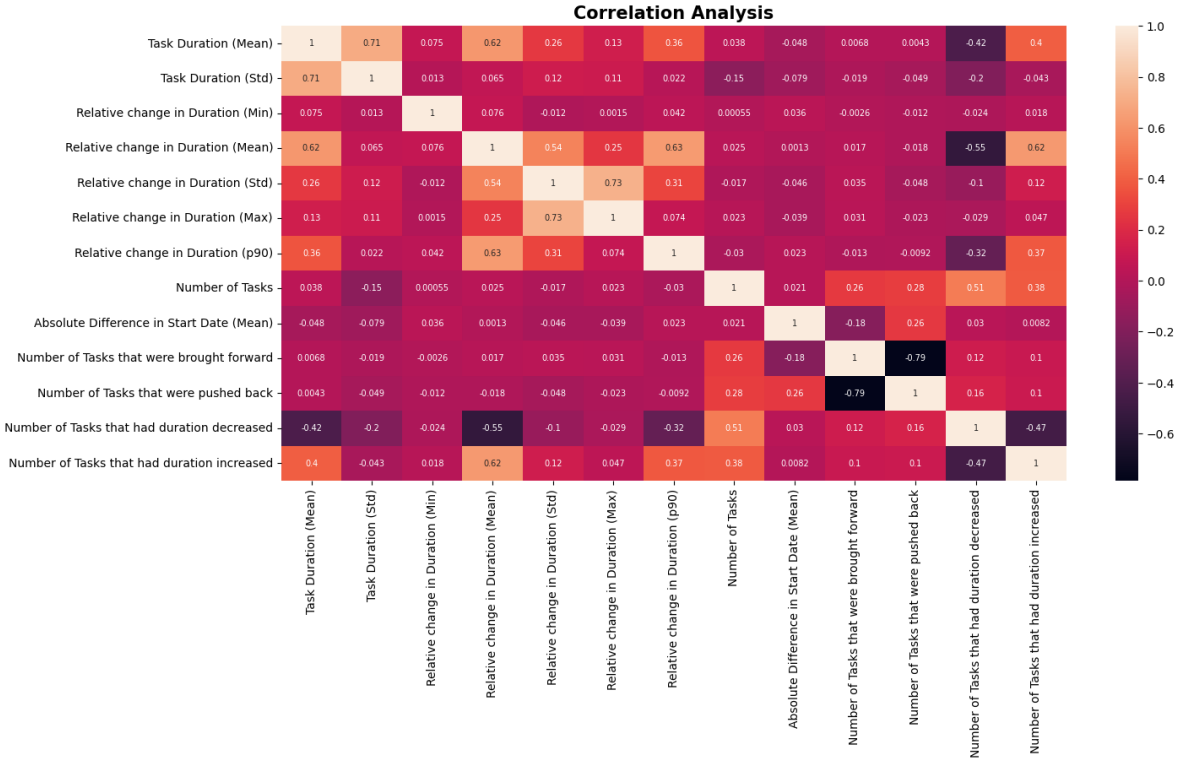



Figure 9: Correlation Analysis of the features in the dataset

Based on the results, it is possible to see which features are highly correlated with one another. We considered a high correlation to be above 0.6 or below -0.6. The Task Duration (Mean) and the Task Duration (Std) are highly correlated, and the first one is also correlated to the Relative Change in Duration (Mean). The latter is also highly correlated to The Relative Change in Duration (p90). The Relative Change in Duration (Max) is correlated to the Relative Change in Duration (Std). The Number of Tasks that were pushed back is also highly correlated to the Number of Tasks that were brought forward. The Number of Tasks that had duration increased is also highly correlated to the Task Duration (Mean).

Based on the results of the correlation analysis, it was decided to remove highly correlated features. The final dataset is composed of the following features for each scenario:

- Number of Tasks.
- Minimum of the Relative Change in Duration.

- Mean of the Relative Change in Duration.
- Maximum of the Relative Change in Duration.
- 90% Percentile (p90) of the Relative Change in Duration.
- Mean Absolute Difference in Start Date (between the deterministic schedule and the scenario).
- Number of Tasks that were brought forward.
- Number of Tasks that had duration decreased.

Since the Maximum of the Relative Change in Duration and the 90% Percentile of the Relative Change in Duration are features that consider higher to extreme cases of relative changes in duration, we chose to perform two cycles of the internal cycle of data science, where the first contains the Maximum of the Relative Change in Duration and the second one contains the 90% Percentile of the Relative Change. Based on the results, we would select the best feature to consider in our problem.

## 4.2 First Internal Cycle of Data Science

The first internal cycle of data science was performed using the following features:

- Number of Tasks.
- Minimum of the Relative Change in Duration.
- Mean of the Relative Change in Duration.
- Maximum of the Relative Change in Duration.
- Mean Absolute Difference in Start Date (between the deterministic schedule and the scenario).
- Number of Tasks that were brought forward.
- Number of Tasks that had duration decreased

### 4.2.1 Clustering Methods

The scenario reduction was performed using six clustering methods and evaluated with three evaluation metrics. Table 6 shows the results of the application of the clustering methods, varying the number of clusters from 2 to 1999. The best number of clusters was selected by the Silhouette Coefficient.

Table 6: Results of the scenario reduction in the first data science cycle

| Model | Number of Clusters | Silhouette Coefficient | Davies-Bouldin score | Calinski-Harabasz score |
|---|---|---|---|---|
| K-Means | 5 | 0.153 | 1.641 | 319.480 |
| K-Medoids | 4 | 0.122 | 2.118 | 277.012 |
| Agglomerative Clustering (Ward) | 782 | 0.169 | 0.856 | 34.529 |
| Agglomerative Clustering (Average) | 3 | 0.358 | 0.995 | 70.425 |
| Agglomerative Clustering (Complete) | 745 | 0.153 | 0.857 | 33.755 |
| Agglomerative Clustering (Single) | 3 | 0.392 | 0.415 | 5.492 |

The results of the clustering methods show that scenario reduction can be successfully made, as all the clusters obtained were below the original number of scenarios. It is possible to see that the methods showed two main results: a very low number of clusters or a higher number (near 750 clusters). The K-Means, K-Medoids, Agglomerative Clustering with Average and Single Linkage resulted in a few clusters from 3 to 5, while Agglomerative Clustering with Ward and Complete Linkage resulted in 782 and 745 number of clusters, respectively. The method with the best Silhouette Coefficient and Davies-Bouldin score was the Agglomerative Clustering Single Linkage, while K-Means presented the best Calinski-Harabasz score.

Even though some clustering methods obtained better evaluation metrics, we considered the number of clusters from these results to be extremely low and, thus, not statistically representative of the original scenarios. From a business perspective, if we reduce in this scale the number of scenarios, we could be removing important ones from the dataset. The other results, close to 750 clusters, seem more realistic and satisfactory for the scenario reduction. For these reasons, they were considered the best results from the clustering methods and chosen for the validation, along with the Agglomerative Clustering Single Linkage, which presented the best Silhouette Coefficient and Davies-Bouldin score.

### 4.2.2 Validation

Firstly, we compared the statistics from each group of scenarios to the original dataset, as shown in Table 7.

Table 7: Comparison of the Original Scenario to the Reduction subsets in the first cycle

| Statistics | 2000 Scenarios Median [Q1 – Q3] | 782 Scenarios Median [Q1 – Q3] | 745 Scenarios Median [Q1 – Q3] | 3 Scenarios Median [Q1 – Q3] |
|---|---|---|---|---|
| Relative change in Duration (Min) | 0.2 [0.13 - 0.2] | 0.2 [0.13 - 0.2] | 0.2 [0.13 - 0.2] | 0.26 [0.2 - 0.26] |
| Relative change in Duration (Mean) | 0.98 [0.97 - 0.99] | 0.98 [0.97 - 0.99] | 0.98 [0.97 - 0.99] | 0.99 [0.98 - 0.99] |
| Relative change in Duration (Max) | 3.67 [3.3 - 4.27] | 3.8 [3.33 - 4.47] | 3.8 [3.33 - 4.53] | 3.53 [3.1 - 4.23] |
| Number of Tasks | 1362 [1350 - 1371] | 1362 [1348 - 1372] | 1361 [1347 - 1372] | 1359 [1313.5 - 1361.5] |
| Absolute Difference in Start Date (Mean) | 79.39 [77.54 - 81.52] | 79.61 [77.59 - 82.11] | 79.69 [77.56 - 82.2] | 88.1 [83.66 - 93.74] |
| Number of Tasks that were brought forward | 390 [369 - 410] | 390 [367 - 412] | 390 [367 - 413] | 364 [354.5 - 378] |
| Number of Tasks that had duration decreased | 710 [697 - 723] | 710 [695 - 723.75] | 710 [695 - 724] | 708 [680.5 - 709] |

Q1: First quartile; Q3: Third Quartile

From the results shown in Table 7, we can see that the subset statistics for the 782 and 745 scenarios are close to the original one, according to the Median and the First and Third Quantiles, while the subset with three scenarios presents a higher difference.

Then, we calculated the Wasserstein Distance, comparing each feature from the reduction to the original set. The results of the Wasserstein Distance for these scenario reduction subsets are presented in Table 8.

Table 8: Wasserstein Distance for Scenario Reduction subsets of the first internal cycle of data science

| Features | Wasserstein Distance | | |
|---|---|---|---|
| | **782 Scenarios** | **745 Scenarios** | **3 Scenarios** |
| Relative change in Duration (Min) | 0.012 | 0.014 | 0.132 |
| Relative change in Duration (Mean) | 0.010 | 0.013 | 0.100 |
| Relative change in Duration (Max) | 0.017 | 0.019 | 0.046 |
| Number of Tasks | 0.015 | 0.019 | 0.200 |
| Absolute Difference in Start Date (Mean) | 0.014 | 0.017 | 0.326 |
| Number of Tasks that were brought forward | 0.011 | 0.015 | 0.126 |
| Number of Tasks that had duration decreased | 0.010 | 0.014 | 0.139 |
| **Mean Wasserstein Distance** | 0.013 | 0.016 | 0.153 |

In Table 8, it is possible to see that, for the clustering with a higher number of scenarios, the Mean Wasserstein Distance is considerably lower than the one for the subset with three scenarios. Based on these results, we could confirm our assumption that the scenario reductions performed by the Agglomerative Clustering with Ward and Complete Linkage successfully found a subset of scenarios that are representative of the original 2000 ones.

To further validate this, we plotted density plots for each subset and compared them to the original scenarios, as shown in Figure 10, Figure 11, and Figure 12.

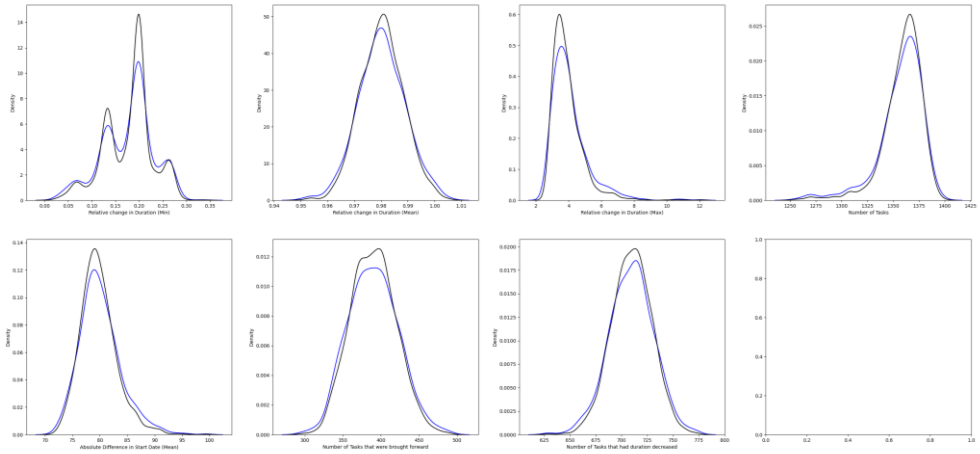


Figure 10: Density Plot of Original Scenarios (K=2000, color black) and Reduced Scenarios by Agglomerative Clustering Ward Linkage (K=782, color blue)

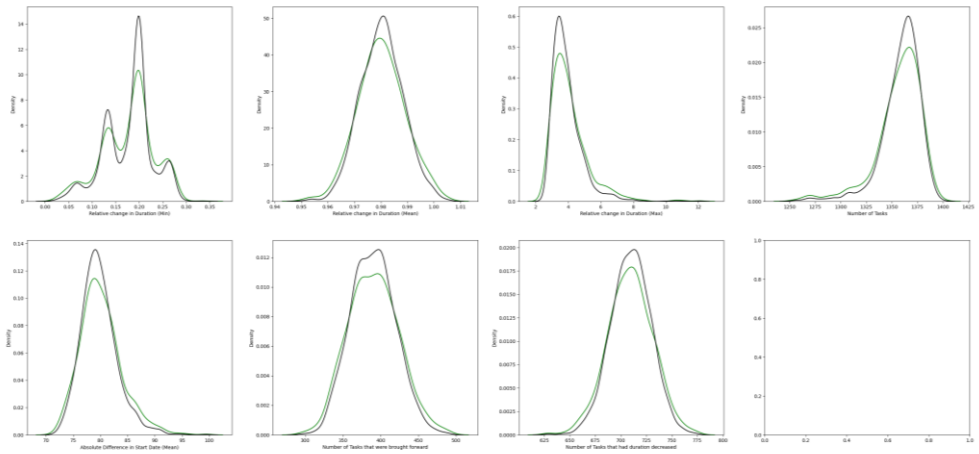


Figure 11: Density Plot of Original Scenarios (K=2000, color black) and Reduced Scenarios by Agglomerative Clustering Complete Linkage (K=745, color green)

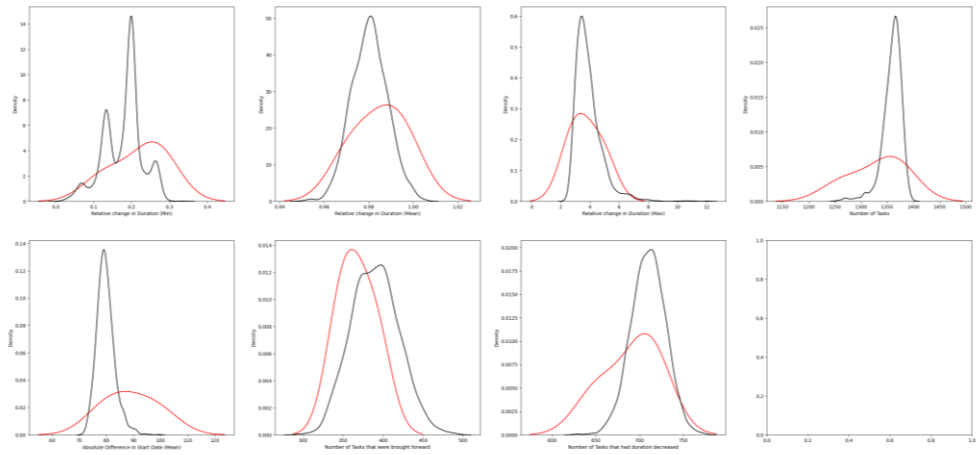


Figure 12: Density Plot of Original Scenarios (K=2000, color black) and Reduced Scenarios by Agglomerative Clustering Single Linkage (K=3, color red)

From the density plots, it is possible to see that for the Ward and Complete subsets, the distributions for each feature are very similar to the 2000 scenarios originally used. For the Single subset, the distribution of the reduced subset and the original one presents considerable differences for most of the features, confirming that the other subsets are a better scenario reduction for this problem.

The Agglomerative Clustering with Ward Linkage presented the lowest Mean Wasserstein Distance and was selected for the analysis of the impact of the scenario reduction in the calculation of the demand for tools and services using the decision support system as part of the output analysis. The demand was calculated using the 2000 scenarios and the 782 scenarios from the scenario reduction for the two groups of tools and rules mentioned in Section 3.3.2.2. Table 9 shows the results of the MAPE for each set and statistics from the calculation of the stochastic demand. The graphical visualization of the demand calculation for each tool and group is shown in APPENDIX II and APPENDIX III.

Table 9: Mean Absolute Percentual Error (MAPE) of the 782 Scenarios using the Agglomerative Clustering with Ward Linkage compared to the 2000 Scenarios

| **Statistics** | **MAPE (%) - 782 Scenarios** | |
|---|---|---|
| | **Group #1** | **Group #2** |
| Stochastic Demand - P10 | 0.47 [0.32 - 0.64] | 1.62 [0.37 - 3.96] |
| Stochastic Demand - P50 | 0.1 [0.06 - 0.14] | 0.22 [0.15 - 0.29] |
| Stochastic Demand - P90 | 0.19 [0.16 - 0.22] | 0.35 [0.1 - 0.63] |
| Stochastic Demand - Mean | 0.04 [0.02 - 0.06] | 0.2 [0.13 - 0.31] |
| Stochastic Demand - Minimum | 5.15 [4.57 - 5.67] | 3.98 [2.34 - 5.53] |
| Stochastic Demand - Maximum | 4.85 [4.5 - 5.19] | 3.79 [2.91 - 4.73] |

*95% confidence interval (percentile bootstrap with 500 resamples)

It is possible to see that the MAPE for each statistic of the demand calculation of each set is low, below 6% error. This shows that the reduction is possible, maintaining the characteristics of the original set. In terms of performance, Table 10 shows the execution time of the calculation of the demand for each set, the reduced subset, and the original set.

Table 10: Comparison of the execution time of the demand calculation in the first cycle

| **# Scenarios** | **Execution time (hours)** | |
|---|---|---|
| | **Group #1** | **Group #2** |
| 2000 | 7.41 | 7.61 |
| 782 | 2.73 | 2.83 |

From Table 10, we conclude that using the subset of scenarios obtained from the scenario reduction, it is possible to reduce approximately 63% of the execution time in the calculation of the demand.

## 4.3 Second Internal Cycle of Data Science

The second internal cycle of data science was performed using the following features:

- Number of Tasks.
- Minimum of the Relative Change in Duration.
- Mean of the Relative Change in Duration.
- 90% Percentile (p90) of the Relative Change in Duration.
- Mean Absolute Difference in Start Date.
- Number of Tasks that were brought forward.
- Number of Tasks that had duration decreased.

### 4.3.1 Clustering Methods

The scenario reduction for this cycle was performed with the same clustering methods and evaluation metrics. Table 11 shows the results of the clustering methods.

Table 11: Results of the scenario reduction in the second data science cycle

| Model | Number of Clusters | Silhouette Coefficient | Davies-Bouldin score | Calinski-Harabasz score |
|---|---|---|---|---|
| K-Means | 3 | 0.161 | 1.904 | 402.814 |
| K-Medoids | 3 | 0.137 | 2.148 | 360.602 |
| Agglomerative Clustering (Ward) | 714 | 0.168 | 0.900 | 34.463 |
| Agglomerative Clustering (Average) | 3 | 0.325 | 0.841 | 11.090 |
| Agglomerative Clustering (Complete) | 880 | 0.152 | 0.796 | 32.778 |
| Agglomerative Clustering (Single) | 3 | 0.368 | 0.460 | 4.288 |

As we saw in the first data science cycle, the methods showed similar results, maintaining a higher number of clusters for the Agglomerative Clustering with Ward and Complete Linkage and a very low number for the rest of the methods. The Agglomerative Clustering Single Linkage still was the method with the best

Silhouette Coefficient and Davies-Bouldin score, while the K-Means presented the best Calinski-Harabasz score.

### 4.3.2 Validation

Following the same steps as the First Internal Cycle of Data Science, a statistical validation was performed, as well as the calculation of the Wasserstein Distance for each feature and the density plots. Details of the statistical validation are shown in APPENDIX I.

The Agglomerative Clustering with Ward Linkage was also selected for the output analysis to observe the impact of the scenario reduction in the calculation of the demand for tools and services. Table 12 shows the results of the MAPE for each set and statistics from the calculation of the stochastic demand. The graphical visualization of the demand calculation for each tool and group is shown in APPENDIX IV and APPENDIX V.

Table 12: Mean Absolute Percentual Error (MAPE) of the 714 Scenarios using the Agglomerative Clustering with Ward Linkage compared to the 2000 Scenarios

| **Statistics** | **MAPE (%) - 714 Scenarios** | |
|---|---|---|
| | **Group #1** | **Group #2** |
| Stochastic Demand - P10 | 0.71 [0.61 - 0.83]* | 2.42 [0.46 - 6.2] |
| Stochastic Demand - P50 | 0.51 [0.31 - 0.84] | 0.44 [0.2 - 0.74] |
| Stochastic Demand - P90 | 0.21 [0.17 - 0.24] | 0.19 [0.03 - 0.4] |
| Stochastic Demand - Mean | 0.03 [0.01 - 0.06] | 0.15 [0.12 - 0.19] |
| Stochastic Demand - Minimum | 6.42 [5.78 - 7.06] | 4.96 [2.66 - 7.12] |
| Stochastic Demand - Maximum | 5.46 [5.02 - 5.89] | 4.01 [3.15 - 4.84] |

*95% confidence interval (percentile bootstrap with 500 resamples)

In this cycle, the MAPE for each statistic of the demand calculation of each set is a little bit higher than the ones calculated in the first cycle but still low, with a higher value below 7% error. Performance-wise, Table 13 shows the execution time of the calculation of the demand for each set, the reduced subset, and the original set.

Table 13: Comparison of the execution time of the demand calculation in the second cycle

| # Scenarios | Execution time (hours) | |
|---|---|---|
| | Group #1 | Group #2 |
| 2000 | 7.41 | 7.61 |
| 714 | 2.49 | 2.56 |

Similar to what was concluded in the first cycle, using the subset of scenarios from the scenario reduction reduces more than 65% of the execution time in the calculation of the demand for this number of scenarios.

## 4.4 Representative Scenarios

To find the minimal representative set of scenarios from the original set, we chose the clustering method and features that obtained the best performance based on the data modeling step.

Based on the clustering results and the Mean Wasserstein distance calculated from each cycle, the Agglomerative Clustering with Ward Linkage using the features from the first cycle resulted in the lowest distance, being considered the best method from our results.

As described in Section 3.4, the process of finding the minimal representative set of scenarios began by applying the clustering method, varying the number of clusters from 2 to 1999, and calculating the Mean Wasserstein Distance for each number of clusters. Figure 13 shows the decrease in this measure with the increase in the number of clusters.

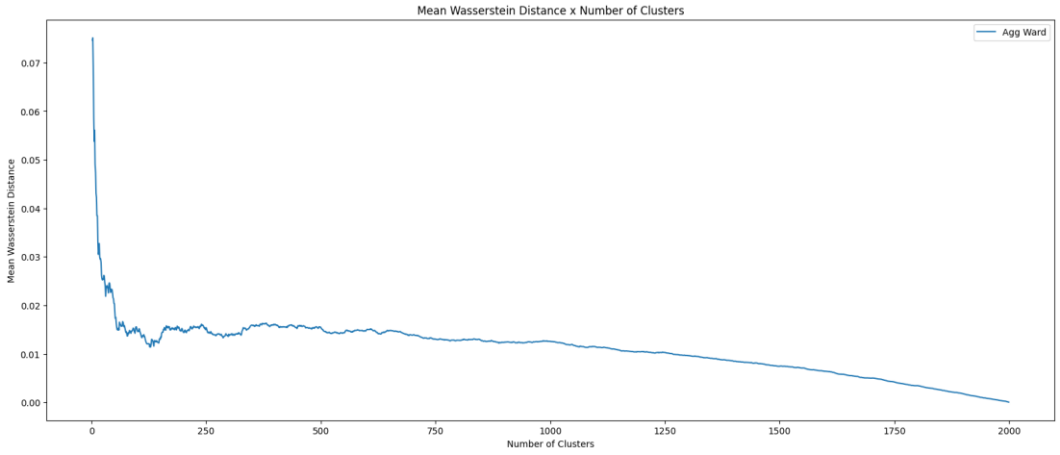



Figure 13: Mean Wasserstein Distance by Number of Clusters for Agglomerative Clustering with Ward Linkage

To identify the number of clusters where the decrease in the Mean Wasserstein Distance is small enough to be considered a minimal representative set, we calculated the decreasing rate and the change in the decreasing rate of this metric, and the absolute change in the decreasing rate is shown in Figure 14.

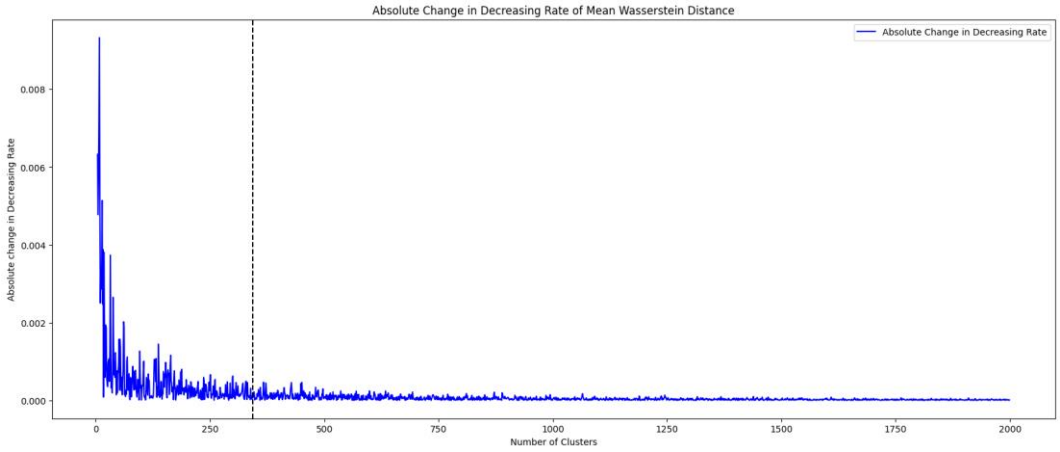



Figure 14: Absolute change in decreasing rate of Mean Wasserstein Distance (the dashed line represents the 343 reduced scenario subset)

Applying the steps of the algorithm in Figure 6, the first number of clusters where we have an absolute change in decreasing rate smaller than $10^{-4}$ is when the number of clusters is equal to 343 clusters.

Following the same validation process, we compared the statistics from the scenario reduction with 343 clusters and the original dataset, as shown in Table 14. We also compared it to the scenario reduction subset from the best clustering method in the first internal data science cycle, which has the same features as the minimal representative subset.

Table 14: Comparison of the Original Scenario to the Minimal Representative Subset

| Statistics | 2000 Scenarios Median [Q1 – Q3] | 782 Scenarios Median [Q1 – Q3] | 343 Scenarios Median [Q1 – Q3] |
|---|---|---|---|
| Relative change in Duration (Min) | 0.2 [0.13 - 0.2] | 0.2 [0.13 - 0.2] | 0.2 [0.13 - 0.2] |
| Relative change in Duration (Mean) | 0.98 [0.97 - 0.99] | 0.98 [0.97 - 0.99] | 0.98 [0.97 - 0.99] |
| Relative change in Duration (Max) | 3.67 [3.3 - 4.27] | 3.8 [3.33 - 4.47] | 3.8 [3.33 - 4.55] |
| Number of Tasks | 1362 [1350 - 1371] | 1362 [1348 - 1372] | 1361 [1348 - 1372] |
| Absolute Difference in Start Date (Mean) | 79.39 [77.54 - 81.52] | 79.61 [77.59 - 82.11] | 79.73 [77.83 - 81.77] |
| Number of Tasks that were brought forward | 390 [369 - 410] | 390 [367 - 412] | 389 [368 - 412.5] |
| Number of Tasks that had duration decreased | 710 [697 - 723] | 710 [695 - 723.75] | 709 [695 - 724] |

Q1: First quartile; Q3: Third Quartile

As we can see from Table 14, the statistics from the minimal representative subset are similar to the original scenario one but slightly farther from the original subset in comparison to the 782 for some features, such as the Number of Tasks or the Absolute Difference in Start Date (Mean). This is expected, as we are considering fewer scenarios. Figure 15 shows the visual comparison of each feature from the minimal representative subset with the original set.

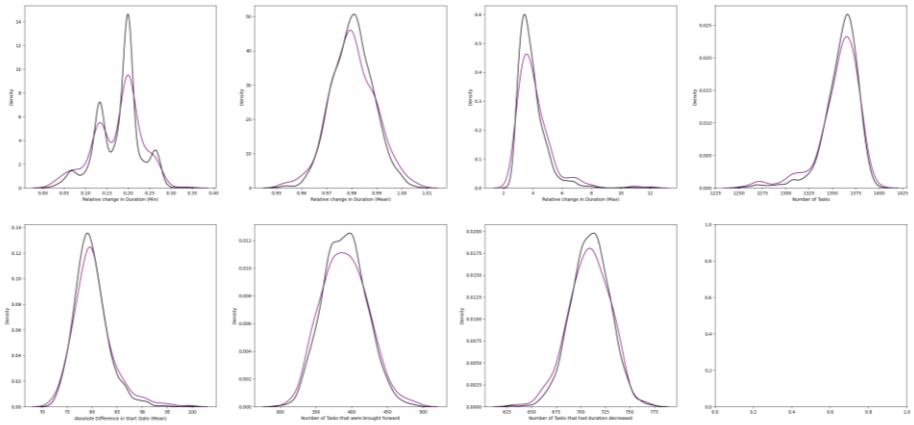


Figure 15: Density Plot of Original Scenarios (K=2000, color black) and Minimal Representative Subset Scenarios (K=343, color purple)

As seen in the results from the first and second data science cycles, Figure 15 shows that the features from the reduced subset are very similar to the original

dataset. This further validates that reduction can maintain the statistical characteristics of the distribution, even for a reduced set of scenarios.

To analyze the impact of this scenario reduction in the demand calculation, we applied this subset to the same group of tools and services previously used in the validation. Table 15 shows the results of the MAPE for each group and statistics from the calculation of the stochastic demand. The graphical visualization of the demand calculation for each tool and group is shown in APPENDIX VI and APPENDIX VII.

Table 15: Mean Absolute Percentual Error (MAPE) of the 343 Scenarios compared to the 2000 Scenarios

| **Statistics** | **MAPE (%) - 343 Scenarios** | |
|---|---|---|
| | **Group #1** | **Group #2** |
| Stochastic Demand - P10 | 1.69 [1.4 - 2]* | 1.54 [0.47 - 3.37] |
| Stochastic Demand - P50 | 0.28 [0.12 - 0.56] | 0.54 [0.28 - 0.92] |
| Stochastic Demand - P90 | 0.22 [0.17 - 0.27] | 0.47 [0.17 - 0.82] |
| Stochastic Demand - Mean | 0.12 [0.1 - 0.15] | 0.3 [0.22 - 0.42] |
| Stochastic Demand - Minimum | 11.03 [9.83 - 12.16] | 9.55 [5.45 - 13.48] |
| Stochastic Demand - Maximum | 10.48 [9.56 - 11.39] | 8.14 [6.49 - 9.93] |

*95% confidence interval (percentile bootstrap with 500 resamples)

As expected, the lower the number of scenarios used for the calculation of the demand, the higher the error obtained from the results. Nonetheless, the MAPE of the demand calculation from the 343 subset is still low for most of the statistics, having the highest value of 11%. In terms of execution time, Table 16 shows the execution time for each group, using the minimal representative scenarios subset and the original set.

Table 16: Comparison of the execution time of the demand calculation using the minimal representative subset

| **# Scenarios** | **Execution time (hours)** | |
|---|---|---|
| | **Group #1** | **Group #2** |
| 2000 | 7.41 | 7.61 |
| 343 | 1.18 | 1.26 |

As we can see from Table 16, the use of the minimal representative subset of scenarios reduces considerably, approximately 84%, the execution time of the

demand calculation. This alternative allows decision-makers to have faster demand results and run the calculation more times, if necessary.

# 5
# Discussion

The proposed methodology followed the steps of data preparation, modeling, and the search for a representative scenarios' subset for the scenario reduction. Using clustering-based methods, the number of clusters obtained varied from a very small number for some methods to a higher amount for others. In the process of validating these results and comparing them to the original set of scenarios, it was concluded that the methods with a higher number of clusters were statistically closer to the original set and thus considered a better fit for our problem.

The Agglomerative Clustering with Ward Linkage obtained better evaluation metrics and was closer to the original set according to the Wasserstein Distance, so it was selected as the best method, with 782 clusters. This reduction resulted in a reduction in the execution time of the demand calculation by approximately 63%, with an error below 6%.

Using this clustering method to find the minimal representative scenario subset, it is possible to reduce the number of clusters even further to 343 clusters. Considering that the processing time of the data is rather small, of approximately 5 minutes, this represents a reduction of the execution time of the demand calculation of 84%, with the highest mean absolute percentual error of 11%. Even though the error is slightly higher, the reduction in the execution time allows decision-makers to have faster demand results and more flexibility to test different tools and service options and set of rules for the schedule and analyze their impact on demand.

As mentioned in Section 2.3, Vieira (2021) also proposed a methodology for selecting scenarios for the stochastic calculation of the demand for tools and services for well-construction tasks performed by offshore rigs. The Set Covering Problem was used to select the scenarios and solved in its classic version using an exact algorithm and a heuristic algorithm.

Their methodology consisted of acquiring the main characteristics of the scenarios, calculating the distances between the scenarios and the graph characterization using the Gower distance, and pruning the graph based on a

maximum distance allowed (cutoff) to limit the solution space. Then, the coverage sets are defined, and the algorithm of the Set Covering Problem is applied. The last step is to assign weights to the representative scenarios based on their representation potential (VIEIRA, 2021).

Using this methodology to perform the scenario reduction for a set with 3000 scenarios, Vieira (2021) obtained a reduction subset with 270 scenarios, using a cutoff of 0.075. This was a 91% reduction in terms of data processed for the calculation of the demand, reducing only 5% of the assertiveness in the group of tools and services used to analyze the results.

The methodology presented by Vieira (2021) and the one in this study are both able to perform successful scenario reductions for this problem, with little assertiveness loss, being possible options for the solution of this problem. Vieira (2021) presented a limitation in the validation of the methodology, as they were not able to perform a statistical validation due to the original scenarios' distributions being unknown. This was overcome in our methodology because, even though the distributions were still unknown, a statistical validation was performed using the Wasserstein metric and visualization tools in order to select representative scenarios in our reduction.

Moreover, Meira et al (2016) also proposed a methodology for scenario reduction to identify representative scenarios in oil fields. To perform the scenario reduction, they used optimization methods, and a mathematical function that modeled the representativeness of the scenarios. This was implemented in a tool called *RMFinder*, which iterates to find a good set of parameters for the set of models in the problem and searches for the set of representative models to minimize their overall cost. Similarly, we proposed an iterative algorithm to find a minimal representative set of scenarios. This was achieved increasing the number of clusters, applying the scenario reduction using clustering methods, and calculating the representativeness metric, which was the Mean Wasserstein Distance, and stopping once the improvement of this metric was too small.

# 6
# Conclusion

The construction and maintenance of marine wells are essential parts of oil exploration and production. Operations such as drilling, assessment, completion, and workover performed by offshore rigs need tools and services to be executed, which is why material resource planning is an important part of the planning of the rig's schedule. Moreover, there are uncertainties present in the operation, as the tasks' start dates and duration are often different from what was initially planned, which needs to be taken into consideration for a better estimation of the demand for tools and services.

This work aimed to reduce the number of scenarios used to calculate the demand for tools and services for well construction tasks of a large Oil and Gas company, finding the most representative ones using clustering-based methods and statistical analysis. The methodology proposed in this study followed three main steps: preparation of the data, modeling using clustering methods, and finding the representative scenarios. The clustering method that obtained the best results was Agglomerative Clustering with Ward Linkage, using the features from the first internal data science cycle. Even though this method did not result in the best Silhouette Coefficient (0.169), it had the lowest Mean Wasserstein Distance (0.013) in the statistical validation, thus being considered the best method with 782 clusters. Using this reduction in the calculation of the demand for tools and services provided a 63% reduction in the execution time of the demand, with a maximum MAPE of approximately 5% in comparison to the demand calculated with the original scenarios.

We used the best clustering method and features to find a minimal representative set of scenarios for our scenario reduction. Using the Mean Wasserstein Distance and analyzing the change in the decreasing rate, the algorithm to find this subset resulted in a reduction to 343 clusters. This subset reduced the execution time of the demand calculation for tools and services by approximately 84%, with the highest MAPE of 11% for the minimum statistic.

Scenario reduction in this context is important for the company in question, as it can help streamline decision-making and the planning of tool and services contracts. As the demand calculation is subject to the group of tools and set of rules that the user wants to evaluate, the decision maker can calculate the demand more times a day or test more tools and rules to be able to correctly estimate the demand for tools and services, according to their knowledge of the business.

The limitations presented in this study include the subjectivity in the calculation of the demand, which requires human input for the selection of tools and services, as well as rules for the allocation, maintenance and dependence conditions, and logistic parameters, one of the limitations. Due to these conditions, it is not possible to evaluate the methodology in every group of tools and set of rules, and our results focus on only two cases. In future research, we propose the application of the methodology in other examples of groups of tools and rules and performing the validation of the scenario reduction with other statistical metrics.

# 7
# References


ABOUELROUS, A.; GABOR, A. F.; ZHANG, Y. Optimizing the inventory and fulfillment of an omnichannel retailer: a stochastic approach with scenario clustering. **Computers & Industrial Engineering**, v. 173, p. 108723, nov. 2022.

ARBELAITZ, O. et al. An extensive comparative study of cluster validity indices. **Pattern Recognition**, v. 46, n. 1, p. 243–256, jan. 2013.

BASSI, H. V.; FERREIRA FILHO, V. J. M.; BAHIENSE, L. Planning and scheduling a fleet of rigs using simulation–optimization. **Computers & Industrial Engineering**, v. 63, n. 4, p. 1074–1088, dez. 2012.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics - Theory and Methods**, v. 3, n. 1, p. 1–27, 1974.

CHAPALOGLOU, S. et al. Data-informed scenario generation for statistically stable energy storage sizing in isolated power systems. **Journal of Energy Storage**, v. 51, p. 104311, jul. 2022.

CONDEIXA, L.; OLIVEIRA, F.; SIDDIQUI, A. S. **Wasserstein-Distance-Based Temporal Clustering for Capacity-Expansion Planning in Power Systems**. 2020 International Conference on Smart Energy Systems and Technologies (SEST). **Anais**... Em: 2020 INTERNATIONAL CONFERENCE ON SMART ENERGY SYSTEMS AND TECHNOLOGIES (SEST). Istanbul, Turkey: IEEE, set. 2020. Disponível em: <https://ieeexplore.ieee.org/document/9203449/>. Acesso em: 15 nov. 2023

DAVIES, D. L.; BOULDIN, D. W. A Cluster Separation Measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. PAMI-1, n. 2, p. 224–227, abr. 1979.

DEVOLD, H. **Oil and gas production handbook: an introduction to oil and gas production**. USA: [s.n.].

DROUVEN, M. G.; GROSSMANN, I. E. Multi-period planning, design, and strategic models for long-term, quality-sensitive shale gas development. **AIChE Journal**, v. 62, n. 7, p. 2296–2323, jul. 2016.

EFRON, B. Bootstrap Methods: Another Look at the Jackknife. 1979.

EFRON, B.; TIBSHIRANI, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. **Statistical Science**, Institute of Mathematical Statistics. v. 1, n. 1, p. 54–75, 1986.

GIL, E.; ARAVENA, I.; CARDENAS, R. Generation Capacity Expansion Planning Under Hydro Uncertainty Using Stochastic Mixed Integer Programming and Scenario Reduction. **IEEE Transactions on Power Systems**, v. 30, n. 4, p. 1838–1847, jul. 2015.

GROWE-KUSKA, N.; HEITSCH, H.; ROMISCH, W. **Scenario reduction and scenario tree construction for power management problems**. 2003 IEEE Bologna Power Tech Conference Proceedings,. **Anais**... Em: 2003 IEEE BOLOGNA POWER TECH. Bologna, Italy: IEEE, 2003. Disponível em: <http://ieeexplore.ieee.org/document/1304379/>. Acesso em: 4 ago. 2023

HALKIDI, M. On Clustering Validation Techniques. 2001.

HEITSCH, H.; MISCH, W. R. Scenario Reduction Algorithms in Stochastic Programming. 2003.

HU, J.; LI, H. A New Clustering Approach for Scenario Reduction in Multi-Stochastic Variable Programming. **IEEE Transactions on Power Systems**, v. 34, n. 5, p. 3813–3825, set. 2019.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys**, v. 31, n. 3, p. 264–323, set. 1999.

JAMES, G. et al. Unsupervised Learning. Em: **An Introduction to Statistical Learning: with Applications in Python**. Cham: Springer International Publishing, 2023. p. 503–556.

KAISER, M. J.; SNYDER, B. The five offshore drilling rig markets. **Marine Policy**, v. 39, p. 201–214, maio 2013.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data: an introduction to cluster analysis**. Hoboken, N.J: Wiley, 2005.

LI, H. et al. A review of scenario analysis methods in planning and operation of modern power systems: Methodologies, applications, and challenges. **Electric Power Systems Research**, v. 205, p. 107722, abr. 2022a.

LI, W. D.; MCMAHON, C. A. A simulated annealing-based optimization approach for integrated process planning and scheduling. **International Journal of Computer Integrated Manufacturing**, v. 20, n. 1, p. 80–95, jan. 2007.

LI, X. et al. Multi-year planning for the integration combining distributed energy system and electric vehicle in neighborhood based on data-driven model. **International Journal of Electrical Power & Energy Systems**, v. 140, p. 108079, set. 2022b.

MACQUEEN, J. Some Methods for Classification and Analysis of Multivariate Observations. **MULTIVARIATE OBSERVATIONS**, 1 jan. 1967.

MAHJOUR, S. K. et al. Scenario reduction methodologies under uncertainties for reservoir development purposes: distance-based clustering and metaheuristic

algorithm. **Journal of Petroleum Exploration and Production Technology**, v. 11, n. 7, p. 3079–3102, jul. 2021.

MAHJOUR, S. K. et al. Evaluation of unsupervised machine learning frameworks to select representative geological realizations for uncertainty quantification. **Journal of Petroleum Science and Engineering**, v. 209, p. 109822, fev. 2022.

MARCHESI, J. F. et al. Otimização do planejamento de projetos: aplicação à construção de poços marítimos em uma indústria de Óleo e Gás. 2019.

MEIRA, L. A. A. et al. Selection of Representative Models for Decision Analysis Under Uncertainty. **Computers & Geosciences**, v. 88, p. 67–82, mar. 2016.

MEIRA, L. A. A. et al. Improving representativeness in a scenario reduction process to aid decision making in petroleum fields. **Journal of Petroleum Science and Engineering**, v. 184, p. 106398, jan. 2020.

OKADA, R. et al. Scenario reduction using machine learning techniques applied to conditional geostatistical simulation. **REM - International Engineering Journal**, v. 72, n. 1 suppl 1, p. 63–68, mar. 2019.

OSMUNDSEN, P.; ROLL, K.; TVETERÅS, R. Exploration Drilling Productivity at the Norwegian Shelf. **Journal of Petroleum Science and Engineering**, v. 73, p. 122–128, 1 ago. 2010.

PANARETOS, V. M.; ZEMEL, Y. Statistical Aspects of Wasserstein Distances. **Annual Review of Statistics and Its Application**, v. 6, n. 1, p. 405–431, 7 mar. 2019.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, nov. 1987.

SANTOS, I. M.; HAMACHER, S.; OLIVEIRA, F. A Systematic Literature review for the rig scheduling problem: Classification and state-of-the-art. **Computers & Chemical Engineering**, v. 153, p. 107443, out. 2021.

SUMAILI, J. et al. **Finding representative wind power scenarios and their probabilities for stochastic models**. 2011 16th International Conference on Intelligent System Applications to Power Systems. **Anais**... Em: 2011 16TH INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEM APPLICATIONS TO POWER SYSTEMS (ISAP). Hersonissos, Greece: IEEE, set. 2011. Disponível em: <http://ieeexplore.ieee.org/document/6082195/>. Acesso em: 5 ago. 2023

SUREJA, N.; CHAWDA, B.; VASANT, A. An improved K-medoids clustering approach based on the crow search algorithm. **Journal of Computational Mathematics and Data Science**, v. 3, p. 100034, jun. 2022.

SUSLICK, S. B.; SCHIOZER, D.; RODRIGUEZ, M. R. Uncertainty and Risk Analysis in Petroleum Exploration and Production. 2009.

VIEIRA, I. F. G. **REDUÇÃO DE CENÁRIOS COM FORMULAÇÃO DE COBERTURA DE CONJUNTOS: UMA APLICAÇÃO NA INDÚSTRIA DE PETRÓLEO**. MESTRE EM ENGENHARIA DE PRODUÇÃO—Rio de Janeiro, Brazil: PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO, 12 abr. 2021.

WU, X. et al. Top 10 algorithms in data mining. **Knowledge and Information Systems**, v. 14, n. 1, p. 1–37, jan. 2008.

## APPENDIX I – Statistical Validation of the Second Internal Cycle of Data Science

Firstly, we compared the statistics from each group of scenarios to the original dataset, as shown in Table 17.

Table 17: Comparison of the Original Scenario to the Reduction subsets in the second cycle

| Statistics | 2000 Scenarios Median [Q1 – Q3] | 714 Scenarios Median [Q1 – Q3] | 880 Scenarios Median [Q1 – Q3] | 3 Scenarios Median [Q1 – Q3] |
|---|---|---|---|---|
| Relative change in Duration (Min) | 0.2 [0.13 - 0.2] | 0.2 [0.13 - 0.2] | 0.2 [0.13 - 0.2] | 0.13 [0.13 - 0.17] |
| Relative change in Duration (Mean) | 0.98 [0.97 - 0.99] | 0.98 [0.97 - 0.99] | 0.98 [0.97 - 0.99] | 0.97 [0.97 - 0.98] |
| Relative change in Duration (p90) | 3.67 [3.3 - 4.27] | 1.28 [1.27 - 1.29] | 1.28 [1.27 - 1.29] | 1.27 [1.26 - 1.3] |
| Number of Tasks | 1362 [1350 - 1371] | 1361 [1347 - 1371] | 1361 [1347 - 1371] | 1364 [1361 - 1372] |
| Absolute Difference in Start Date (Mean) | 79.39 [77.54 - 81.52] | 79.49 [77.45 - 81.98] | 79.49 [77.41 - 81.96] | 94.79 [88.3 - 97.09] |
| Number of Tasks that were brought forward | 390 [369 - 410] | 389 [367 - 411] | 389.5 [367 - 411] | 349 [347 - 370] |
| Number of Tasks that had duration decreased | 710 [697 - 723] | 710 [694.25 - 723] | 709 [695 - 724] | 710 [709.5 - 739.5] |

Q1: first quartile; Q3: third quartile

As seen in Table 17 the results are similar to the ones observed in the first cycle, where the subsets statistics for the 714 and 880 scenarios are extremely close to the original one, according to the Median and the First and Third Quantiles, while the subset with 3 scenarios presents a higher difference.

The Wasserstein Distance for these scenario reduction subsets are presented in Table 18.

Table 18: Wasserstein Distance for Scenario Reduction subsets of the second cycle

| Features | Wasserstein Distance | | |
|---|---|---|---|
| | **714 Scenarios** | **880 Scenarios** | **3 Scenarios** |
| Relative change in Duration (Min) | 0.016 | 0.014 | 0.108 |
| Relative change in Duration (Mean) | 0.013 | 0.014 | 0.090 |
| Relative change in Duration (p90) | 0.016 | 0.019 | 0.185 |
| Number of Tasks | 0.018 | 0.017 | 0.065 |
| Absolute Difference in Start Date (Mean) | 0.012 | 0.013 | 0.436 |
| Number of Tasks that were brought forward | 0.011 | 0.012 | 0.153 |
| Number of Tasks that had duration decreased | 0.013 | 0.014 | 0.142 |
| **Mean Wasserstein Distance** | 0.014 | 0.014 | 0.168 |

These results are also similar to the ones observed in the first cycle, where the clustering with a higher number of scenarios have a lower metric than the subset with 3 scenarios, which indicates that they are a better scenario reduction subset for our original scenarios.

We also plotted density plots for each subset and compared them to the original scenarios to validate our findings, as shown in Figures 16, 17 and 18.

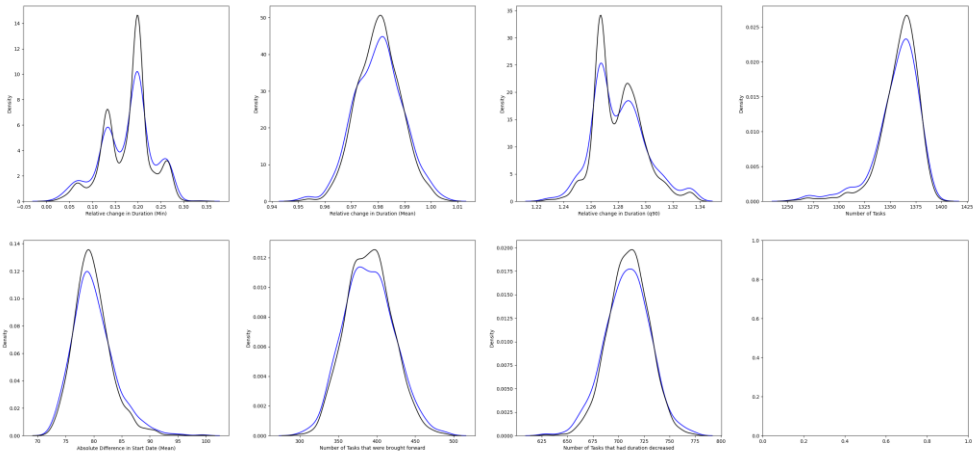


Figure 16: Density Plot of Original Scenarios (K=2000, color black) and Reduced Scenarios by Agglomerative Clustering Ward Linkage (K=714, color blue)

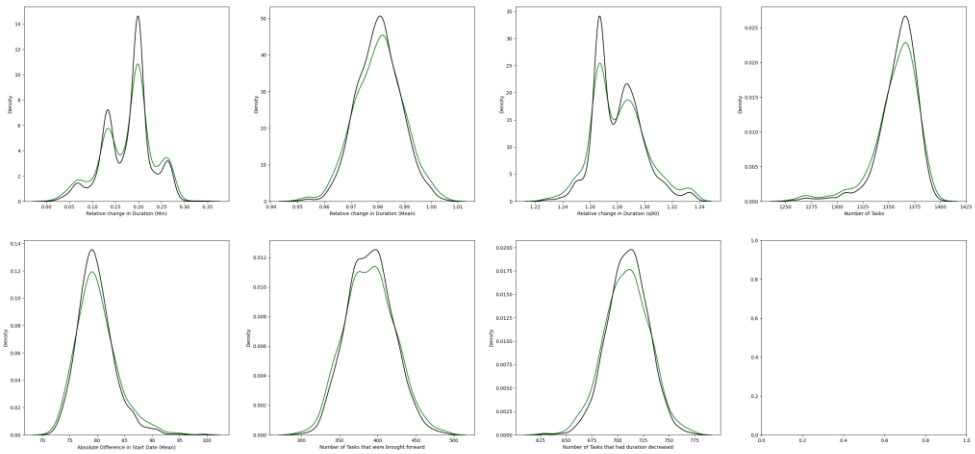


Figure 17: Density Plot of Original Scenarios (K=2000, color black) and Reduced Scenarios by Agglomerative Clustering Complete Linkage (K=880, color green)

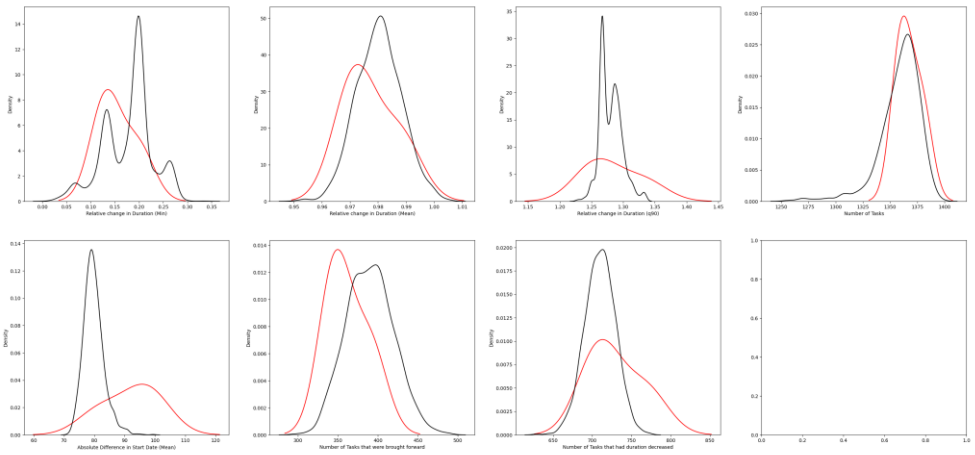
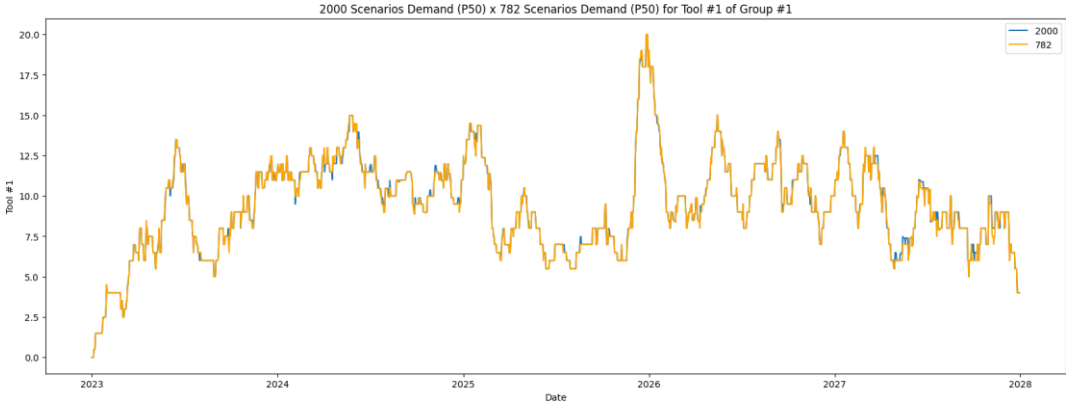
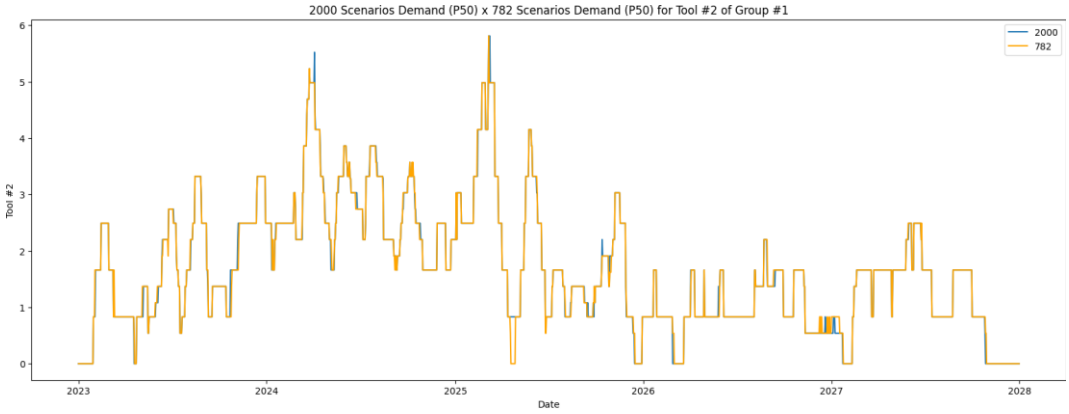


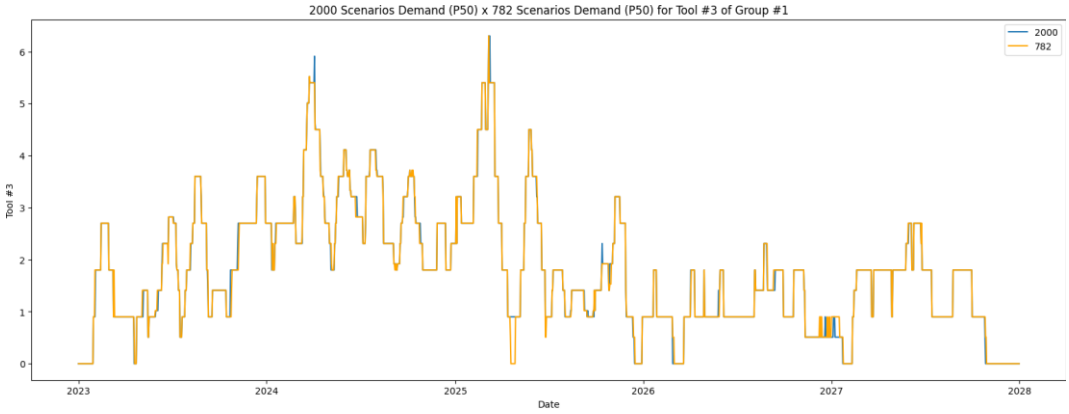
Figure 18: Density Plot of Original Scenarios (K=2000, color black) and Reduced Scenarios by Agglomerative Clustering Single Linkage (K=3, color red)

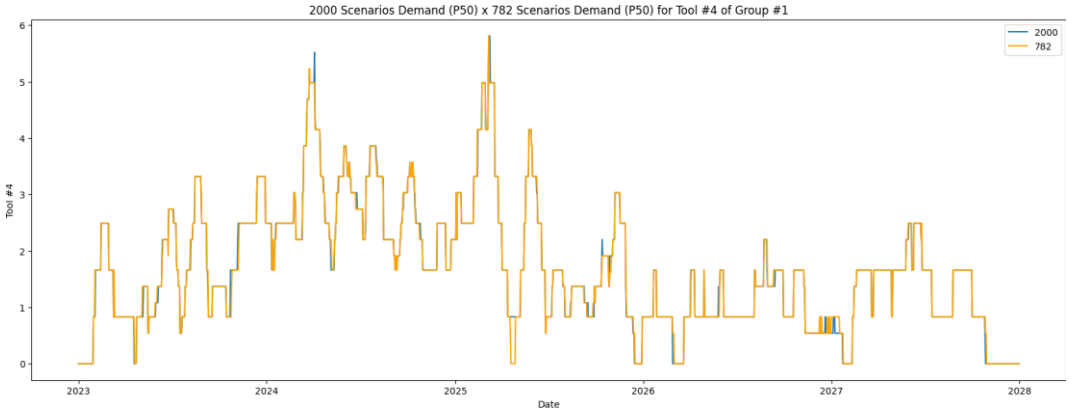
As observed in the first cycle, the density plot shows similar results, where it is possible to see that for the Ward and Complete subsets, the distributions for each feature are very similar to the 2000 scenarios originally used, and for the Single subset, the differences between distributions are more noticeable.

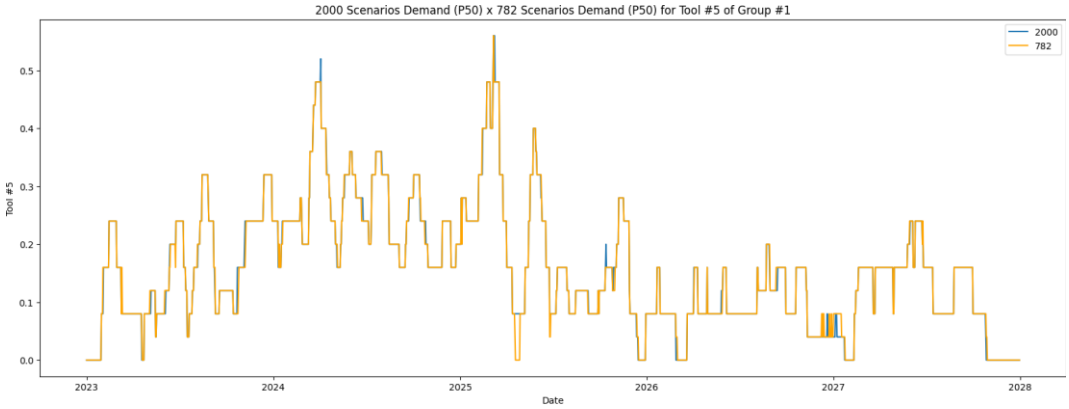
# APPENDIX II – Comparison of P50 of Demand Calculation for Original Set and 782 Scenarios for Group 1

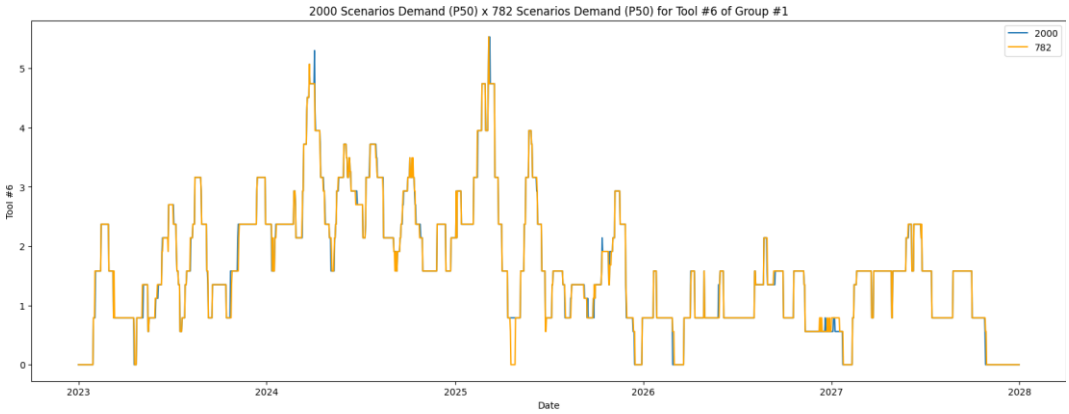


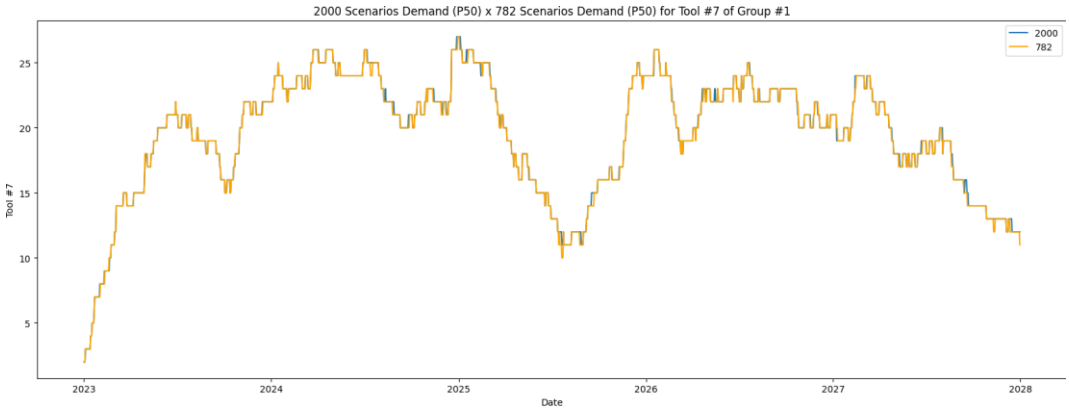



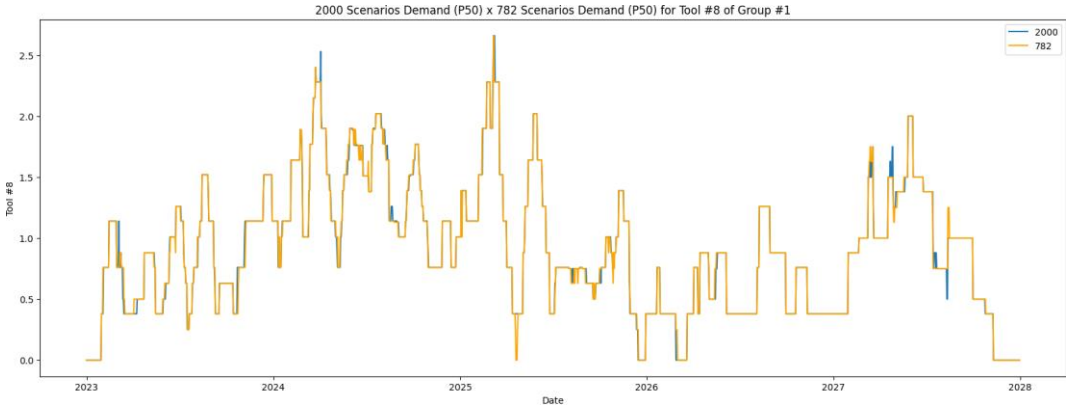



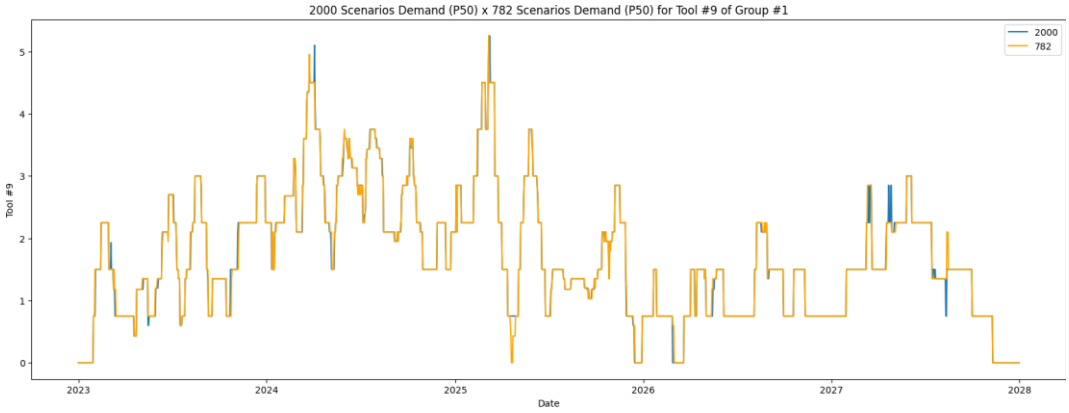



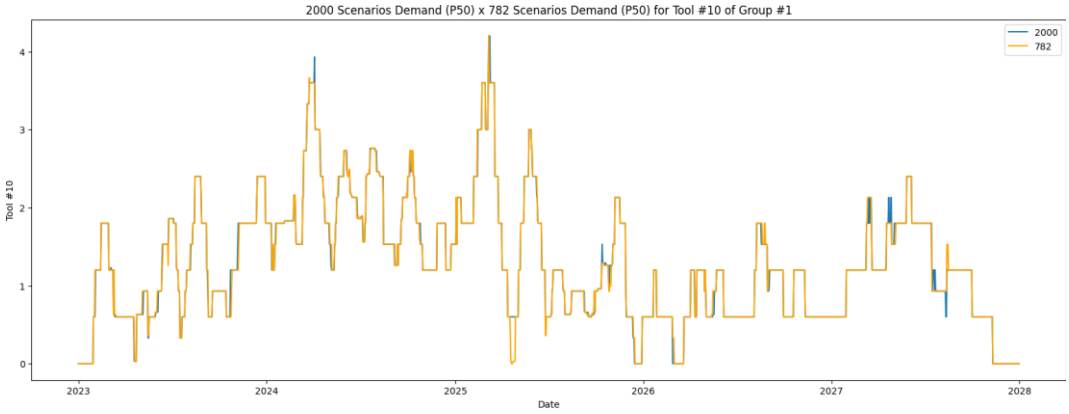

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #5 of Group #1
2000
782
Tool #5
Date
2023
2024
2025
2026
2027
2028

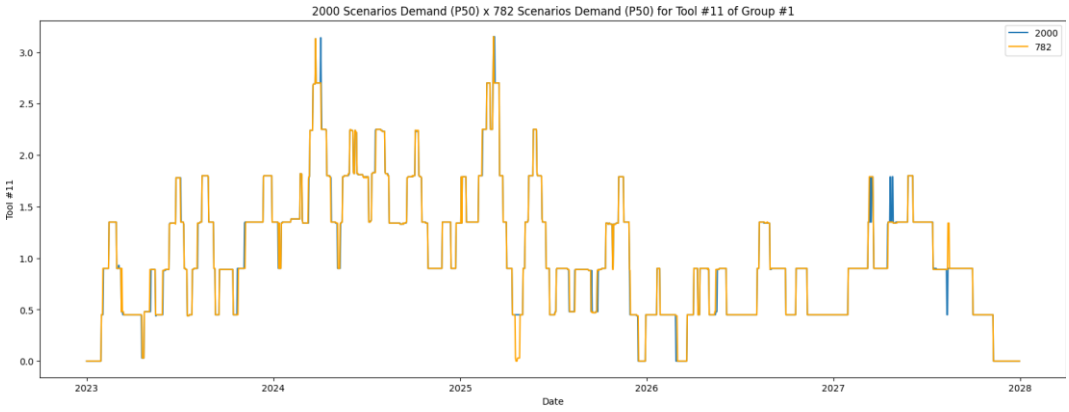

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #6 of Group #1
2000
782
Tool #6
Date
2023
2024
2025
2026
2027
2028

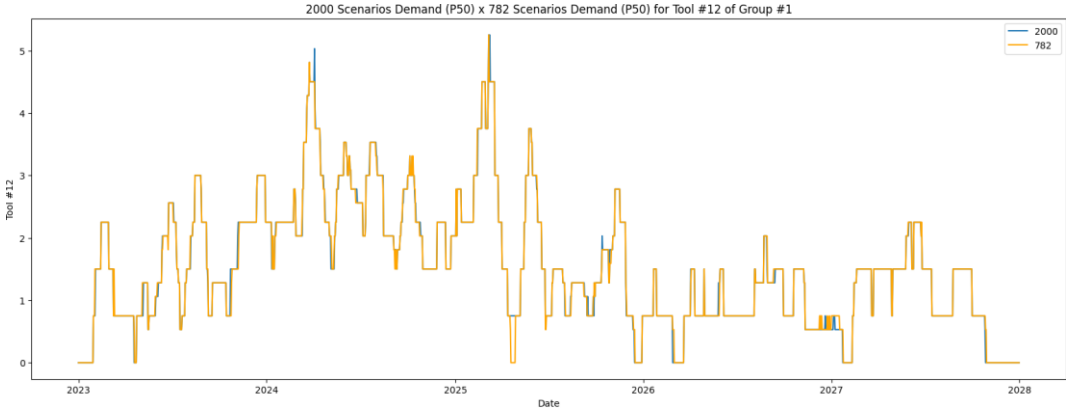

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #7 of Group #1
2000
782
Tool #7
Date
2023
2024
2025
2026
2027
2028

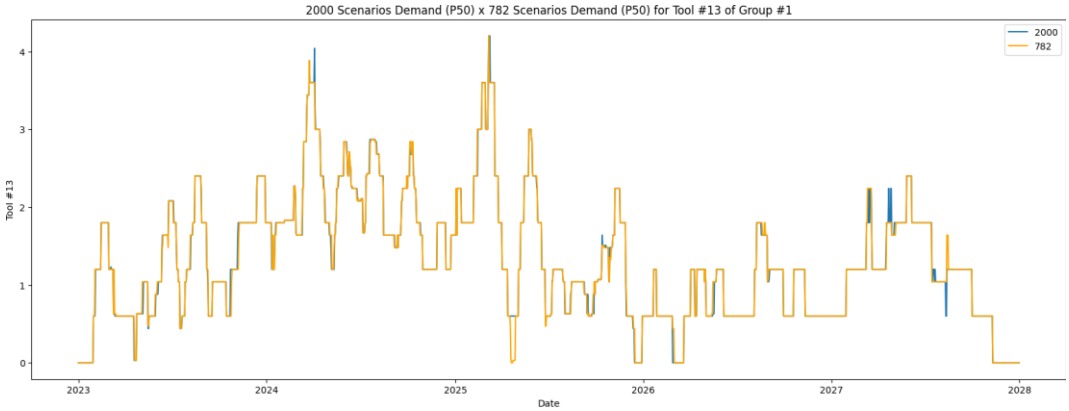

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #8 of Group #1
2000
782
Tool #8
Date
2023
2024
2025
2026
2027
2028

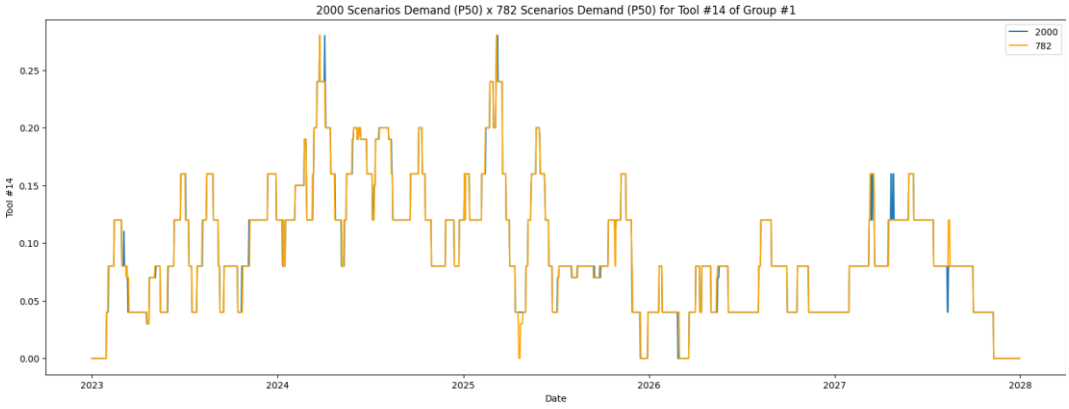

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #9 of Group #1
2000
782
Tool #9
Date
2023
2024
2025
2026
2027
2028

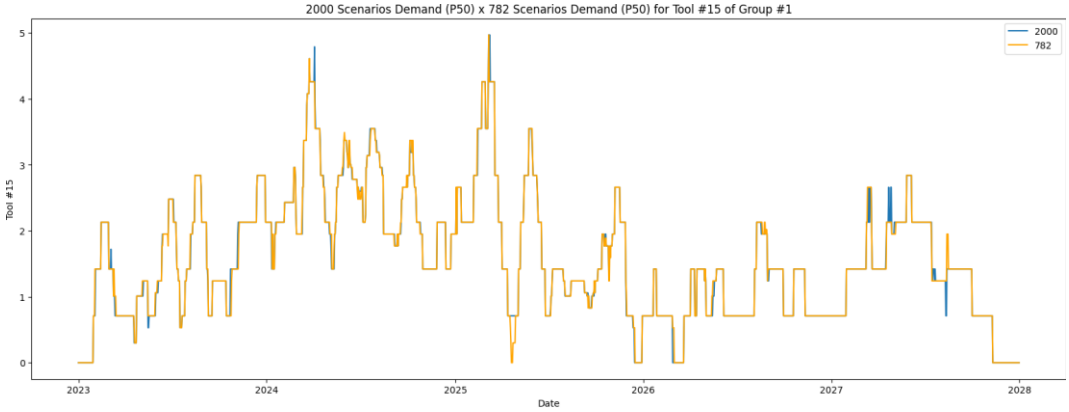

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #10 of Group #1
2000
782
Tool #10
Date
2023
2024
2025
2026
2027
2028

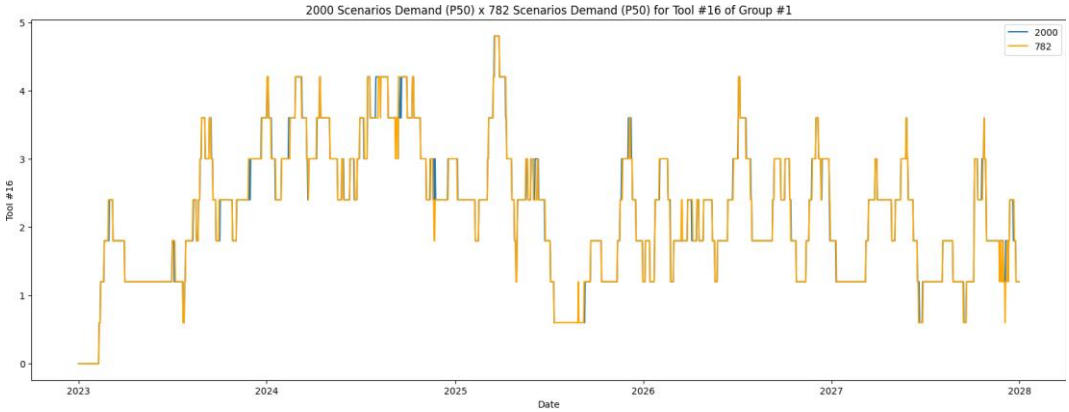

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #11 of Group #1
2000
782
Tool #11
Date
2023
2024
2025
2026
2027
2028

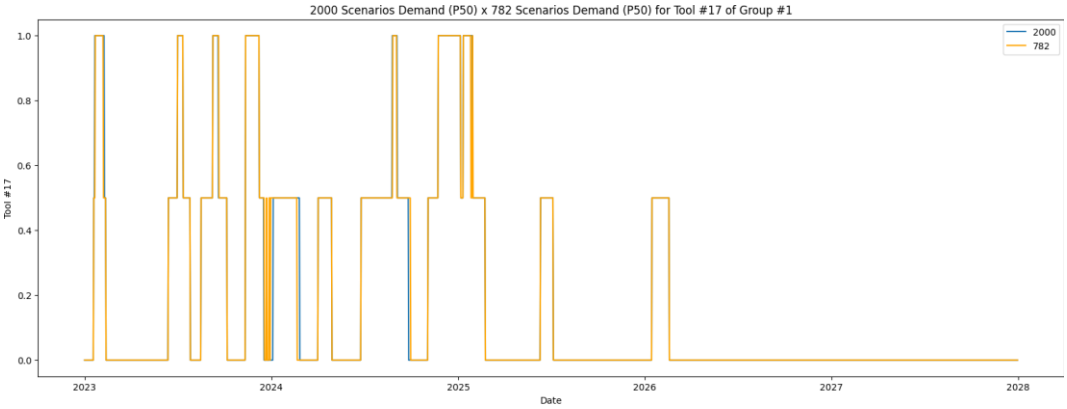

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #12 of Group #1
2000
782
Tool #12
Date
2023
2024
2025
2026
2027
2028

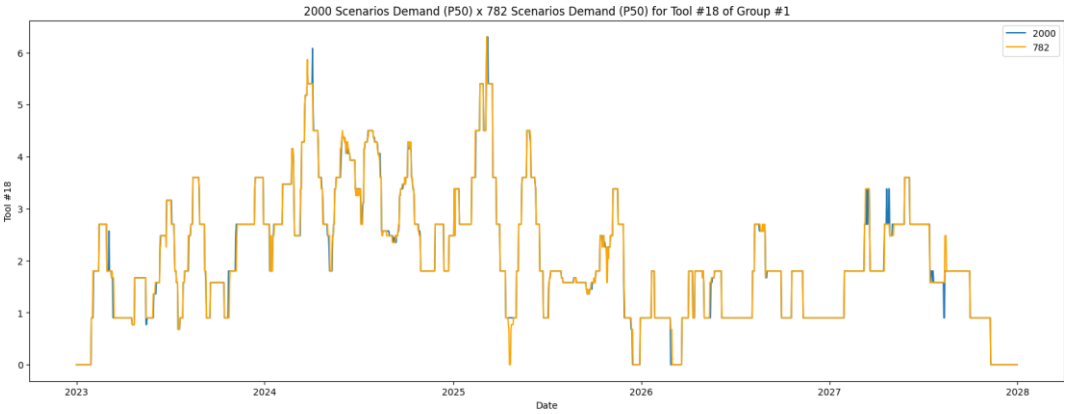

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #13 of Group #1
2000
782
Tool #13
Date
0
1
2
3
4
2023
2024
2025
2026
2027
2028

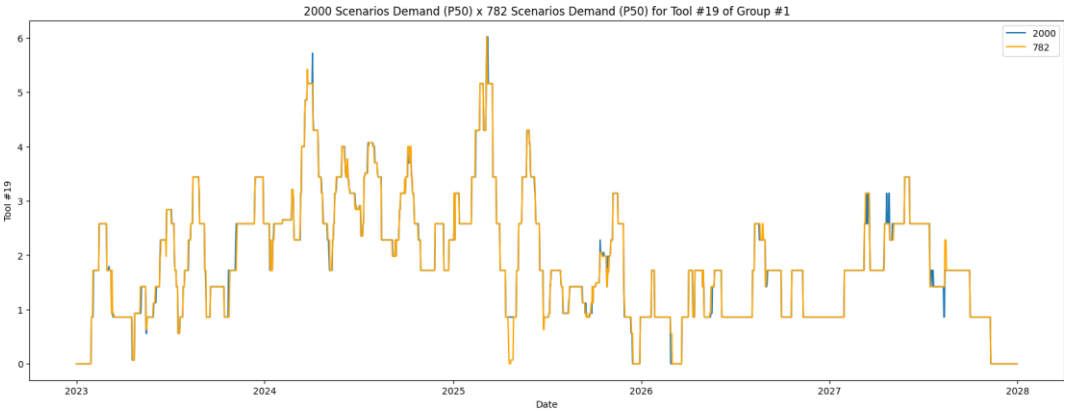

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #14 of Group #1
2000
782
Tool #14
Date
0.00
0.05
0.10
0.15
0.20
0.25
2023
2024
2025
2026
2027
2028

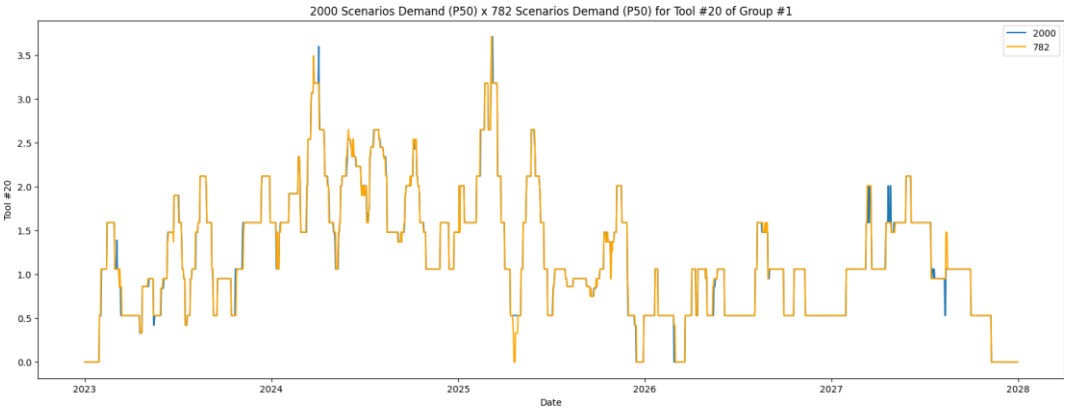

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #15 of Group #1
2000
782
Tool #15
Date
0
1
2
3
4
5
2023
2024
2025
2026
2027
2028

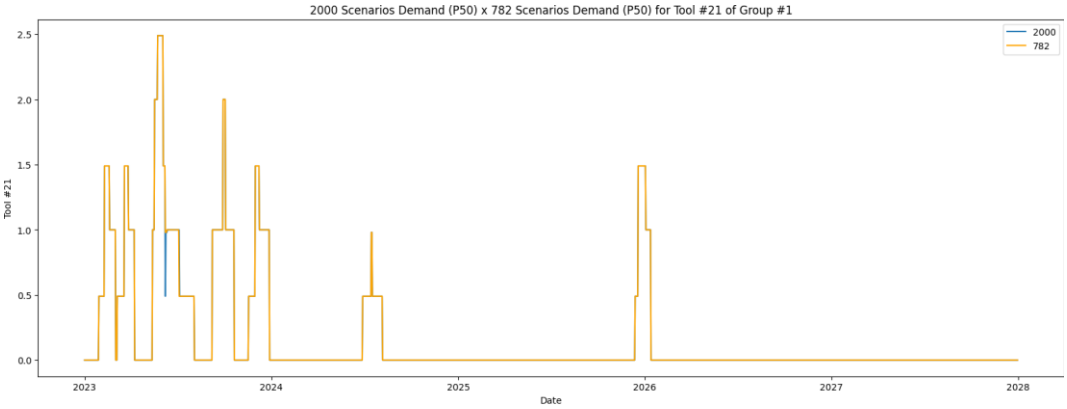

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #16 of Group #1
2000
782
Tool #16
Date
0
1
2
3
4
5
2023
2024
2025
2026
2027
2028

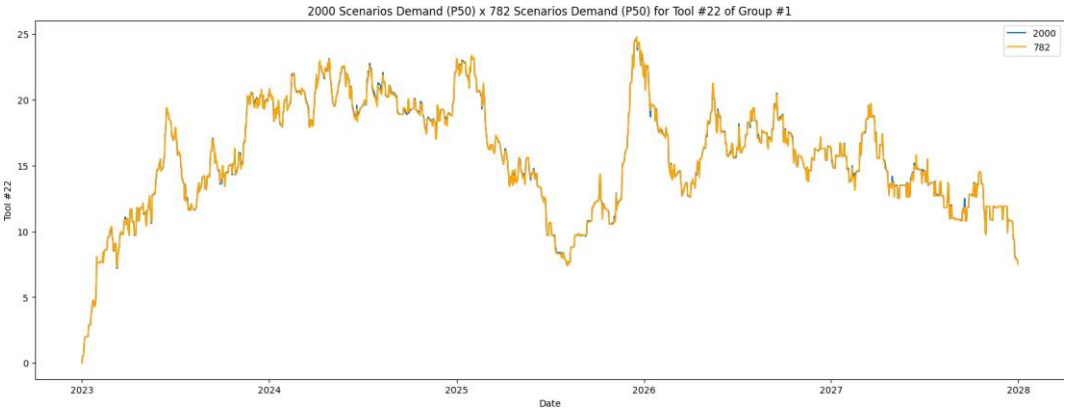

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #17 of Group #1
2000
782
Tool #17
Date
2023
2024
2025
2026
2027
2028


2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #18 of Group #1
2000
782
Tool #18
Date
2023
2024
2025
2026
2027
2028

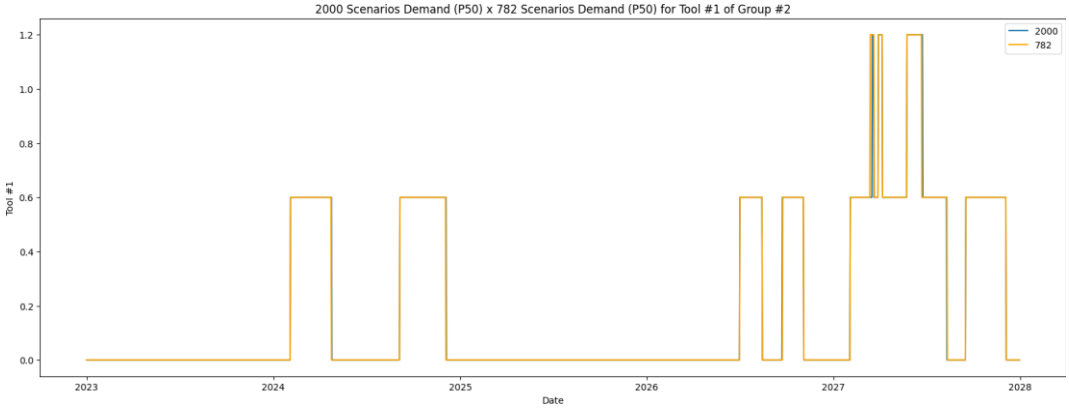

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #19 of Group #1
2000
782
Tool #19
Date
2023
2024
2025
2026
2027
2028

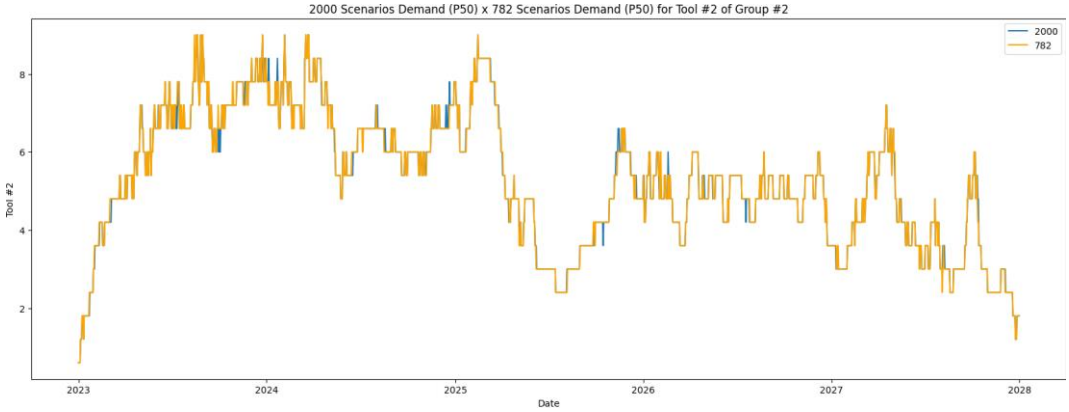

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #20 of Group #1
2000
782
Tool #20
Date
2023
2024
2025
2026
2027
2028

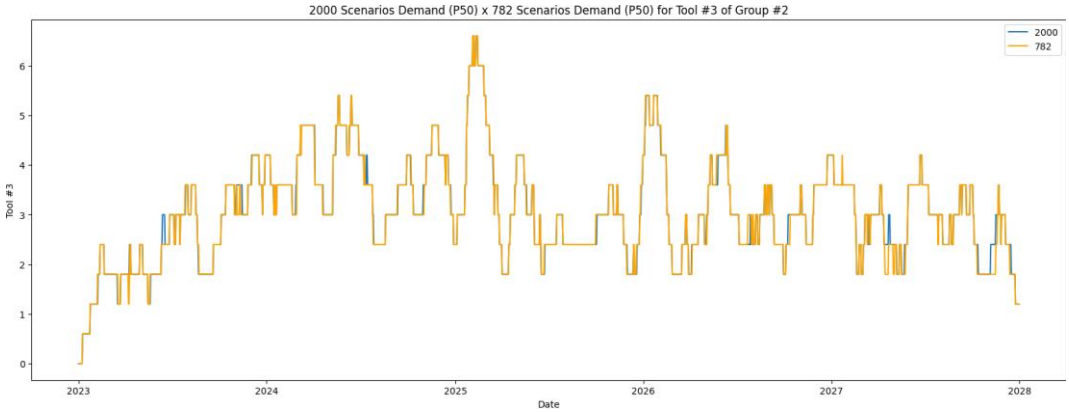

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #21 of Group #1
2000
782
Tool #21
Date
2023
2024
2025
2026
2027
2028

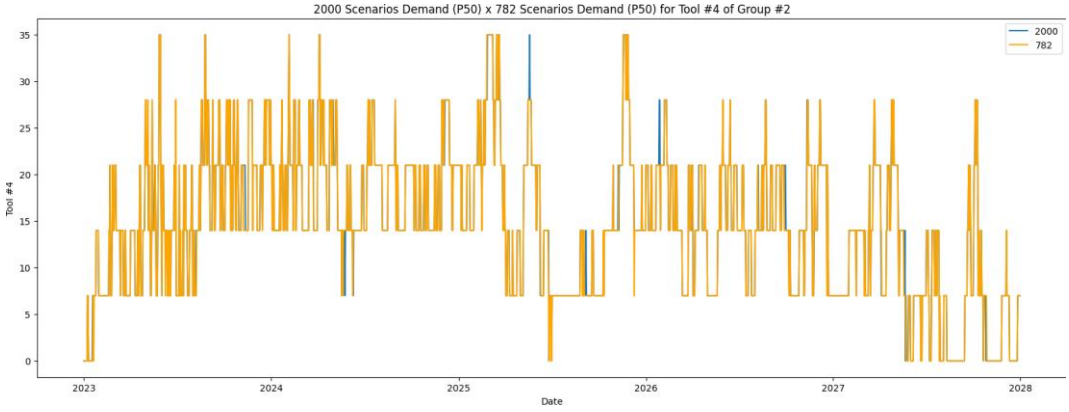

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #22 of Group #1
2000
782
Tool #22
Date
2023
2024
2025
2026
2027
2028

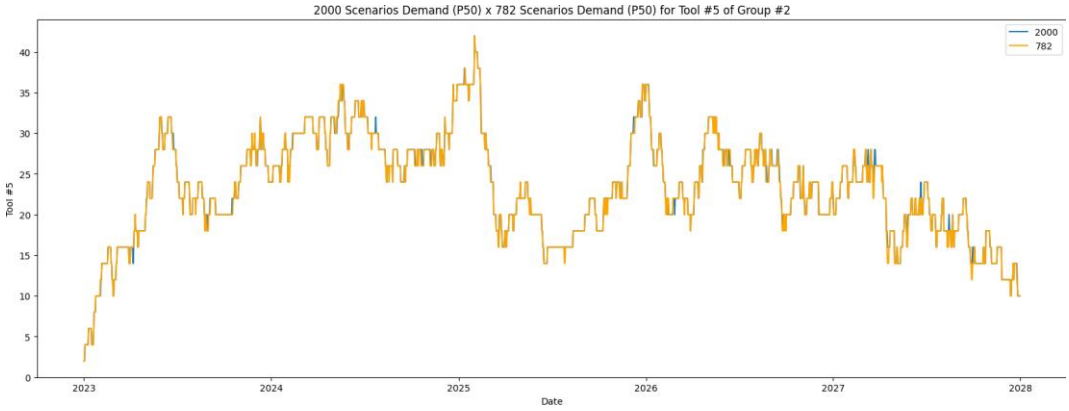
# APPENDIX III – Comparison of P50 of Demand Calculation for Original Set and 782 Scenarios for Group 2

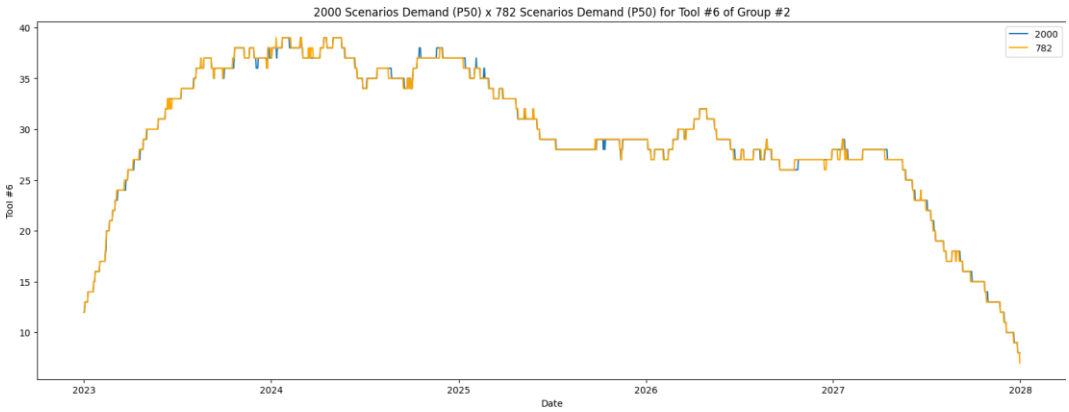


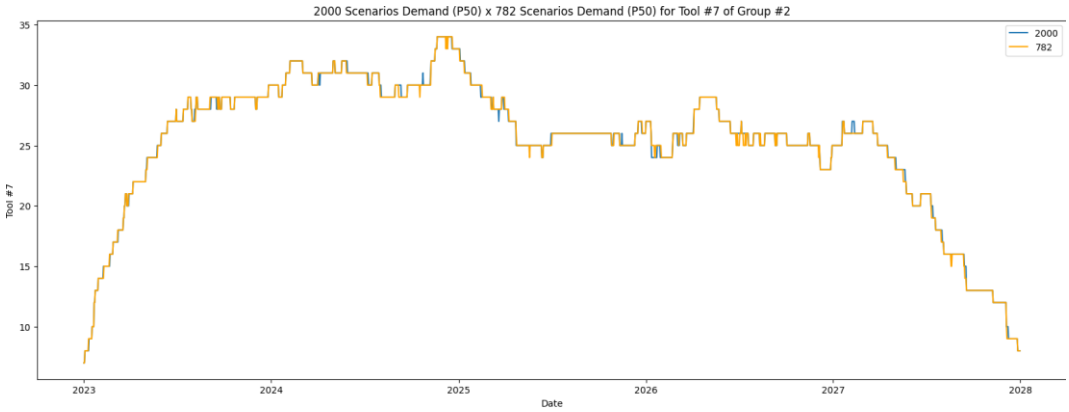



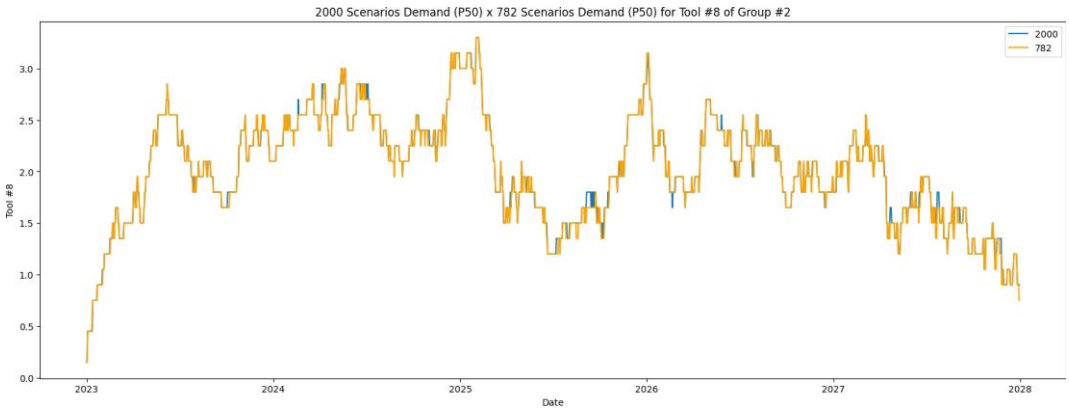



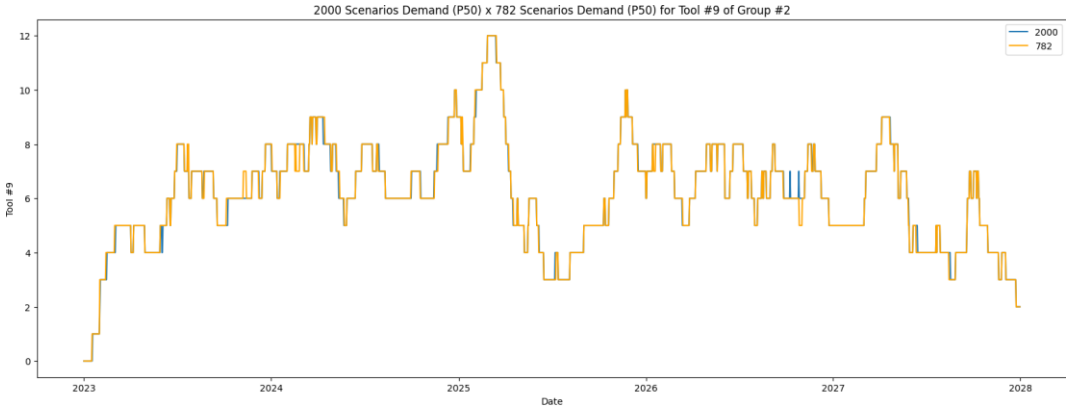



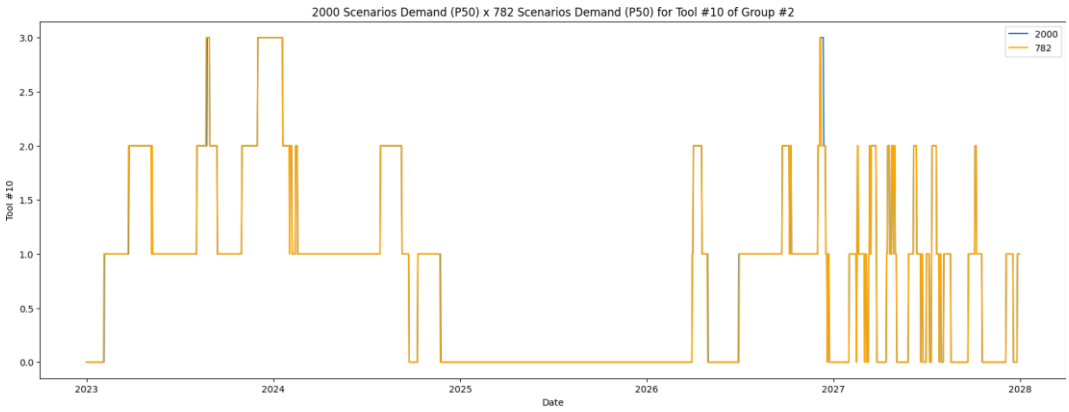

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #5 of Group #2
2000
782
Tool #5
Date
0
5
10
15
20
25
30
35
40
2023
2024
2025
2026
2027
2028


2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #6 of Group #2
2000
782
Tool #6
Date
10
15
20
25
30
35
40
2023
2024
2025
2026
2027
2028

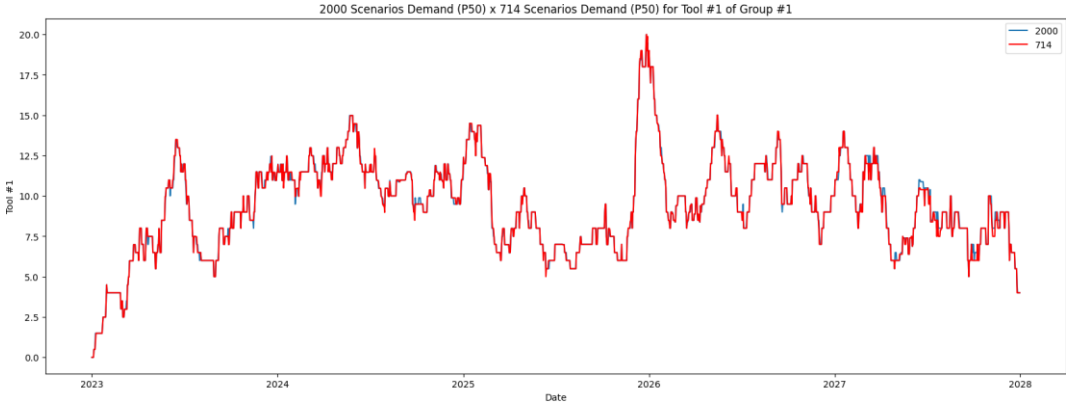

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #7 of Group #2
2000
782
Tool #7
Date
10
15
20
25
30
35
2023
2024
2025
2026
2027
2028

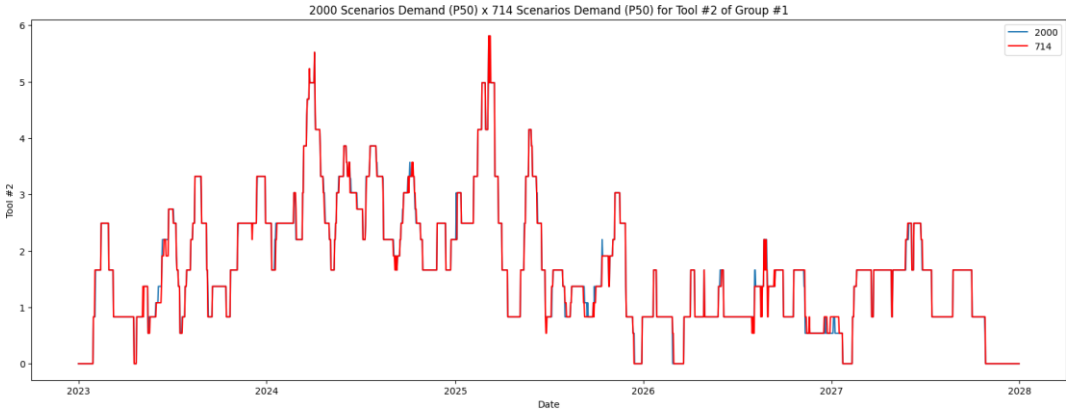

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #8 of Group #2
2000
782
Tool #8
Date
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2023
2024
2025
2026
2027
2028

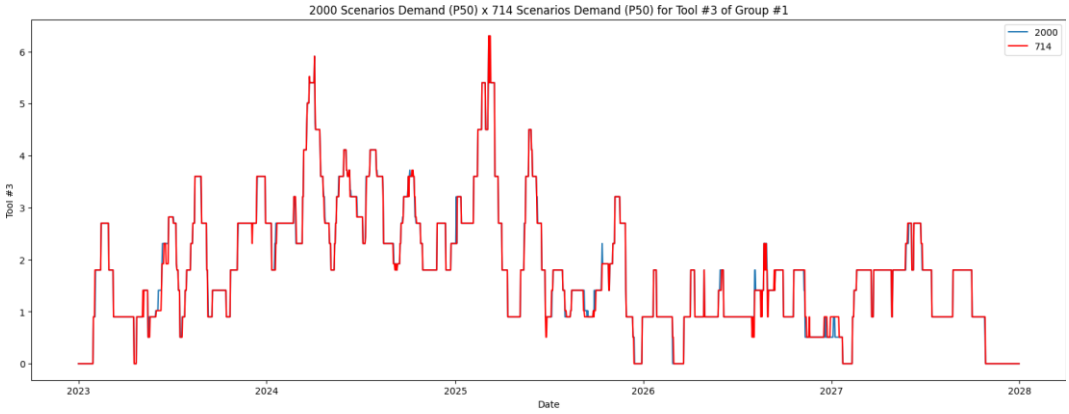

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #9 of Group #2
2000
782
Tool #9
Date
0
2
4
6
8
10
12
2023
2024
2025
2026
2027
2028

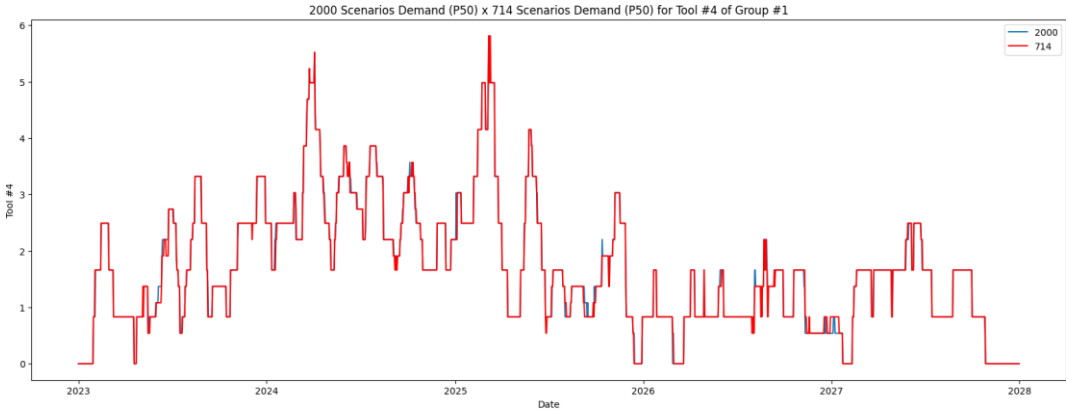

2000 Scenarios Demand (P50) x 782 Scenarios Demand (P50) for Tool #10 of Group #2
2000
782
Tool #10
Date
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2023
2024
2025
2026
2027
2028

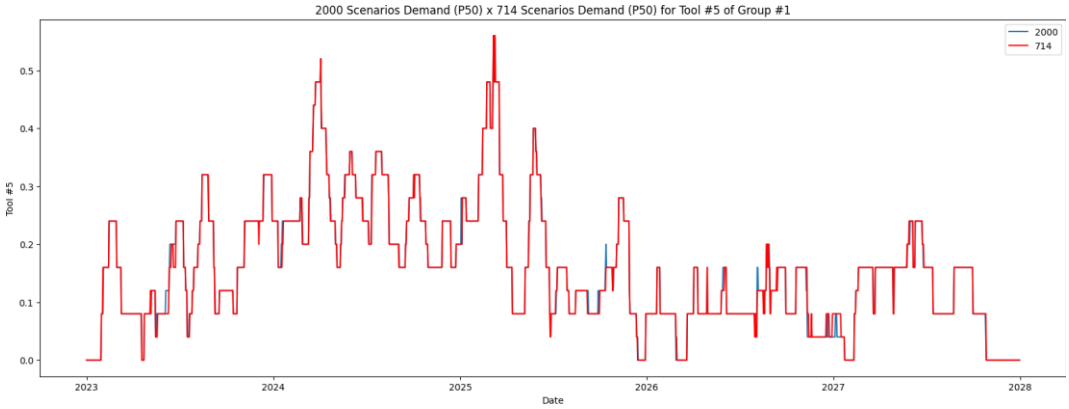
# APPENDIX IV – Comparison of P50 of Demand Calculation for Original Set and 714 Scenarios for Group 1

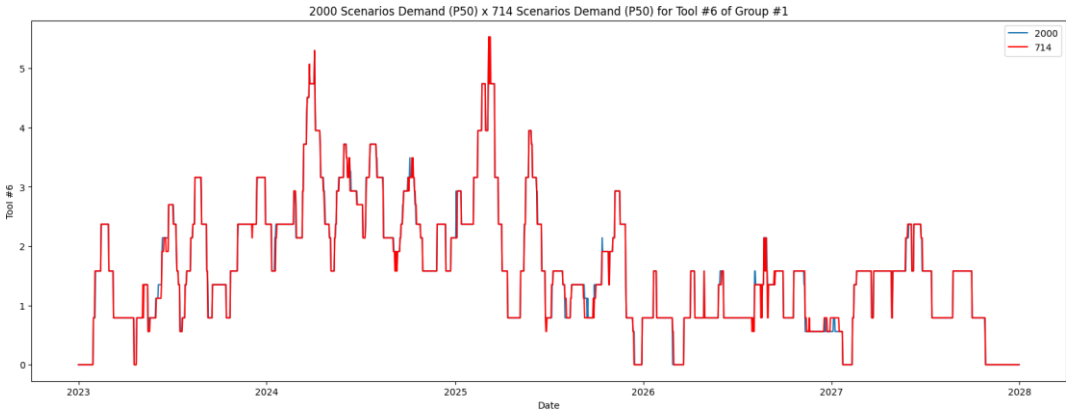


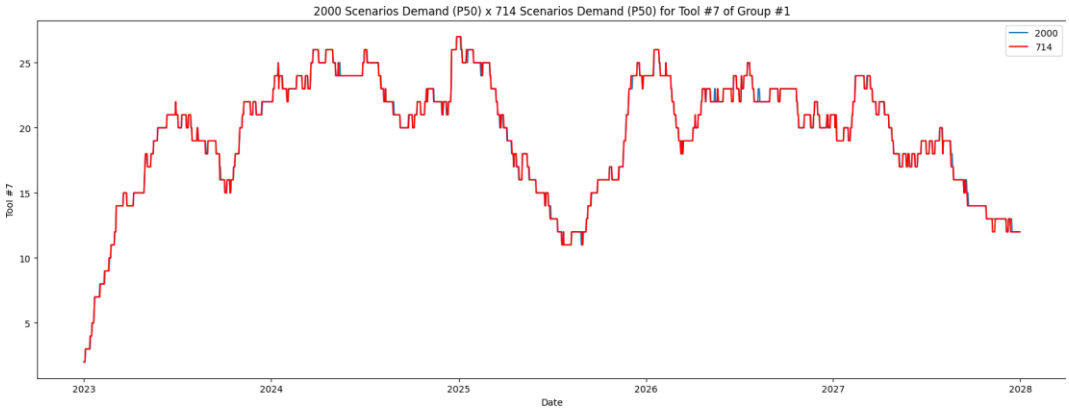



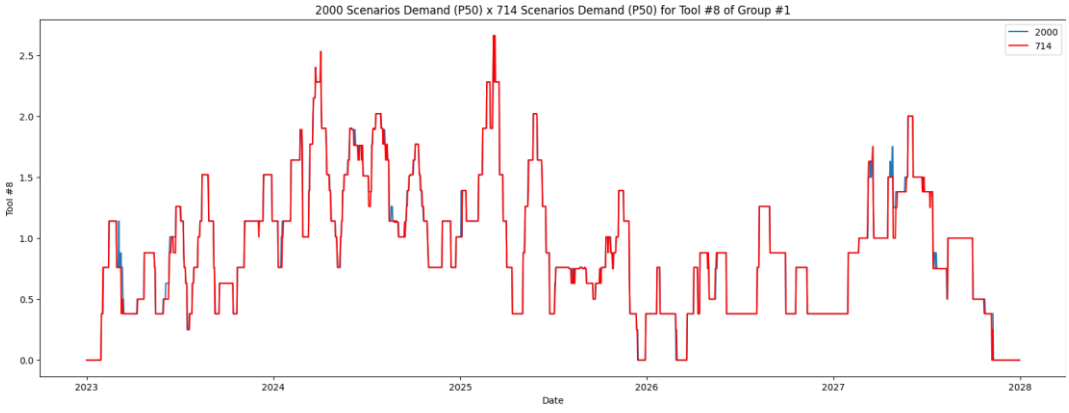



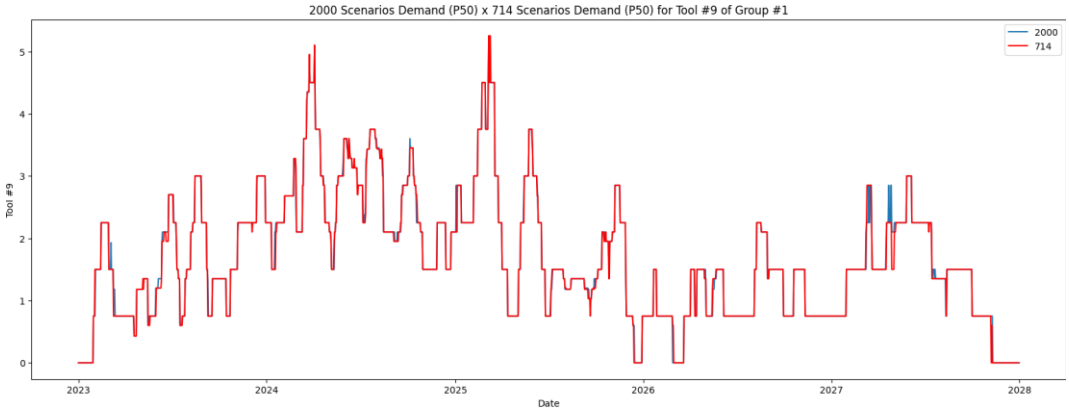



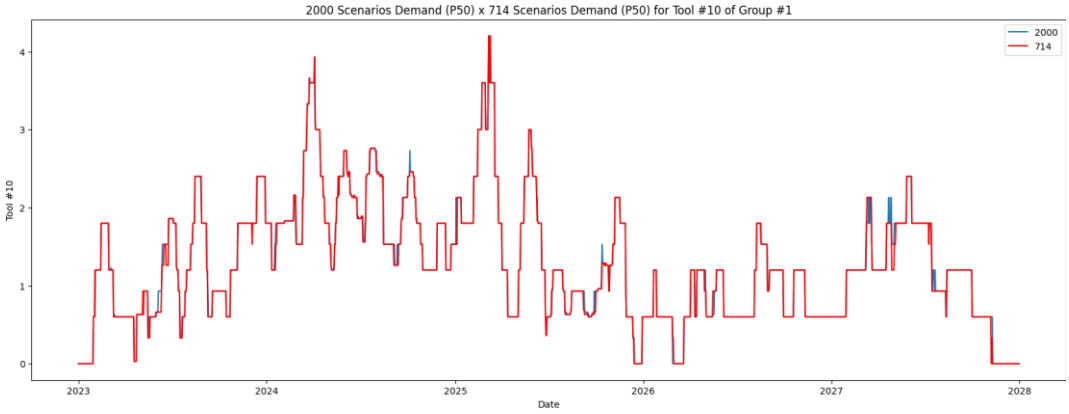

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #5 of Group #1
2000
714
Tool #5
Date
0.0
0.1
0.2
0.3
0.4
0.5
2023
2024
2025
2026
2027
2028

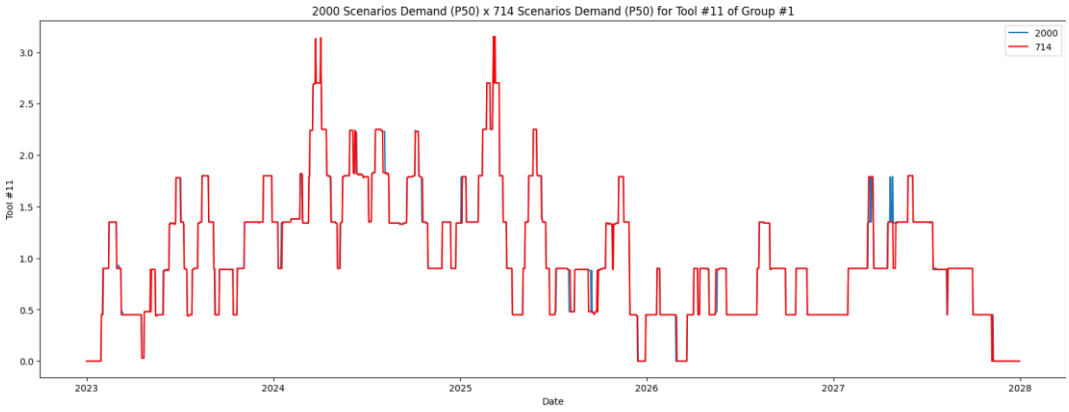

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #6 of Group #1
2000
714
Tool #6
Date
0
1
2
3
4
5
2023
2024
2025
2026
2027
2028

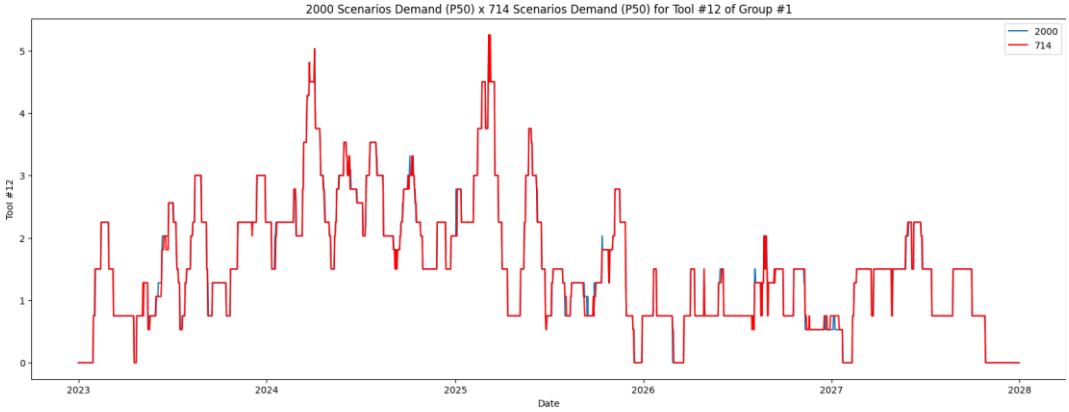

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #7 of Group #1
2000
714
Tool #7
Date
5
10
15
20
25
2023
2024
2025
2026
2027
2028

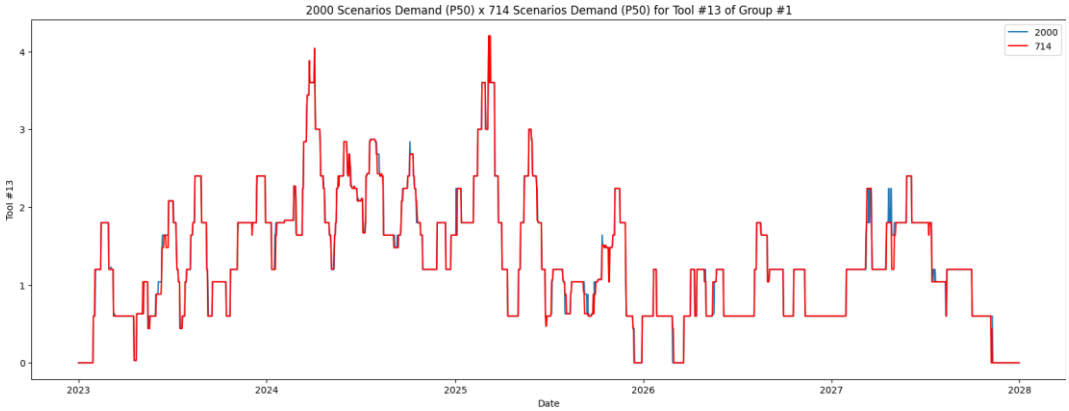

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #8 of Group #1
2000
714
Tool #8
Date
0.0
0.5
1.0
1.5
2.0
2.5
2023
2024
2025
2026
2027
2028

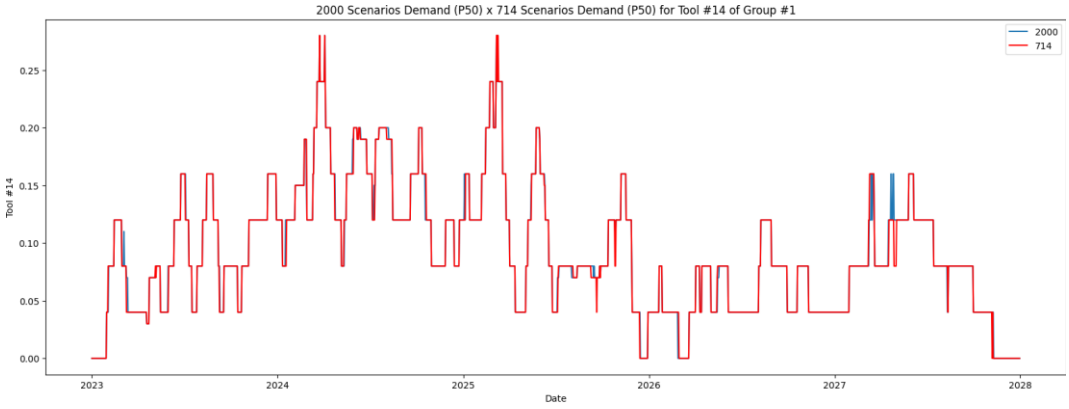

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #9 of Group #1
2000
714
Tool #9
Date
2023
2024
2025
2026
2027
2028

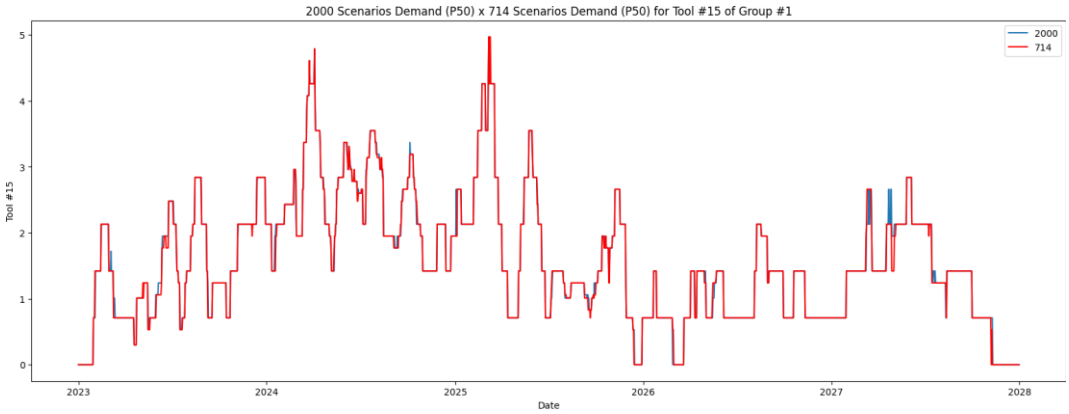

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #10 of Group #1
2000
714
Tool #10
Date
2023
2024
2025
2026
2027
2028

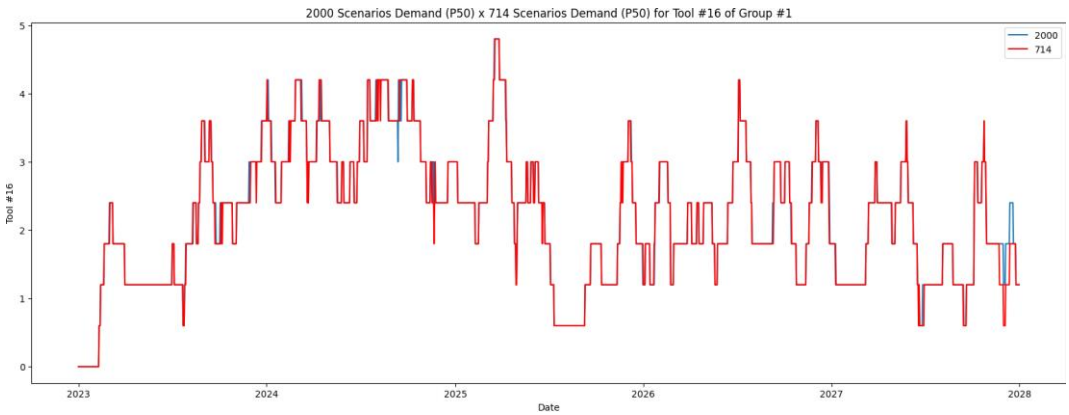

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #11 of Group #1
2000
714
Tool #11
Date
2023
2024
2025
2026
2027
2028

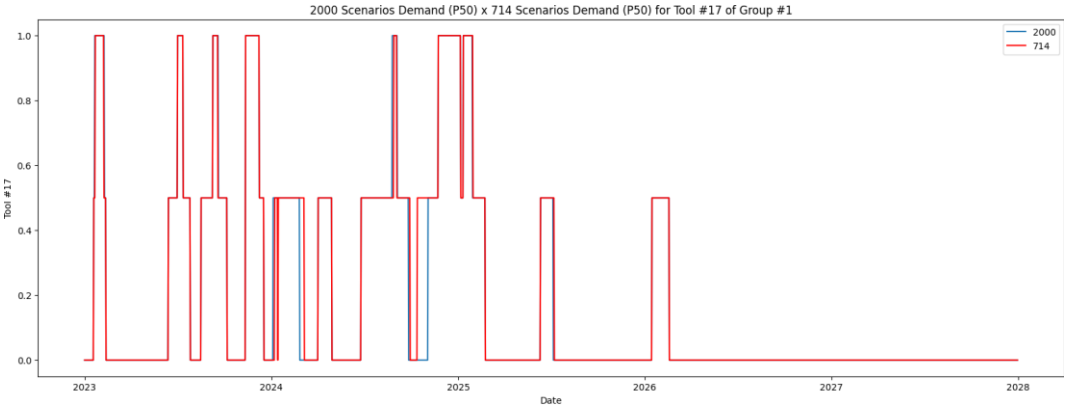

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #12 of Group #1
2000
714
Tool #12
Date
2023
2024
2025
2026
2027
2028

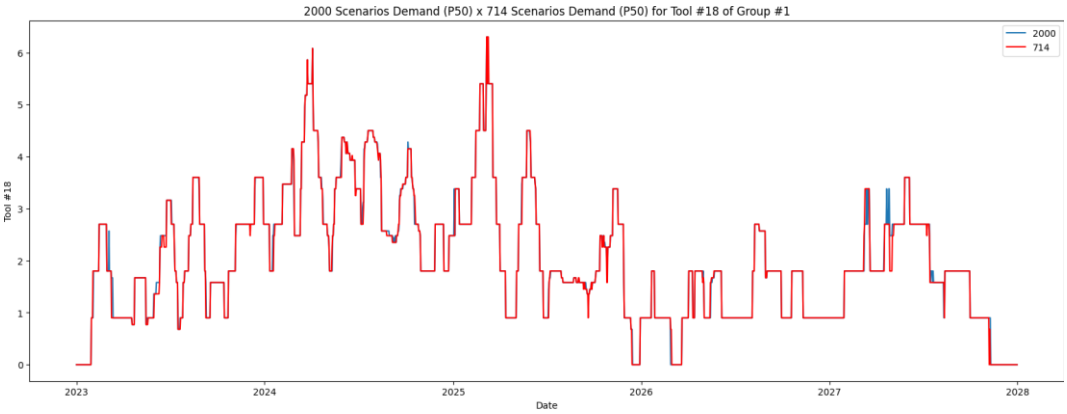

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #13 of Group #1
2000
714
Tool #13
Date
2023
2024
2025
2026
2027
2028

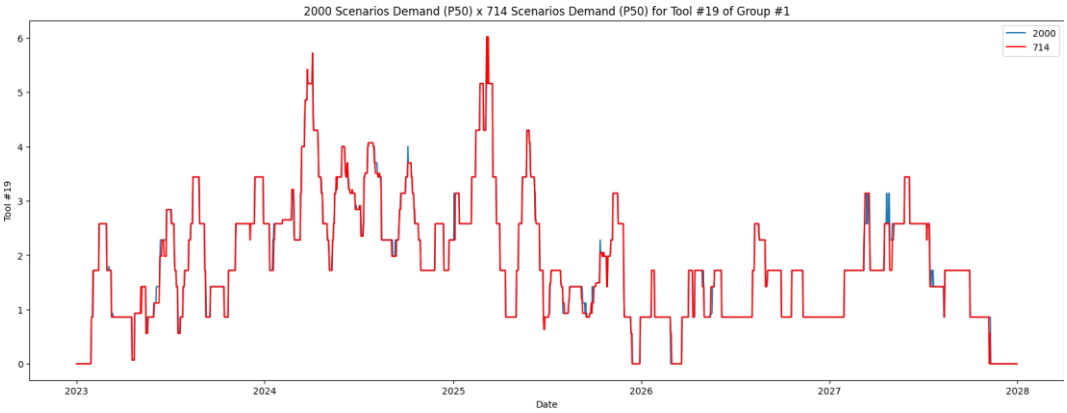

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #14 of Group #1
2000
714
Tool #14
Date
2023
2024
2025
2026
2027
2028

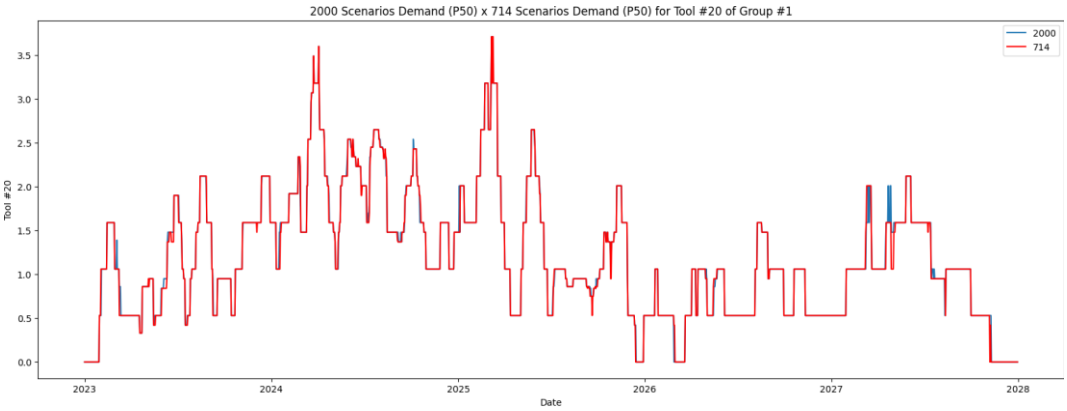

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #15 of Group #1
2000
714
Tool #15
Date
2023
2024
2025
2026
2027
2028

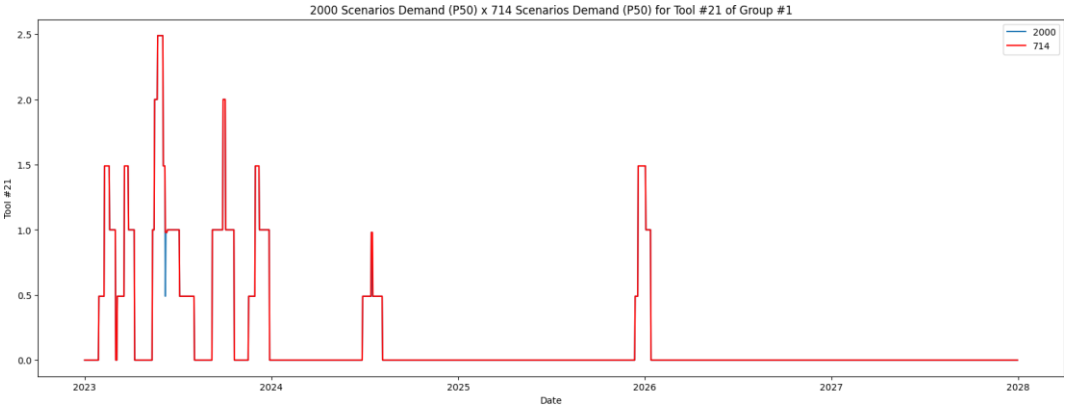

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #16 of Group #1
2000
714
Tool #16
Date
2023
2024
2025
2026
2027
2028

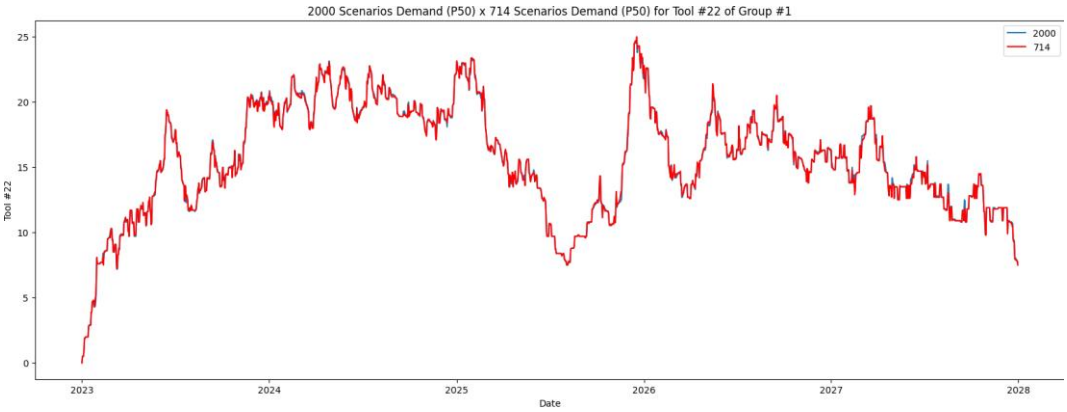

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #17 of Group #1
2000
714
Tool #17
Date


2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #18 of Group #1
2000
714
Tool #18
Date

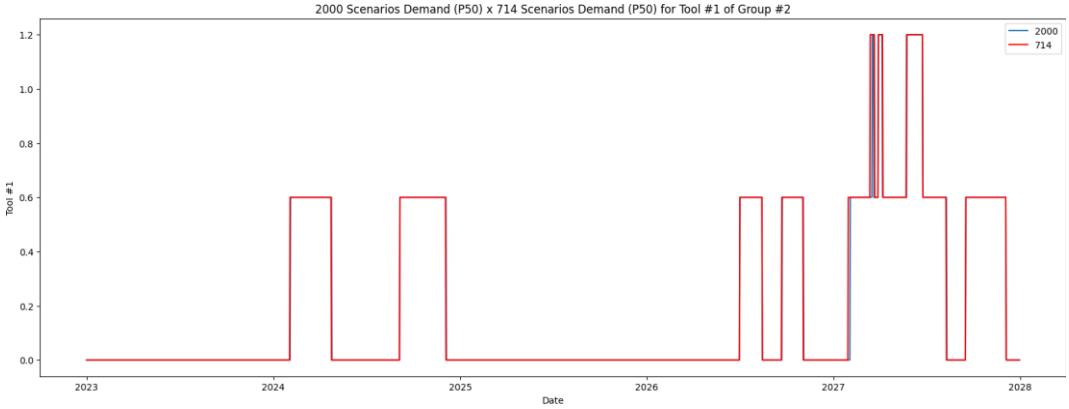

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #19 of Group #1
2000
714
Tool #19
Date

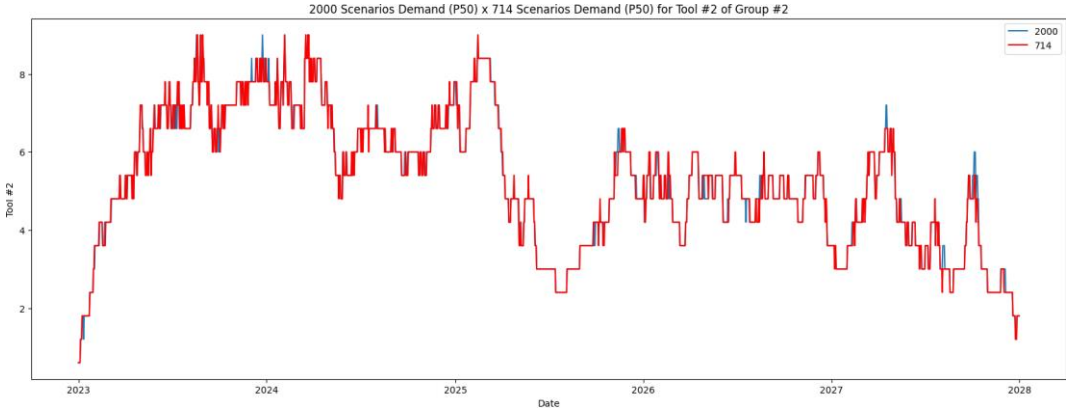

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #20 of Group #1
2000
714
Tool #20
Date

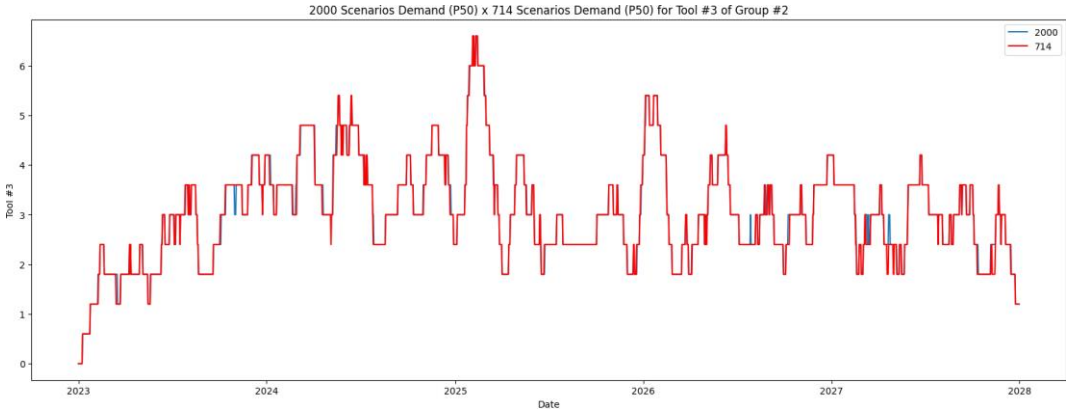

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #21 of Group #1
2000
714
Tool #21
Date
2023
2024
2025
2026
2027
2028

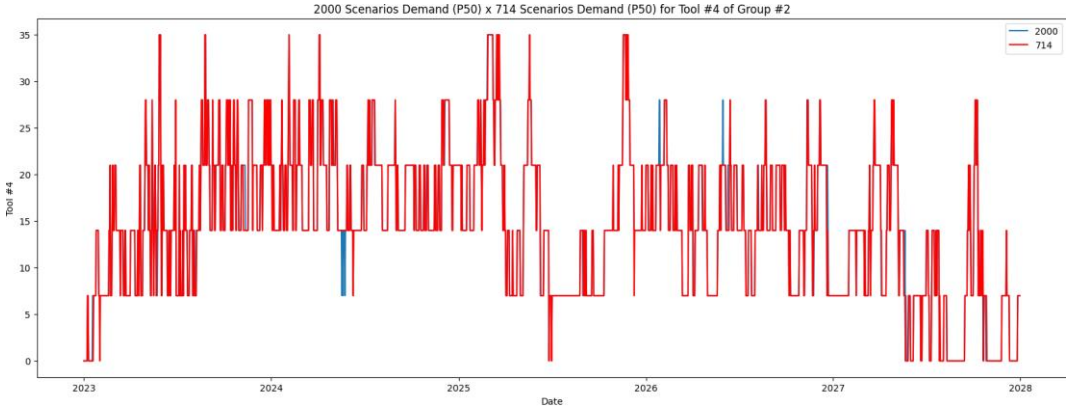

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #22 of Group #1
2000
714
Tool #22
Date
2023
2024
2025
2026
2027
2028

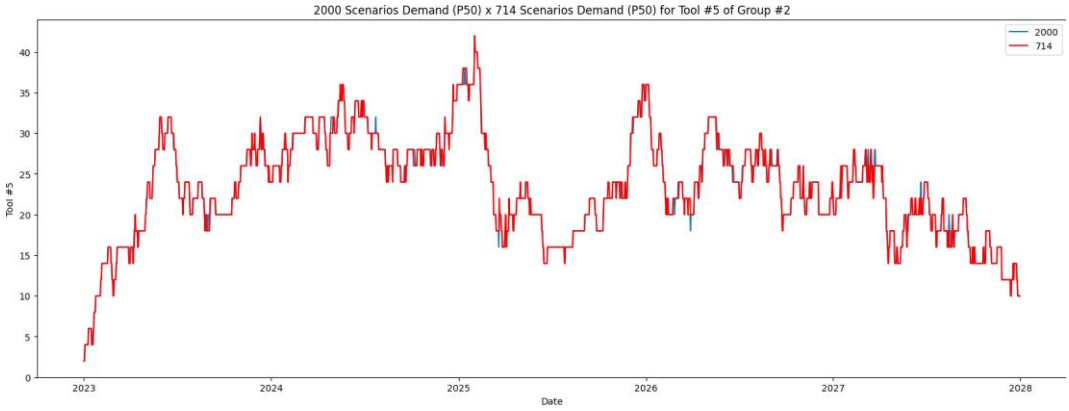
# APPENDIX V – Comparison of P50 of Demand Calculation for Original Set and 714 Scenarios for Group 2

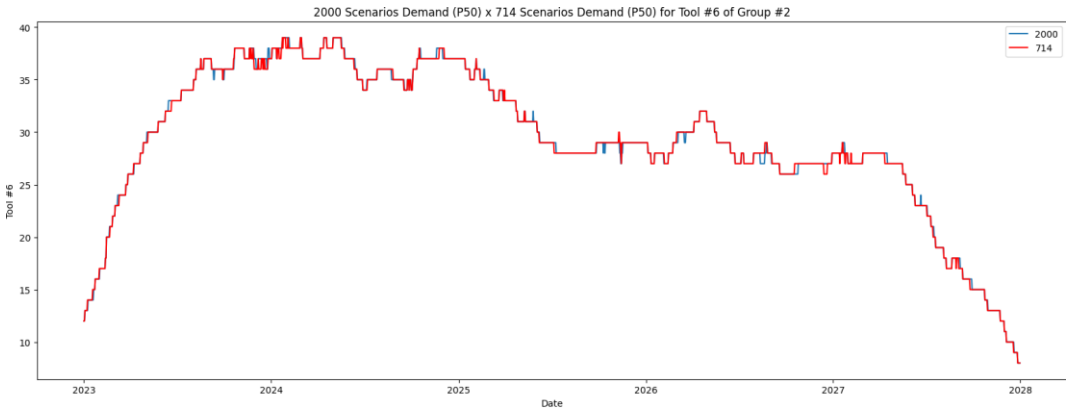


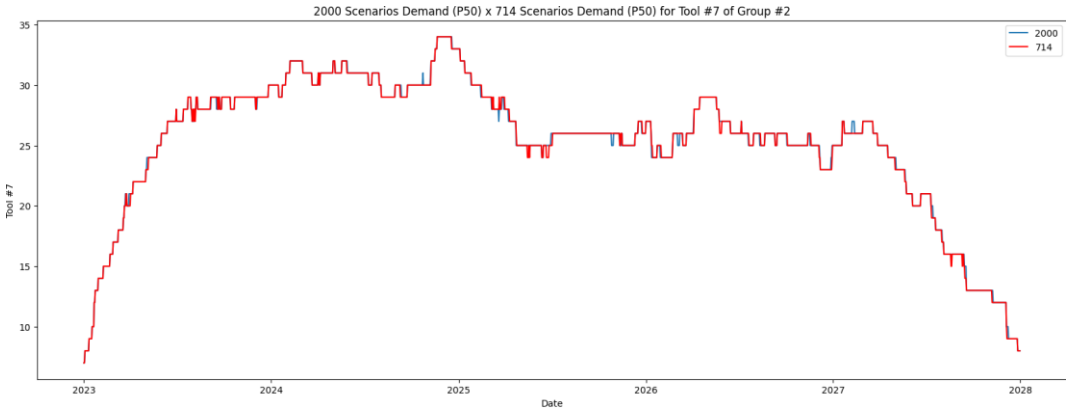



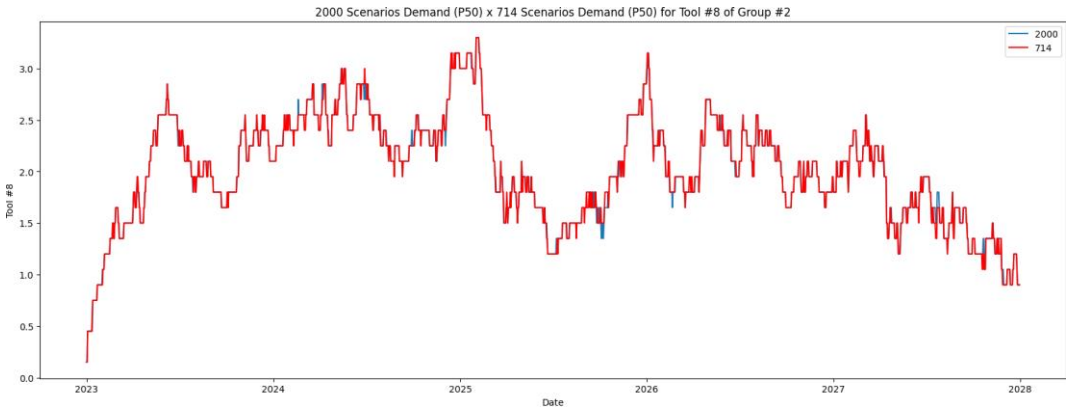



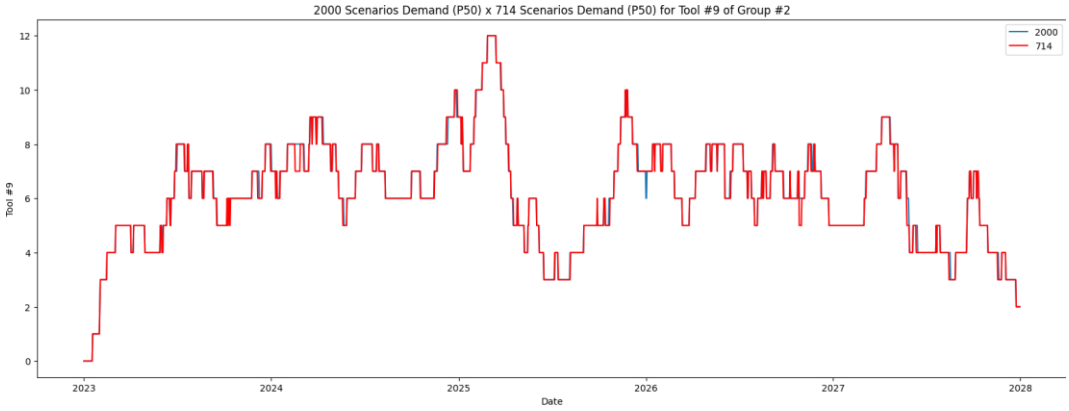



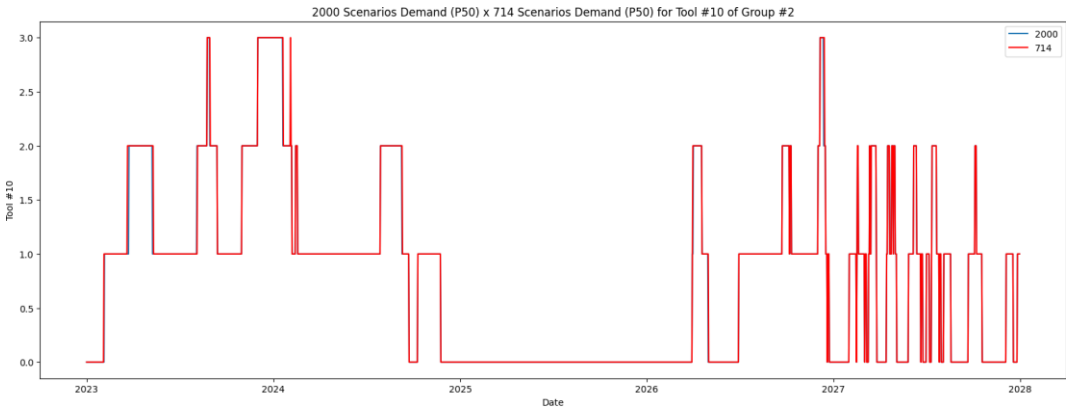

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #5 of Group #2
2000
714
Tool #5
Date
0
5
10
15
20
25
30
35
40
2023
2024
2025
2026
2027
2028


2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #6 of Group #2
2000
714
Tool #6
Date
10
15
20
25
30
35
40
2023
2024
2025
2026
2027
2028

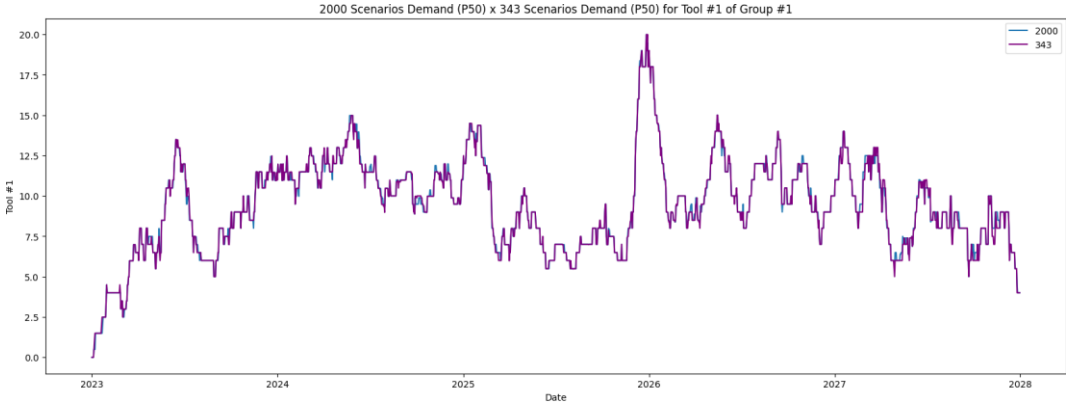

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #7 of Group #2
2000
714
Tool #7
Date
10
15
20
25
30
35
2023
2024
2025
2026
2027
2028

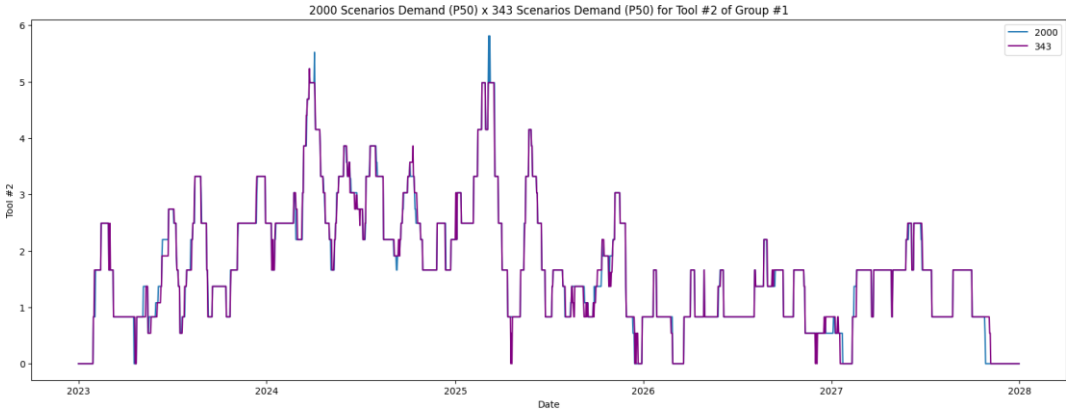

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #8 of Group #2
2000
714
Tool #8
Date
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2023
2024
2025
2026
2027
2028

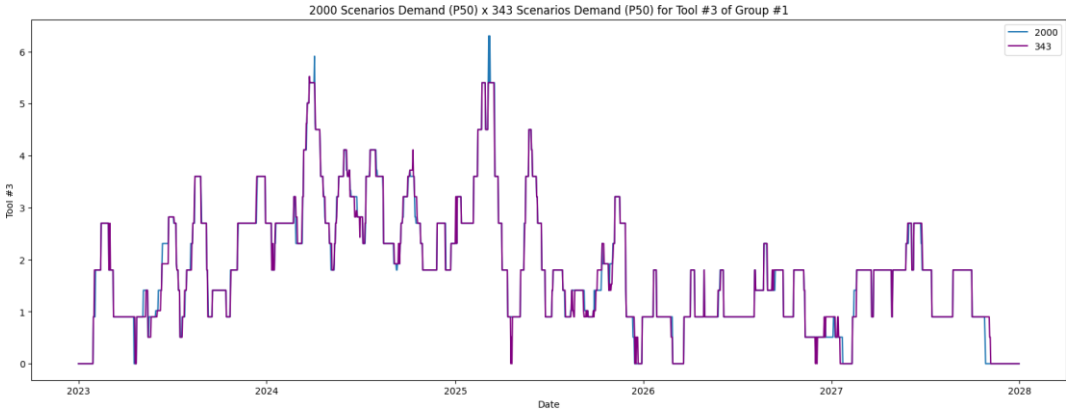

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #9 of Group #2
2000
714
Tool #9
Date
0
2
4
6
8
10
12
2023
2024
2025
2026
2027
2028

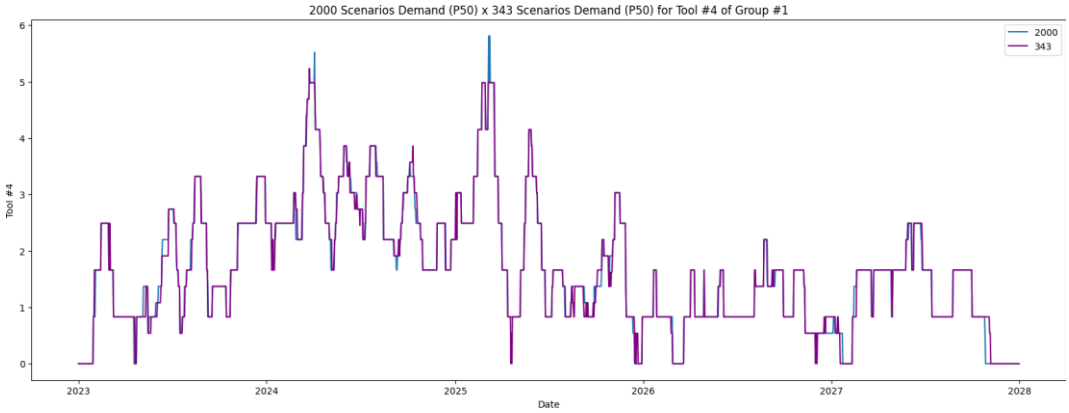

2000 Scenarios Demand (P50) x 714 Scenarios Demand (P50) for Tool #10 of Group #2
2000
714
Tool #10
Date
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2023
2024
2025
2026
2027
2028

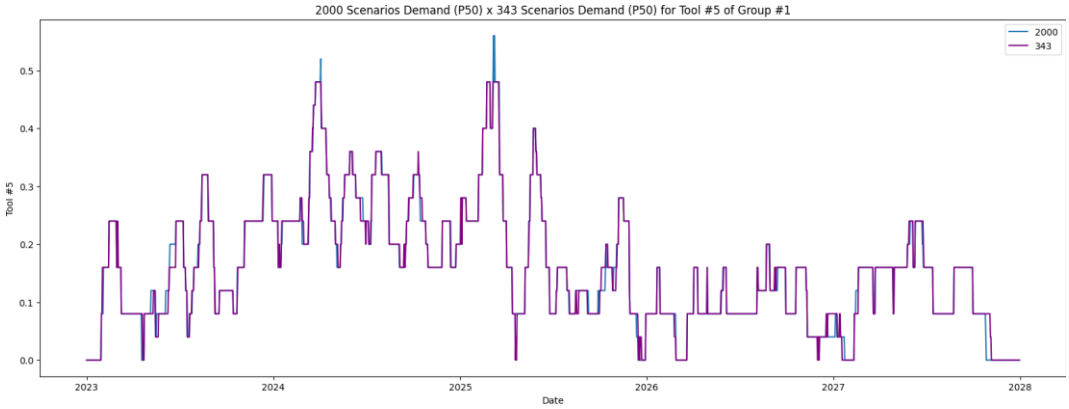
# APPENDIX VI – Comparison of P50 of Demand Calculation for Original Set and 343 Scenarios for Group 1

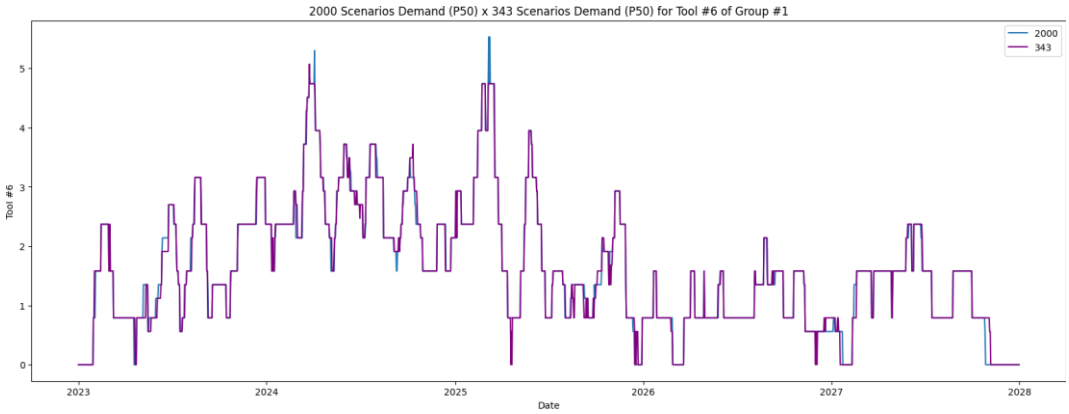


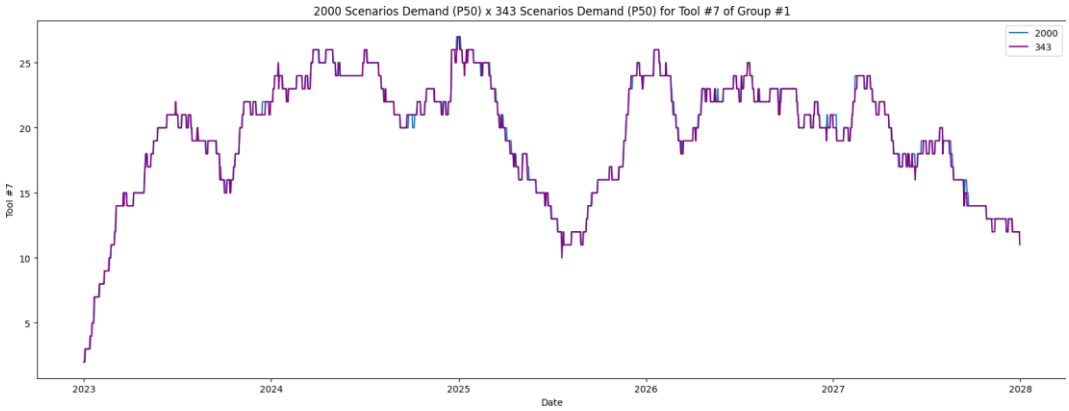



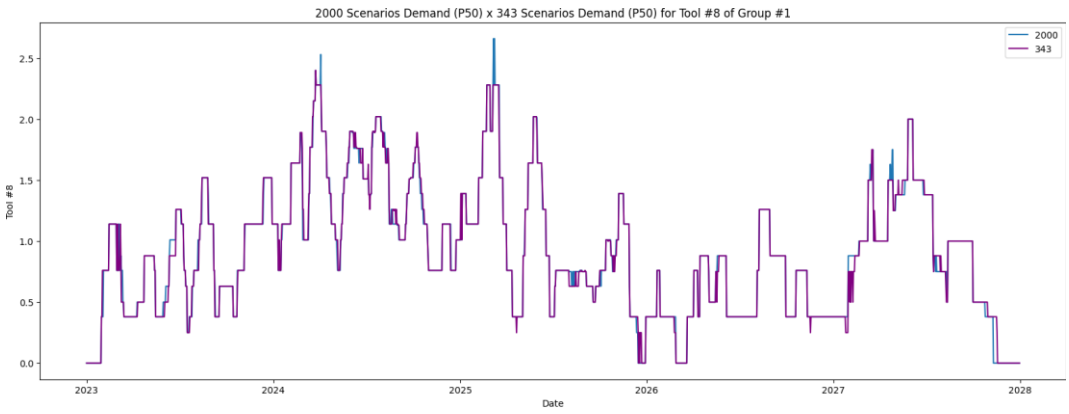



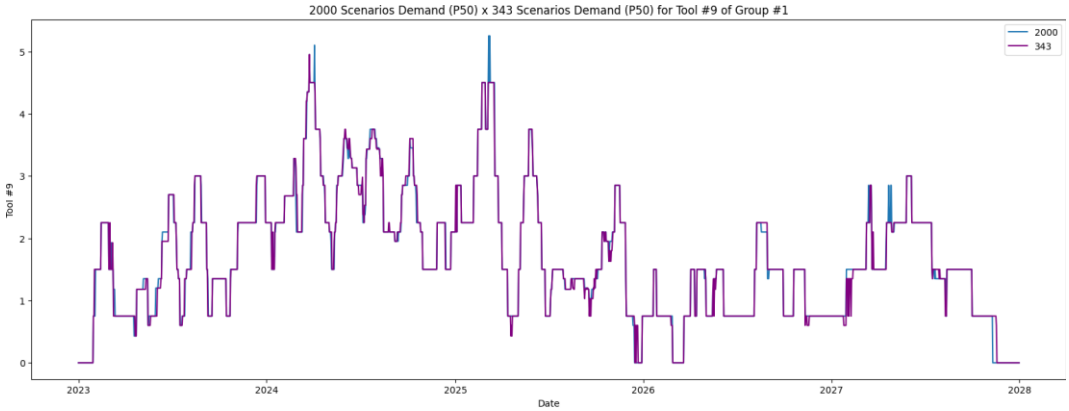



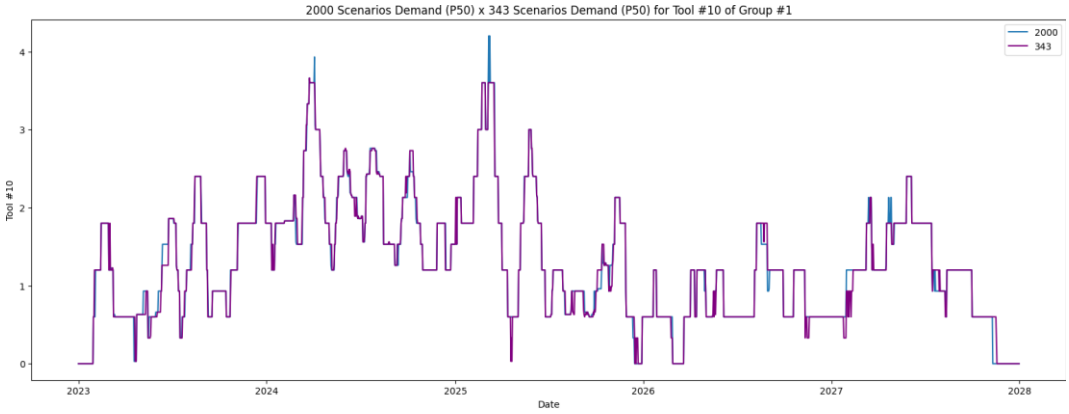

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #5 of Group #1
2000
343
Tool #5
Date

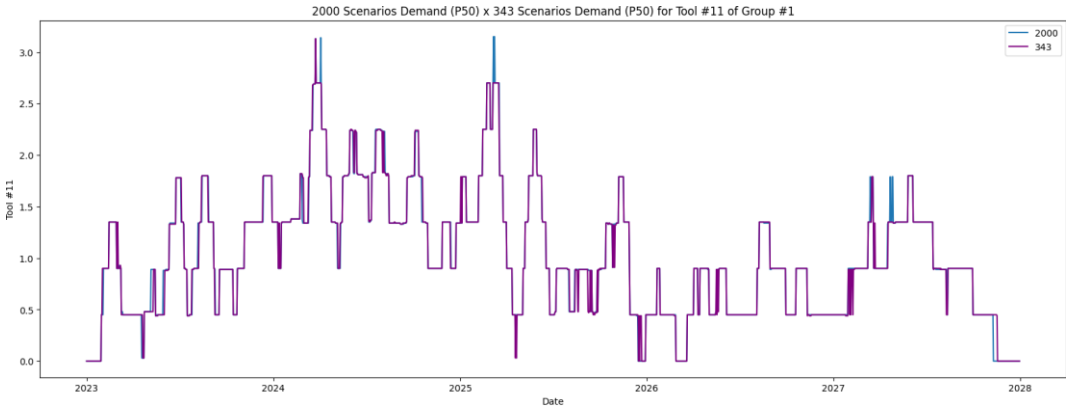

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #6 of Group #1
2000
343
Tool #6
Date

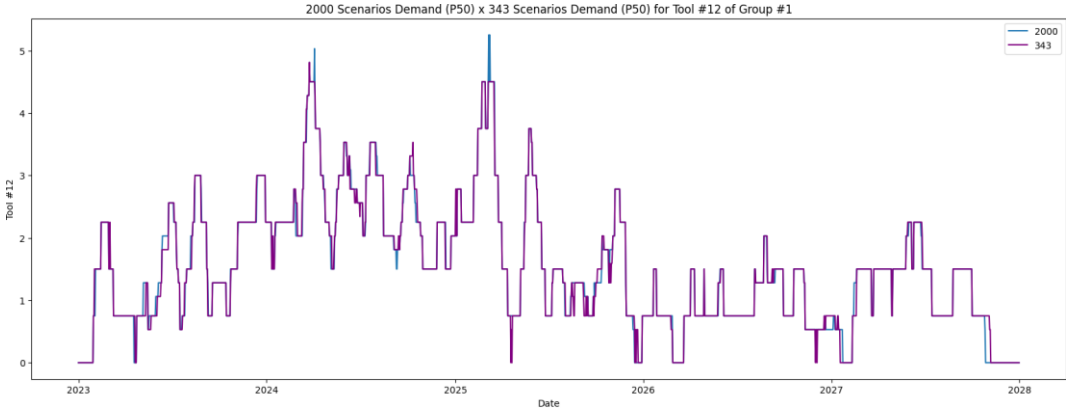

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #7 of Group #1
2000
343
Tool #7
Date

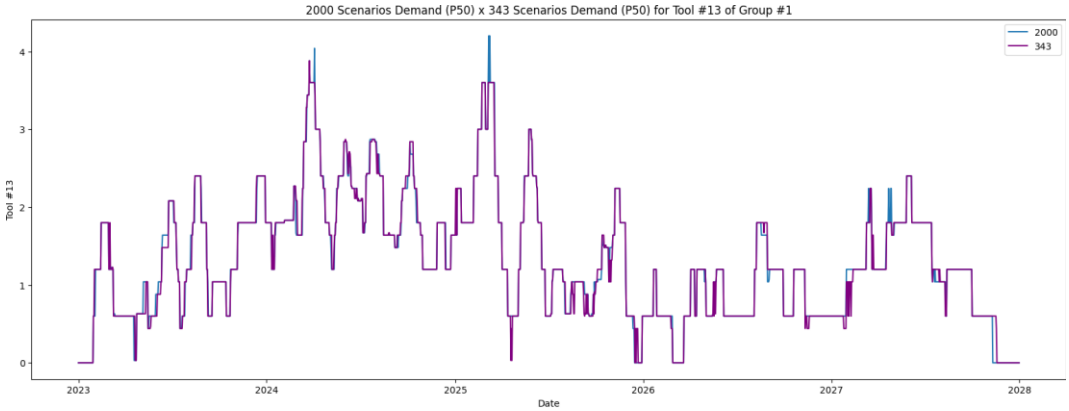

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #8 of Group #1
2000
343
Tool #8
Date

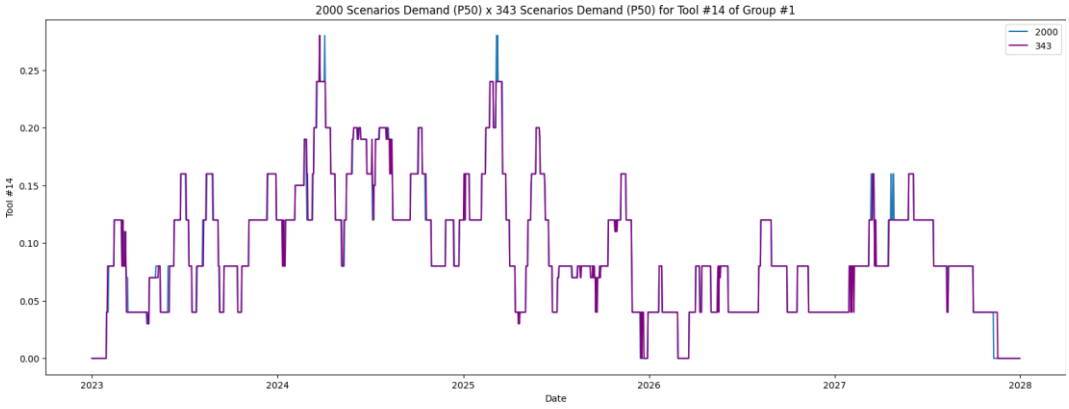

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #9 of Group #1
2000
343
Tool #9
Date
0
1
2
3
4
5
2023
2024
2025
2026
2027
2028

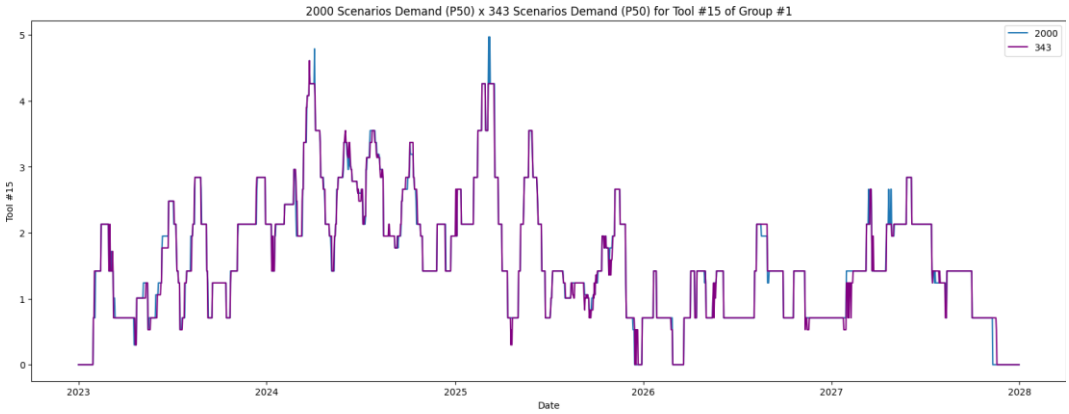

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #10 of Group #1
2000
343
Tool #10
Date
0
1
2
3
4
2023
2024
2025
2026
2027
2028

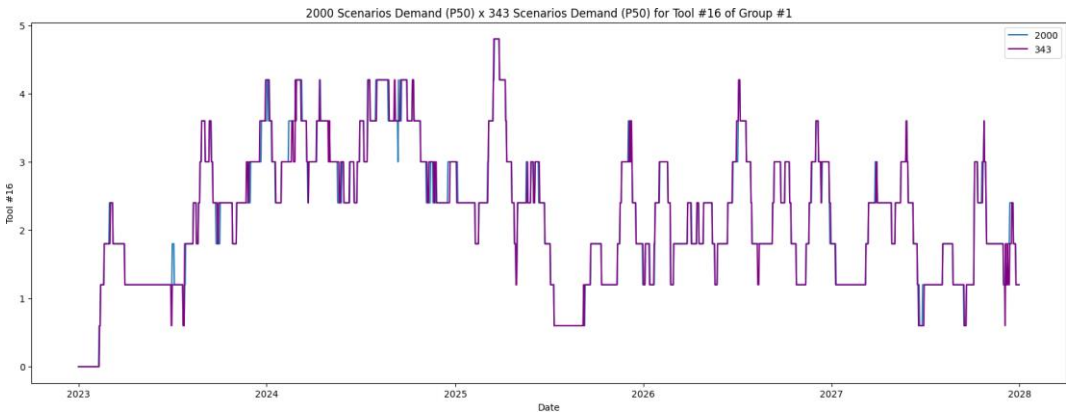

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #11 of Group #1
2000
343
Tool #11
Date
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2023
2024
2025
2026
2027
2028

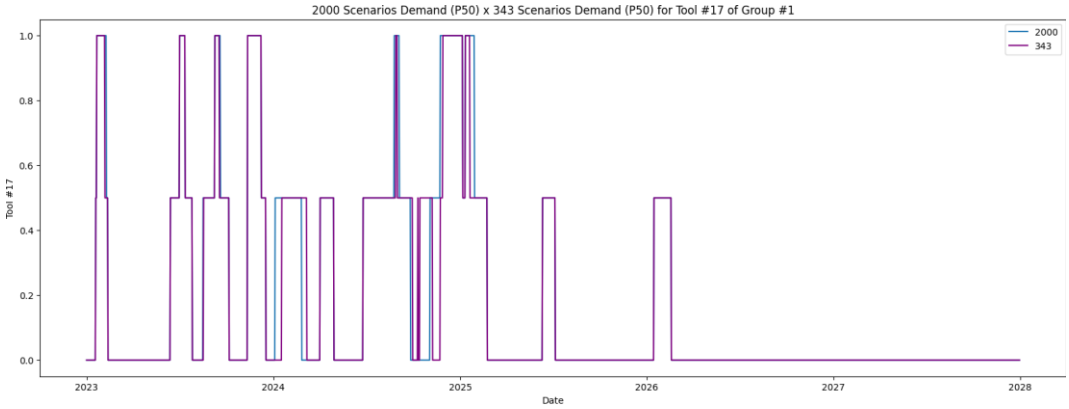

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #12 of Group #1
2000
343
Tool #12
Date
0
1
2
3
4
5
2023
2024
2025
2026
2027
2028

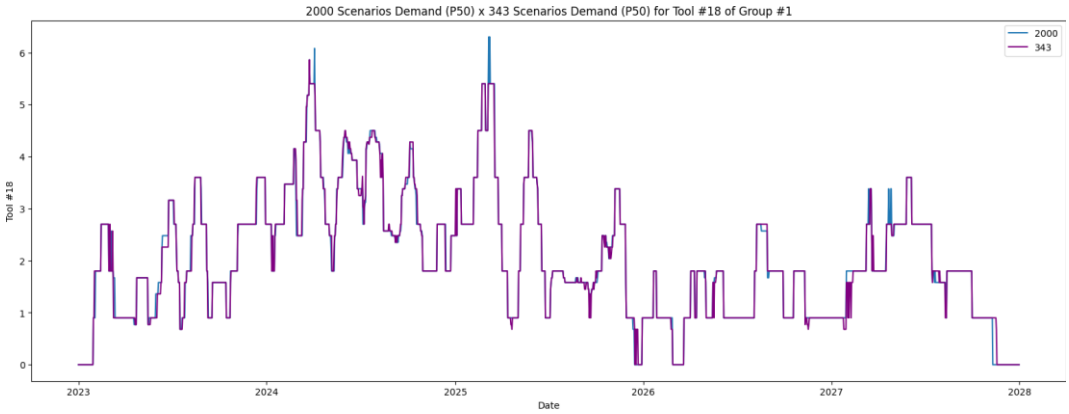

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #13 of Group #1
2000
343
Tool #13
Date
2023
2024
2025
2026
2027
2028

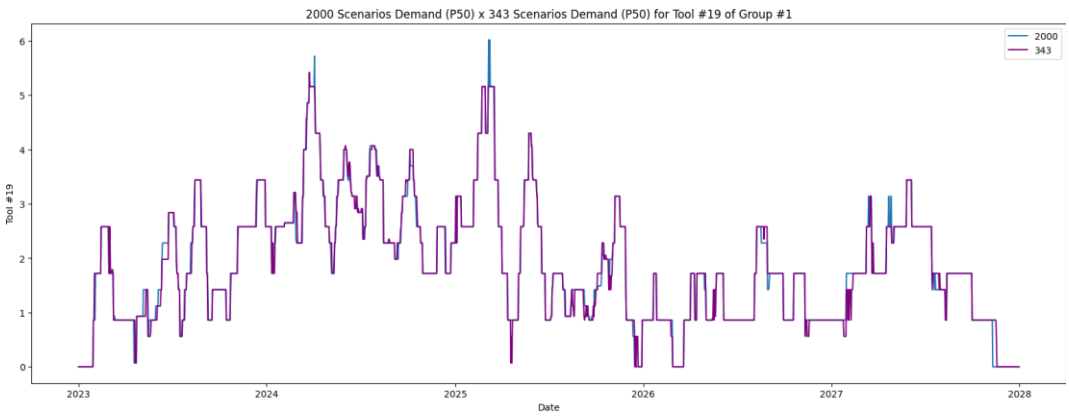

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #14 of Group #1
2000
343
Tool #14
Date
2023
2024
2025
2026
2027
2028

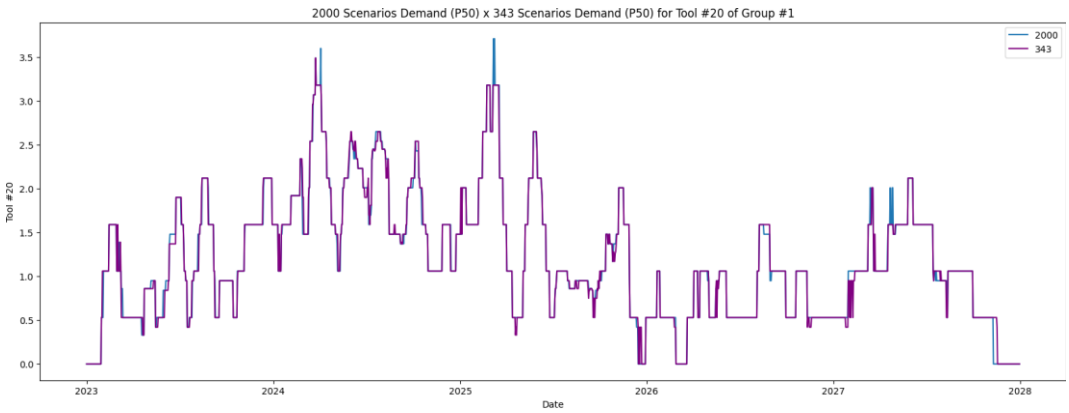

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #15 of Group #1
2000
343
Tool #15
Date
2023
2024
2025
2026
2027
2028

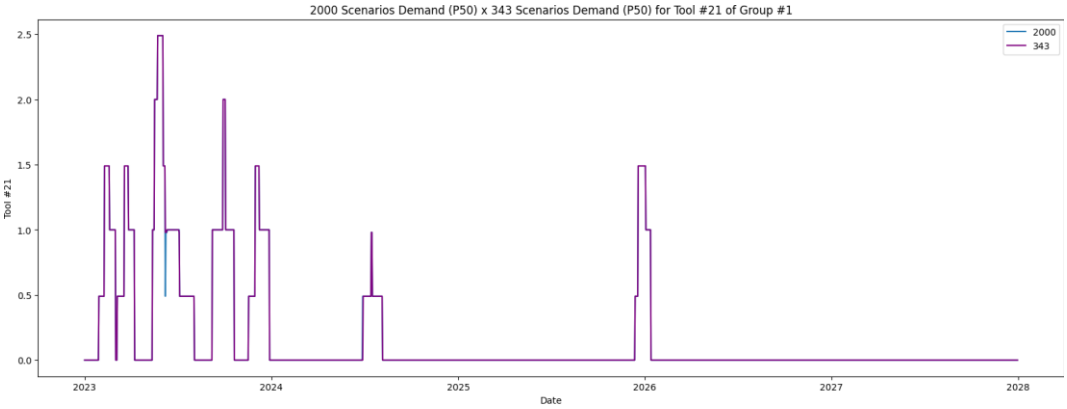

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #16 of Group #1
2000
343
Tool #16
Date
2023
2024
2025
2026
2027
2028

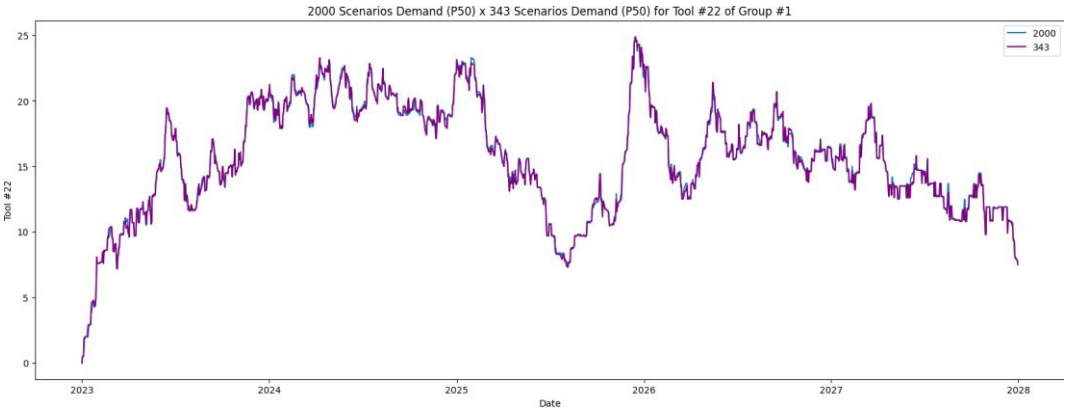

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #17 of Group #1
2000
343
Tool #17
Date
2023
2024
2025
2026
2027
2028


2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #18 of Group #1
2000
343
Tool #18
Date
2023
2024
2025
2026
2027
2028

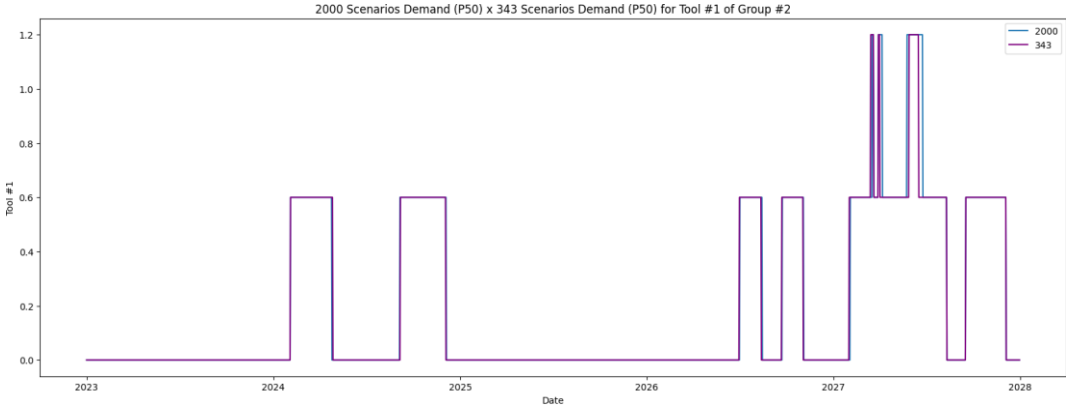

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #19 of Group #1
2000
343
Tool #19
Date
2023
2024
2025
2026
2027
2028

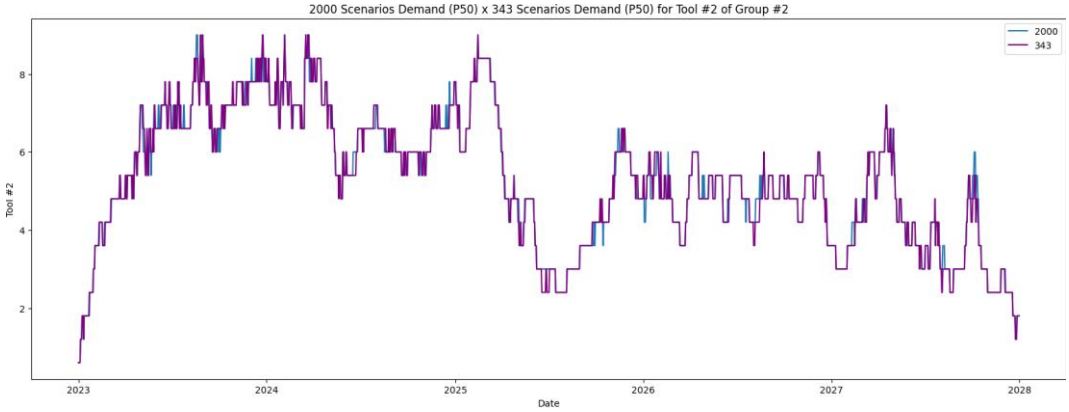

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #20 of Group #1
2000
343
Tool #20
Date
2023
2024
2025
2026
2027
2028

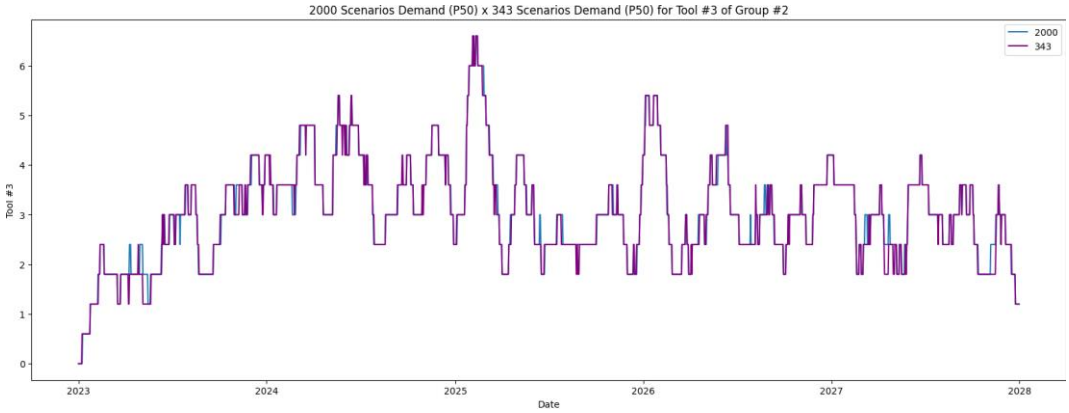

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #21 of Group #1
2000
343
Tool #21
Date
2023
2024
2025
2026
2027
2028

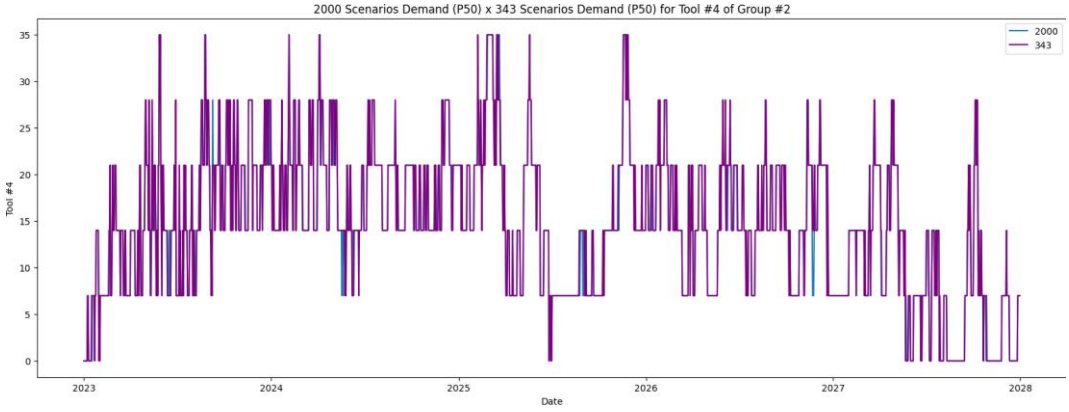

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #22 of Group #1
2000
343
Tool #22
Date
2023
2024
2025
2026
2027
2028

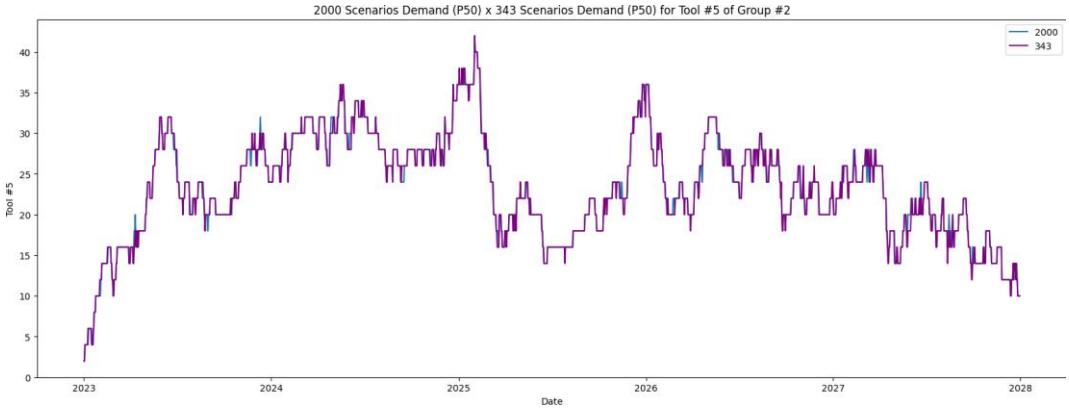
## APPENDIX VII – Comparison of P50 of Demand Calculation for Original Set and 343 Scenarios for Group 2

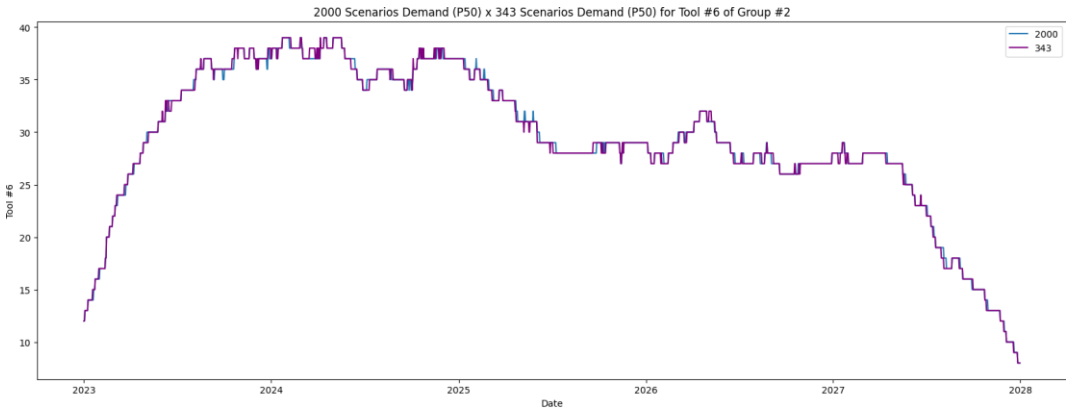


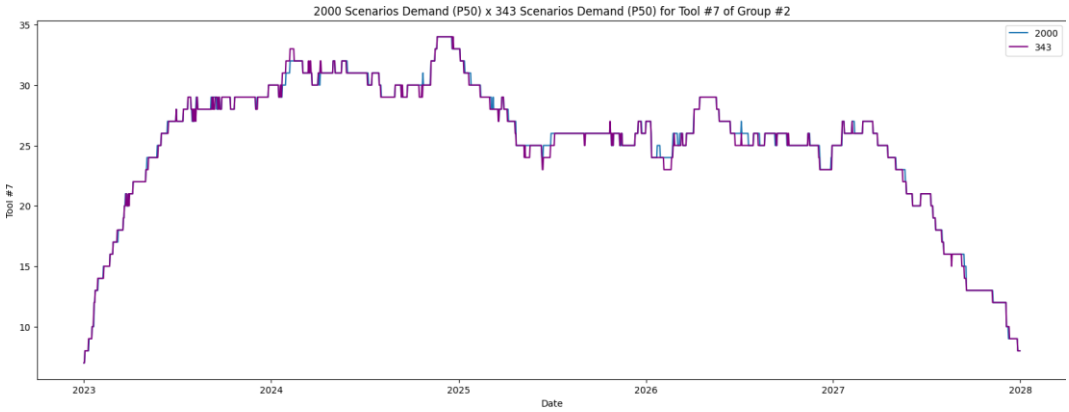



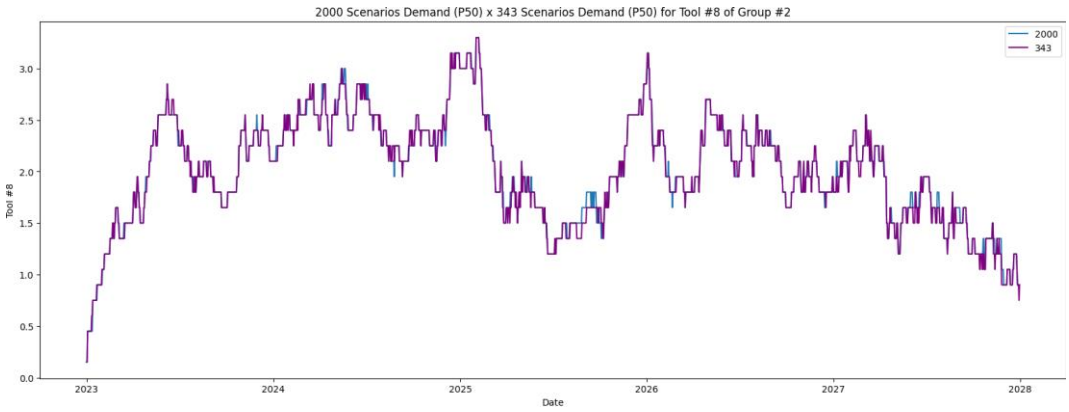



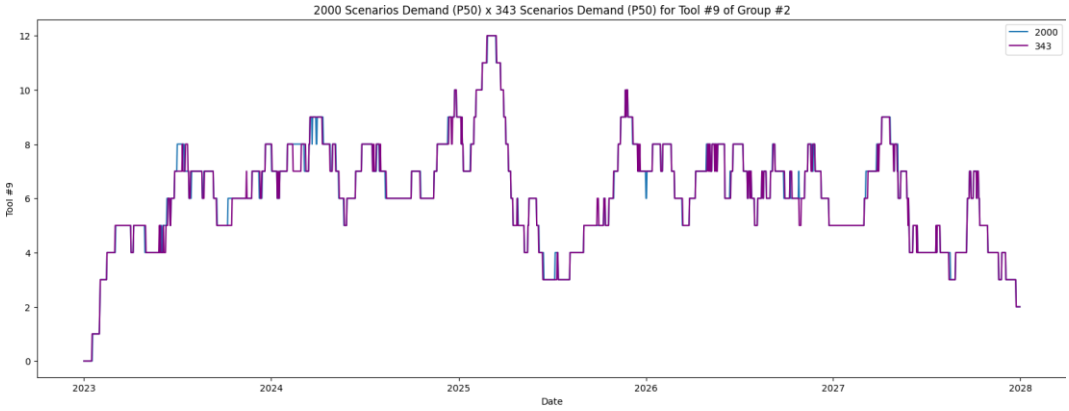



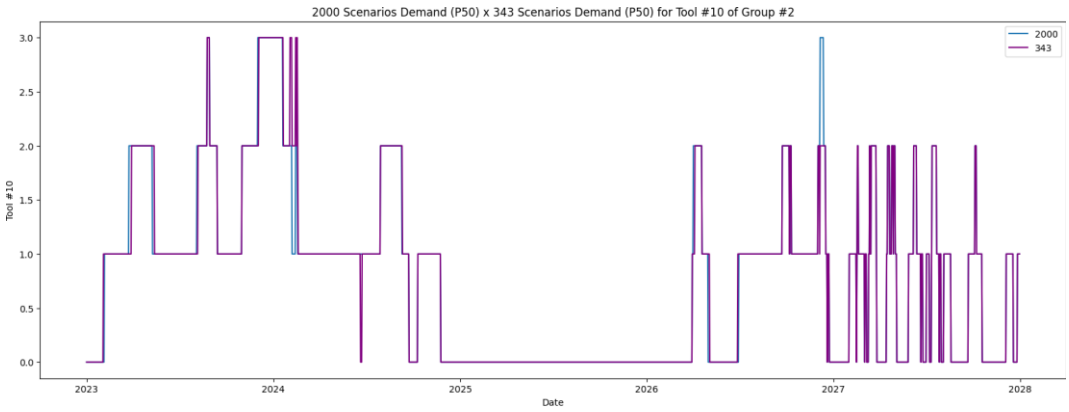

2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #5 of Group #2
2000
343
Tool #5
Date
0
5
10
15
20
25
30
35
40
2023
2024
2025
2026
2027
2028


2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #6 of Group #2
2000
343
Tool #6
Date
10
15
20
25
30
35
40
2023
2024
2025
2026
2027
2028


2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #7 of Group #2
2000
343
Tool #7
Date
10
15
20
25
30
35
2023
2024
2025
2026
2027
2028


2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #8 of Group #2
2000
343
Tool #8
Date
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2023
2024
2025
2026
2027
2028


2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #9 of Group #2
2000
343
Tool #9
Date
0
2
4
6
8
10
12
2023
2024
2025
2026
2027
2028


2000 Scenarios Demand (P50) x 343 Scenarios Demand (P50) for Tool #10 of Group #2
2000
343
Tool #10
Date
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2023
2024
2025
2026
2027
2028