

Cost-Driven Scenario Reduction via Data-Space K-means and Nearest Economic Selection for Stochastic Programming

Kanisorn Sritapak and Udom Janjarassuk

Department of Industrial Engineering, School of Engineering
King Mongkut's Institute of Technology Ladkrabang
Bangkok, 10520, Thailand
64601209@kmitl.ac.th, udom.ja@kmitl.ac.th

Abstract

Scenario reduction is an essential process for solving large-scale two-stage stochastic optimization models, as the number of realizations directly affects complexity and computational burden. Traditional methods often focus primarily on statistical distributions between the original and reduced sets, or struggle with the trade-off between solution quality and computation time. This study proposes a novel scenario reduction approach termed K-means clustering with Nearest Economic Selection (K-NES). The methodology combines data space partitioning with a recourse evaluation phase, utilizing the expected value problem as a proxy to estimate costs and identify optimal economic medoids. Furthermore, a progressive iteration framework is presented to systematically scale the subset size. The proposed algorithm was implemented in Python using the Gurobi solver and tested on two benchmark cases: Aircraft Allocation (GBD) and Electricity Planning (LandS). Experimental results show that K-NES outperforms conventional techniques, such as Fast Forward Selection, K-means clustering, and Monte Carlo sampling, in both computational efficiency and solution quality. Even when scenarios were reduced to 1-4% of the sample size, the algorithm achieved relative gaps of 0.014% and 0.001% in GBD and LandS, respectively. Our approach requires approximately one second for the reduction step. Under the progressive iteration framework, K-NES met the stopping criterion of a relative gap of less than or equal to 0.001% in 95% and 90% of the evaluations for GBD and LandS, respectively. These findings highlight the effectiveness of K-NES as a practical and scalable tool for addressing uncertainty in large-scale operations.

Keywords

Two-stage stochastic optimization, Scenario reduction, K-means clustering, Cost-driven heuristic.

1. Introduction

Uncertainty in parameters poses a significant challenge to the reliability of mathematical optimization models (Ning and You 2019). To mitigate this issue, stochastic optimization has emerged as a powerful paradigm that systematically embeds probabilistic elements into the decision-making process. Specifically, the two-stage stochastic program (TSSP) with recourse is widely utilized (Torres et al. 2022). Under the TSSP framework, the problem is divided into two phases: first-stage decisions are formulated before the uncertainty is realized, while second-stage recourse actions are executed after the random outcomes are observed (Li and Grossmann 2021). In practice, continuous probability distributions are approximated by a finite set of discrete scenarios (Gonzalez-Castellanos and Pozo 2021). Although a large number of scenarios provide a more accurate representation of the target distribution, the mathematical formulation is significantly expanded, leading to high computational costs (Kammammettu and Li 2023).

Consequently, scenario reduction is commonly performed to extract a representative subset, allowing the problem to be solved efficiently (Heitsch and Römisich 2003).

Traditional scenario reduction techniques focus on approximating probability distributions by emphasizing the structural similarity between the original and reduced sets (Dupačová et al. 2003). Sampling methods are also popular but have limitations in terms of stability and reliability (Linderoth et al. 2006). Consequently, the actual economic impact on the objective function is overlooked in both approaches. Although problem-driven reduction algorithms yield high-quality solutions, they incur significantly higher computational overhead compared to conventional methods (Keutchan et al. 2023). Recently, machine learning approaches have offered computational acceleration but remain hindered by extensive offline training and specialized architectures (Dong et al. 2024). A further complication arises as the target number of reduced scenarios must be predefined by decision-makers before algorithm execution (Bertsimas and Mundry 2022). Therefore, the motivation for this study stems from the observation that integrating a distance-driven approach with an economic selection process remains largely unexplored in the existing literature.

Accordingly, a scenario reduction method called K-means clustering with Nearest Economic Selection (K-NES) is proposed in this study. Initially, the scenarios are partitioned based on structural similarity. Then, the expected value problem (EVP) is solved using the mean values of the stochastic parameters as a proxy to estimate recourse costs and identify the economic medoids. Furthermore, the algorithm is integrated within a progressive iteration framework to systematically determine the appropriate subset size. This design minimizes the reliance on manual trial-and-error often associated with conventional techniques. To demonstrate its computational capability, the performance of the K-NES approach is benchmarked against conventional reduction methods using standard problems.

1.1 Objectives

The primary objectives of this study are threefold.

- 1) To propose a novel scenario reduction approach (K-NES) that integrates data-space characteristics with cost-driven metrics.
- 2) To evaluate the solution quality and computational efficiency of the proposed approach against traditional methods using standard benchmark problems in two-stage stochastic optimization.
- 3) To validate the practical applicability and reliability of the proposed approach within an automated progressive iteration system for two-stage stochastic optimization.

2. Literature Review

Scenario reduction is widely employed as a fundamental preprocessing step in two-stage stochastic programs to mitigate the computational complexity of these models. Pioneering work in this domain introduced probability-metric algorithms to efficiently identify representative sub-distributions (Dupačová et al. 2003). Subsequent research established bounds and accelerated variants for forward and backward selection heuristics (Heitsch and Römisich 2003). Clustering algorithms, particularly K-means and its advanced initialization technique (K-means++), have been used to partition the realizations into representative clusters based on spatial proximity (Arthur and Vassilvitskii 2007). Furthermore, a constant-factor approximation scheme based on the K-median clustering problem was developed to effectively tackle this computational challenge (Rujeerapaiboon et al. 2022). Another prevalent approach involves generating finite discrete subsets directly from the true distribution using Monte Carlo sampling under the Sample Average Approximation (SAA) framework (Linderoth et al. 2006). Generally, these methodologies are highly favored in practical applications due to their straightforward implementation and rapid execution time.

More advanced methodologies have framed the scenario selection process as a mixed-integer linear programming (MILP) problem to directly minimize the Wasserstein distance (Li and Li 2016). Alternatively, entropy-regularized optimal transport via the Sinkhorn distance was introduced (Kammammettu and Li 2023). Other MILP-based models have been formulated to incorporate the output performance of decision-making into the reduction procedure (Li and Floudas 2014). Similarly, a cost-space scenario clustering framework was introduced to partition scenarios based on their opportunity costs (Keutchan et al. 2023). For instance, a problem-dependent scenario reduction method was developed to select a subset that minimizes the loss of information regarding the objective function (Bertsimas and Mundry 2022). Additionally, adaptive partition-based algorithms iteratively refine subsets using explicit optimality gaps (Gonzalez-Castellanos and Pozo 2021). This ensures a highly accurate representation and significantly improves solution quality. Recently, an unsupervised neural network framework based on the Wasserstein distance has been

explored for scenario reduction (Dong et al. 2024). Additionally, the application of autoencoders has been investigated to encode latent representations for pattern recognition in time-series data (Liang and Tang 2021). These models provide an acceleration in computational speed after being properly trained.

However, a critical research gap remains in the existing literature. While traditional distance-driven and sampling techniques are favored for their simplicity, they inherently disregard the economic impact on the objective function. Conversely, problem-driven algorithms incur prohibitive computational overhead and complexity when applied to large-scale instances. In addition, machine learning models require offline training time. Furthermore, a fundamental challenge across all these approaches is their operation as an open-loop system, where the target subset size must be predetermined by decision-makers. Consequently, an inefficient trial-and-error process is executed.

Therefore, the K-NES algorithm is formulated to integrate structural clustering with problem-driven approaches for the scenario reduction process. By utilizing the EVP as a proxy to estimate recourse costs, the algorithm identifies economic medoids without the heavy computational burden of solving the complete recourse model. Moreover, a progressive iteration framework is implemented and executed alongside SAA to automatically regulate the size of the reduced set. This effectively eliminates the need for extensive manual parameter tuning. Thus, the proposed methodology is highly scalable and efficient for two-stage stochastic problems.

3. Methods

3.1 Problem Formulation and Solution Methodology

To address the optimization problem under uncertainty, a stochastic programming (SP) framework is adopted and formulated as a two-stage stochastic linear program with fixed recourse (Birge and Louveaux 2011). Accordingly, the first-stage decisions (x) are made before the realization of the random vector (ξ). Second-stage recourse decisions (y) are then taken to adjust for the actual outcomes. The classical formulation can be expressed as Eq. (1).

$$\begin{aligned} \min_x \quad & c^T x + E_{\xi}[Q(x, \xi(\omega))] \\ \text{s.t.} \quad & Ax = b, \\ & x \geq 0 \end{aligned} \quad (1)$$

where $Q(x, \xi(\omega))$ represents the second-stage value function for a predetermined decision x under a specific realization of the random event ω . It is formulated as $Q(x, \xi(\omega)) = \min \{q(\omega)^T y \mid Wy = h(\omega) - T(\omega)x, y \geq 0\}$. Based on this, the objective function is constructed to minimize the deterministic initial cost ($c^T x$) and the expected recourse cost, which is taken over all realizations of ω . This is subject to the first-stage constraints, where A and b represent the coefficient matrix and the right-hand side vector, respectively. Regarding the stochastic constraints, $T(\omega)$ denotes the technology matrix, W is the recourse matrix, and $h(\omega)$ indicates the right-hand side vector.

In real-world applications, computing the exact mathematical expectation over a continuous probability distribution is computationally intractable. Typically, the random vector is assumed to have a finite number of realizations. Let k denote the index of scenarios with an associated probability p_k . Moreover, solving a problem with a large number of discrete scenarios remains computationally demanding. In practice, the SAA method is applied by generating a finite sample set consisting of independent and identically distributed scenarios via Monte Carlo sampling. Consequently, the problem can be transformed into a deterministic equivalent formulation in the extensive form. By replacing the true distribution with this empirical sample, and assuming an equal probability of each scenario, the expected second-stage cost is approximated by the sample average. The extensive form is formulated as Eq. (2).

$$\begin{aligned} \min \quad & c^T x + \frac{1}{n} \sum_{k=1}^n q_k^T y_k \\ \text{s.t.} \quad & Ax = b, \\ & T_k x + W y_k = h_k, \quad k = 1, \dots, n; \\ & y_k \geq 0, \quad k = 1, \dots, n; \\ & x \geq 0 \end{aligned} \quad (2)$$

where n is the total size of the sample set. In this study, a large-scale SAA baseline is generated to establish a ground truth. Specifically, the exact baseline solution is obtained by solving the full sample set (n). This solution is utilized to benchmark the performance of the scenario reduction algorithms. To ensure statistical reliability, the evaluation is performed across multiple independent replications.

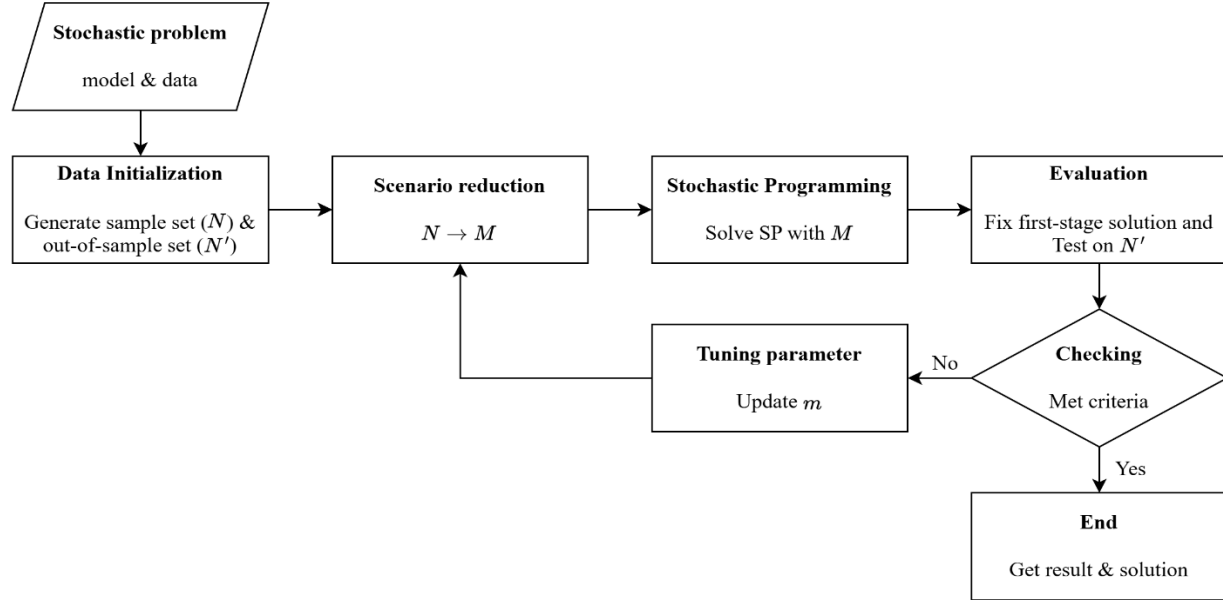


Figure 1. The proposed progressive iteration framework for solving two-stage stochastic optimization

Figure 1 illustrates the solution methodology and the proposed progressive iteration framework. Starting with the original stochastic problem, a sample set (N) and an out-of-sample (OOS) set (N') of sizes n and n' are generated via Monte Carlo sampling. Next, we apply a scenario reduction process to extract a smaller set (M) of size m from N . Subsequently, the stochastic programming model is solved using this reduced set. After that, the obtained first-stage solution is fixed and evaluated on N' to determine the OOS cost and computational time. Normally, a conventional approach would end at this step. However, our automated framework continues by checking the predefined stopping criteria, which depend on solution quality and resource limitations. If these criteria are not met, the system tunes the parameter by increasing m and loops back into the scenario reduction process. The loop finally stops when all criteria are satisfied.

3.2 Scenario Reduction

A scenario reduction technique is essential to ensure that large-scale stochastic problems can be solved within a reasonable computational time. The main objective of this process is to extract a representative dataset that preserves both the original probability distribution and the actual economic impact on the objective function. Accordingly, the K-NES algorithm is proposed. While conventional methods primarily focus on the structural similarity of the raw data, the proposed K-NES incorporates an economic evaluation into the selection process. The overall procedure consists of three main steps, as presented in Algorithm 1.

In the economic cost evaluation step, the cost of each scenario is calculated to evaluate its economic impact. Initially, the mean of the stochastic parameters is computed, as shown in Eq. (3). Then, the EVP is solved, as demonstrated in Eq. (4) (Birge and Louveaux 2011). Subsequently, the first-stage decision variable x within the feasible region X , denoted as x_{EVP} , is obtained. This variable is then fixed across all scenarios to compute the recourse cost (f_s) for each realization s , as presented in Eq. (5). These evaluated costs for each realization act as economic indicators, representing the outcome if a particular scenario is chosen, and they are further utilized in the selection step.

$$\bar{\xi} = \sum_{s=1}^n p_s \xi_s \quad (3)$$

$$\min_{x \in X} \{c^T x + Q(x, \bar{\xi})\} \quad (4)$$

$$f_s = Q(x_{EVP}, \xi_s) \quad (5)$$

Next, the K-means clustering algorithm is applied to partition the n scenarios into m clusters based on their distance to the cluster centroids (Arthur and Vassilvitskii 2007). Following this, the probability mass (p'_l) of cluster l is adjusted by aggregating the probabilities of its constituent members, as calculated in Eq. (6).

$$p'_l = \sum_{s \in C_l} p_s \quad (6)$$

where s and p_s represent the scenario index and its corresponding probability, respectively, and C_l is the set of scenario indices belonging to that cluster. Finally, a representative scenario is selected for each cluster using the economic cost (f_s) evaluated in the first step. Within each cluster l , the economic medoid (s_l^*) is identified by finding the specific scenario that minimizes the sum of Euclidean distances to the economic costs of all other scenarios in the same cluster, as demonstrated in Eq. (7).

$$s_l^* = \arg \min_{i \in C_l} \sum_{j \in C_l} \|f_i - f_j\| \quad (7)$$

Algorithm 1: K-means clustering with Nearest Economic Selection (K-NES)

Input: Initial sample set $N = \{\xi_1, \dots, \xi_n\}$ with probabilities p_s , target subset size m .

Output: Reduced scenario set M and updated probabilities p'_l .

// Step 1: Economic cost evaluation

1. $\bar{\xi} \leftarrow \sum_{s=1}^n p_s \xi_s$
2. $x_{EVP} \leftarrow \arg \min_{x \in X} \{c^T x + Q(x, \bar{\xi})\}$
3. **for** $s = 1, \dots, n$ **do**
4. $f_s \leftarrow Q(x_{EVP}, \xi_s)$ // Compute economic cost
5. **end for**

// Step 2: Structural clustering

6. Separate $\{\xi_1, \dots, \xi_n\}$ into m clusters $\{C_1, \dots, C_m\}$ via K-means clustering
7. **for** $l = 1, \dots, m$ **do**
8. $p'_l \leftarrow \sum_{s \in C_l} p_s$ // Aggregate cluster probability
9. **end for**

// Step 3: Representative selection

10. $M \leftarrow \emptyset$
11. **for** $l = 1, \dots, m$ **do**
12. $s_l^* \leftarrow \arg \min_{i \in C_l} \sum_{j \in C_l} \|f_i - f_j\|$ // Identify economic medoid
13. $\xi_l^* \leftarrow \xi_{s_l^*}$ // Extract representative vector
14. $M \leftarrow M \cup \{\xi_l^*\}$
15. **end for**
16. **return** M and $\{p'_1, \dots, p'_m\}$

4. Results and Discussion

4.1 Experimental Setup and Benchmark Problems

The computational performance of the proposed K-NES algorithm is evaluated using two standard benchmark problems in two-stage stochastic optimization. The first benchmark is the Aircraft Allocation (GBD) (Dantzig 1963). The second instance is the Electricity Planning (LandS) (Louveaux and Smeers 1988). The structural characteristics and specific data parameters for both instances, summarized in Table 1, are adopted from the literature (Linderoth et al. 2006). For the experimental setup, the Sample Average Approximation (SAA) approach was adopted. An initial sample set of 5,000 scenarios and a large OOS set of 200,000 scenarios were generated from the true underlying probability distribution using Monte Carlo sampling. Scenario reduction was subsequently performed on the sample set to extract the reduced subsets, and the true performance was evaluated against the OOS set.

Table 1. Structural characteristics of the benchmark problems

Characteristics	Aircraft Allocation (GBD)	Electricity Planning (LandS)
Objective	Maximize expected profit	Minimize expected total cost
# Original Scenarios	6.5×10^5	10^6
# 1st-stage Vars.	17 (Aircraft allocation)	4 (Technology capacities)
# 1st-stage Cons.	4 (Aircraft availability)	2 (Capacity and budget limits)
# 2nd-stage Vars.	10 (Carried/bumped passengers)	12 (Electricity production modes)
# 2nd-stage Cons.	5 (Demand balance)	7 (Capacity and demand limits)
# Random parameters	5 (Passenger demands)	3 (Electricity demands)

All computational experiments were executed in Python on a machine equipped with an Intel® Core™ i5-1035G1 CPU @ 1.00GHz and 8.0 GB of RAM. The formulations of the two-stage stochastic programs were solved using the Gurobi solver. To ensure statistical reliability, the performance of the proposed K-NES method was benchmarked against the Fast Forward Selection (FFS) implementation of Dupačová's greedy heuristic (Bertsimas and Mundru 2022), standard K-means clustering (Arthur and Vassilvitskii 2007), and Monte Carlo (MC) sampling with replacement (Bertsimas and Mundru 2022). All methods were tested over 20 independent replications, and their performances were measured in terms of solution quality (OOS cost, relative gap) and computation time, including progressive iteration efficiency. Specifically, the relative gap is calculated to quantify the percentage deviation of the OOS cost obtained from a scenario reduction method ($z_{oos}(\hat{x})$) relative to the baseline OOS cost derived from the full sample set ($z_{oos}(x^*)$), as formulated in Eq. (8).

$$\text{Relative gap (\%)} = \left| \frac{z_{oos}(\hat{x}) - z_{oos}(x^*)}{z_{oos}(x^*)} \right| \times 100 \quad (8)$$

For the performance evaluation of methods, the experiments are divided into two parts. First, Sections 4.2 and 4.3 test the methods with fixed subset sizes (m) to benchmark solution quality and reduction time. Second, Section 4.4 evaluates the proposed progressive iteration framework to demonstrate its practical applicability.

4.2 Solution Quality Performance

In terms of solution quality, the results regarding the OOS cost and relative gap are presented. First, the OOS cost evaluates the expected performance and the structural stability of the reduced scenarios (M) when exposed to a massive set of realizations. This metric ensures the reliability of the outcomes and closely approximates the true system cost across multiple replications. Additionally, the average and standard deviation of all results are summarized in Table 2. Second, the relative gap measures the deviation of the solution obtained from the scenario reduction method relative to the deterministic equivalent method, which solves the full sample set without any reduction process.

Table 2. Out-of-Sample cost performance comparison of scenario reduction methods (averaged over 20 independent replications)

Problem	Method	$m = 50$		$m = 100$		$m = 200$	
		OOS cost	Std. Dev.	OOS cost	Std. Dev.	OOS cost	Std. Dev.
GBD	FFS	1656.617	2.231	1656.437	2.287	1655.971	1.741

	K-means	1658.637	2.376	1657.581	1.768	1657.637	1.834
	MC	1661.002	5.382	1658.515	4.785	1656.294	1.667
	K-NES	1656.969	1.779	1656.487	1.830	1655.677	1.961
LandS	FFS	225.677	0.116	225.669	0.114	225.665	0.114
	K-means	225.680	0.113	225.668	0.116	225.665	0.113
	MC	225.808	0.253	225.706	0.116	225.695	0.113
	K-NES	225.681	0.116	225.668	0.114	225.665	0.114

As shown in Table 2, K-NES and FFS consistently yielded superior OOS costs compared to other approaches at equivalent scenario sizes for the GBD problem. Notably, the proposed algorithm achieved the lowest expected OOS cost of 1655.677 at $m = 200$. Moreover, the OOS costs incurred by FFS (1655.971), MC sampling (1656.294), and K-means (1657.637) are higher than those of our technique. Furthermore, the standard deviation of K-NES remained low across all subset levels, which demonstrates the robustness of the method. A similar trend is presented in the LandS problem. Although the cost of LandS is smoother and inherently produces lower variability across all methods, K-NES, FFS, and K-means continue to provide highly stable performance (achieving an absolute cost of 225.665 at $m = 200$), whereas MC sampling exhibits inferior performance in this dataset (225.695 at $m = 200$).

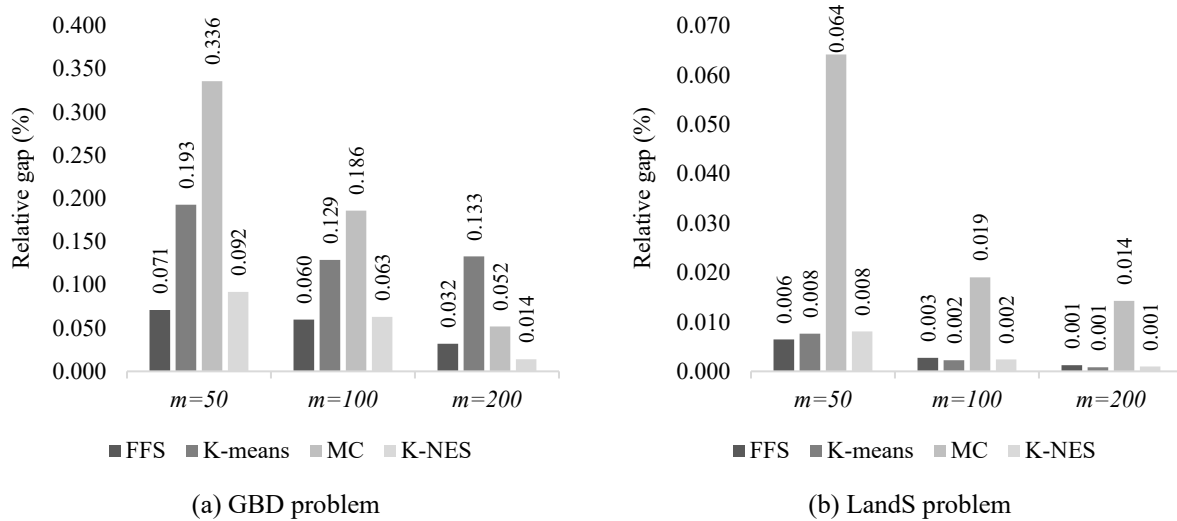


Figure 2. Relative gap of scenario reduction methods (averaged over 20 independent replications)

As demonstrated by Figure 2, the relative optimality gap further illustrates the solution quality performance of each method, which aligns with the OOS cost trend. Moreover, the capability of all methods generally improves as the number of scenarios increases. In the GBD problem, standard K-means exhibits a substantial structural deficiency because the calculated centroids may not represent realizable scenarios, and the random process in MC is unsatisfactory, resulting in a high relative gap at all levels by 0.133%, 0.052%, respectively at $m = 200$. Conversely, FFS yielded the best results at $m = 50$ and 100, but at $m = 200$, K-NES achieved a remarkably low error of 0.014%, outperforming the FFS method's 0.032%. In the LandS problem, FFS, K-means, and K-NES consistently deliver highly precise solutions with a 0.001% gap at $m = 200$. Therefore, analysis of the results in terms of solution quality confirms that K-NES can reliably maintain the key economic properties of traditional distribution, performing efficiently, and providing highly accurate results in both environments.

4.3 Computational Time

In terms of computational efficiency, the reduction time is an essential factor for the practical applicability of scenario reduction algorithms in stochastic optimization. This section focuses on this metric, as the times required for solving the stochastic program and OOS evaluation depend on the subset size and the defined OOS size, remaining relatively

constant across all methods. As illustrated in Figure 3, the processing time for the FFS approach grows exponentially as m increases in both the GBD and LandS test cases. Conversely, the proposed K-NES framework achieves a highly efficient execution time that mirrors the speed of K-means and MC. For instance, at $m = 200$, K-NES completes the procedure in approximately one second, while FFS consumes between 51 and 55 seconds depending on the dataset. This confirms that our method is a scalable solution for large-scale operations compared to traditional techniques.

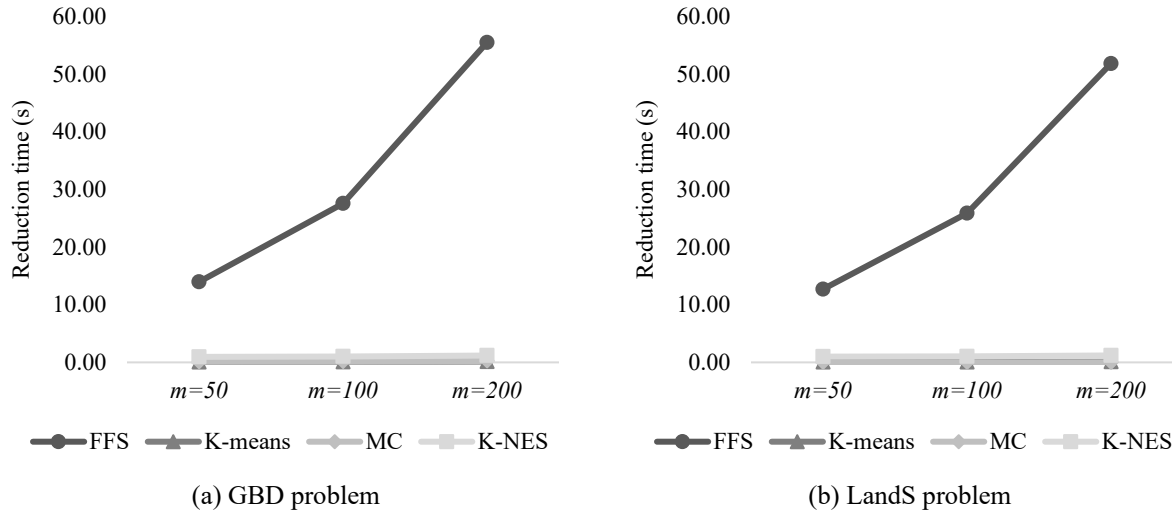


Figure 3. Scenario reduction time of each method (averaged over 20 independent replications)

4.4 Progressive Iteration Performance

To validate practical applicability in automated decision-making systems, a progressive iteration procedure is evaluated. Generally, decision-makers do not know the optimal value of m in advance and often rely on trial-and-error approaches until an acceptable solution is found. In this experiment, the initial subset size is set to $m = 50$. If the relative gap exceeds the stringent optimality threshold of 0.001%, m is systematically doubled in the subsequent iterations. This process continues until either the stopping criterion of a relative gap $\leq 0.001\%$ is met, or a maximum resource limit is reached, which is defined as $m = 200$ or a cumulative run time exceeding 900 seconds. This automated system guarantees a final solution without requiring manual parameter tuning. The algorithmic performance and stopping criteria distribution are summarized in Table 3.

Table 3. Summary of progressive iteration results and stopping conditions (evaluated across 20 independent replications)

Problem	Method	Stopping at m (%)				Stopping Criteria (%)	
		$m = 50$	$m = 100$	$m = 200$	Average m	Relative Gap $< 0.001\%$	Resource Limit
GBD	FFS	15.00	5.00	80.00	172.50	20.00	80.00
	K-means	0.00	0.00	100.00	200.00	0.00	100.00
	MC	15.00	10.00	75.00	167.50	40.00	60.00
	K-NES	10.00	15.00	75.00	170.00	95.00	5.00
LandS	FFS	0.00	15.00	85.00	185.00	40.00	60.00
	K-means	5.00	20.00	75.00	172.50	85.00	15.00
	MC	0.00	0.00	100.00	200.00	0.00	100.00
	K-NES	5.00	40.00	55.00	152.50	90.00	10.00

As presented in Table 3, under this automated iteration system, K-NES demonstrates superior convergence capabilities. Notably, since no replications exceeded the 900-second time limit, all resource limit terminations were strictly due to reaching the maximum $m = 200$. In the GBD instance, our proposed method successfully met the stringent relative gap criterion in 95% of the replications, requiring an average of 170 scenarios. Although MC sampling recorded a slightly lower average of 167.5, it satisfied the quality condition in only 40% of the replications. In contrast, K-means completely failed to meet the quality criterion, consistently hitting the maximum resource limit. Meanwhile, the FFS managed to meet the relative gap condition in merely 20% of the replications with an average m of 172.5.

Furthermore, K-NES requires the lowest average number of scenarios for decision-making compared to FFS, K-means, and MC sampling in the LandS problem. In addition, K-NES achieved a 90% success rate in meeting the relative gap criterion, outperforming K-means (85%), FFS (40%), and MC (0%). In conclusion, this experiment demonstrates that K-NES serves as a highly reliable approach, rapidly identifying optimal stability with fewer progressive steps and significantly decreasing the overall computational burden in automated environments.

4.5 Discussion

The results from both problems demonstrate that the K-NES method is highly proficient in bridging data-space characteristics with cost-driven metrics during the scenario reduction procedure for solving stochastic optimization. Generally, decision-makers are forced to accept a significant trade-off between solution quality and computational complexity. For instance, the FFS approach provides high performance but requires substantial computational time. Conversely, clustering methods like K-means offer rapid execution but often yield poor, unrealistic solutions due to the generation of centroids, which can lead to unsuitable outcomes in certain environments. Furthermore, both FFS and K-means are fundamentally distance-based approaches that ignore the problem's objective cost, while MC sampling remains random and unstructured. Our algorithm consistently exhibited a lower standard deviation across all tests, confirming its superior accuracy and stability. Additionally, the EVP is solved only once, which directly reduces the overall computational burden. Ultimately, the K-NES algorithm overcomes these fundamental limitations and effectively ensures high-fidelity bounds without incurring a prohibitive runtime.

In this study, the proposed framework was implemented within the Sample Average Approximation (SAA) approach, as deriving an exact analytical solution for large-scale real-world applications is computationally intractable. This advancement holds substantial practical value from an industrial engineering and operations management perspective because K-NES achieves high-quality solutions using a reduced size of merely 1% to 4% of the sample size (n). Modern decision-support systems, such as those utilized in GBD or LandS, operate under high-dimensional uncertainties and require rapid, automated capabilities without the need for manual parameter tuning. The progressive iteration performance confirms that K-NES is a highly reliable algorithm for such automated environments.

Despite these advantages, the proposed framework presents certain limitations. While standard Monte Carlo sampling was deliberately used to establish a fair comparative baseline, its inherent randomness may occasionally yield a poorly representative set. Regarding scalability, the K-NES algorithm requires solving the proxy objective cost for every initial scenario. Consequently, the total execution time will inevitably increase as the sample size grows. Furthermore, in practical implementations, operators are still required to predefine n , the OOS size, and the initial reduced size, parameters that directly dictate the overall runtime.

5. Conclusion

This paper proposes the K-NES framework, a novel scenario reduction approach for two-stage stochastic optimization that integrates data-space characteristics from K-means clustering with cost-driven metrics. Evaluated on the GBD and LandS benchmark problems, K-NES achieved superior out-of-sample costs and highly precise relative optimality gaps compared to traditional methods such as FFS, K-means, and Monte Carlo sampling. Simultaneously, it maintains a rapid execution time comparable to fast data-space methods, while effectively breaking the computational bottleneck associated with heuristics like the FFS approach. Furthermore, when implemented in a progressive iteration system, K-NES proved to be highly reliable for automated environments. It successfully met stringent optimality criteria while requiring a consistently low average number of scenarios, achieving success rates of 95% and 90% for the GBD and LandS instances, respectively.

These findings confirm that the proposed method serves as a highly robust alternative scenario reduction technique that balances both underlying distribution data and economic cost structures. By comprehensively evaluating solution quality, computational efficiency, and practical applicability within an automated progressive iteration system, this study successfully fulfills its core research objectives. Future research will focus on integrating machine learning techniques into the scenario reduction procedure to create an intelligent closed-loop system capable of autonomously determining the optimal number of scenarios. This progression will further enhance and support decision-making systems for large-scale applied stochastic optimization.

References

- Arthur, D. and Vassilvitskii, S., k-means++: the advantages of careful seeding, *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027-1035, New Orleans, USA, January 7-9, 2007.
- Bertsimas, D. and Mundru, N., Optimization-based scenario reduction for data-driven two-stage stochastic optimization, *Operations Research*, vol. 71, no. 4, pp. 1343-1361, 2022.
- Birge, J. R. and Louveaux, F., *Introduction to Stochastic Programming*, 2nd ed., Springer, New York, NY, USA, 2011.
- Dantzig, G. B., The allocation of aircraft to routes under uncertain demand, *Linear programming and extensions*, Princeton University Press, 1963.
- Dong, X., Sun, Y., Malik, S. M., Pu, T., Li, Y. and Wang, X., Scenario reduction network based on Wasserstein distance with regularization, *IEEE Transactions on Power Systems*, vol. 39, no. 1, 2024.
- Dupačová, J., Gröwe-Kuska, N. and Römisches, W., Scenario reduction in stochastic programming: An approach using probability metric, *Mathematical Programming*, vol. 95, pp. 493-511, 2003.
- Gonzalez-Castellanos, A. and Pozo, D., An optimal scenario reduction method for stochastic power system problems, *2021 IEEE Madrid PowerTech*, Madrid, Spain, June 28 - July 2, 2021.
- Heitsch, H. and Römisches, W., Scenario reduction algorithms in stochastic programming, *Computational Optimization and Applications*, vol. 24, pp. 187-206, 2003.
- Kammammettu, S. and Li, Z., Scenario reduction and scenario tree generation for stochastic programming using Sinkhorn distance, *Computers and Chemical Engineering*, vol. 170, Art. no. 108122, 2023.
- Keutchan, J., Ortmann, J. and Rei, W., Problem-driven scenario clustering in stochastic optimization, *Computational Management Science*, vol. 20, Art. no. 13, 2023.
- Li, C. and Grossmann, I. E., A review of stochastic programming methods for optimization of process systems under uncertainty, *Frontiers in Chemical Engineering*, vol. 2, Art. no. 622241, 2021.
- Li, Z. and Floudas, C. A., Optimal scenario reduction framework based on distance of uncertainty distribution and output performance: I. Single reduction via mixed integer linear optimization, *Computers and Chemical Engineering*, vol. 70, pp. 50-66, 2014.
- Li, Z. and Li, Z., Linear programming-based scenario reduction using transportation distance, *Computers & Chemical Engineering*, vol. 88, pp. 50-58, 2016.
- Liang, J. and Tang, W., Scenario reduction for stochastic day-ahead scheduling: A mixed autoencoder based time-series clustering approach, *IEEE Transactions on Smart Grid*, vol. 12, no. 3, pp. 2652-2662, 2021.
- Linderoth, J., Shapiro, A. and Wright, S., The empirical behavior of sampling methods for stochastic programming, *Annals of Operations Research*, vol. 142, pp. 215-241, 2006.
- Louveaux, F. V. and Smeers, Y., Optimal investments for electricity generation: A stochastic model and a test problem, *Numerical techniques for stochastic optimization problems*, Springer-Verlag, Berlin, pp. 445-452, 1988.
- Ning, C. and You, F., Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming, *Computers & Chemical Engineering*, vol. 125, pp. 434-448, 2019.
- Rujeerapaiboon, N., Schindler, K., Kuhn, D. and Wiesemann, W., Scenario reduction revisited: Fundamental limits and guarantees, *Mathematical Programming*, vol. 191, pp. 207-242, 2022.
- Torres, J. J., Li, C., Apap, R. M. and Grossmann, I. E., A review on the performance of linear and mixed integer two-stage stochastic programming software, *Algorithms*, vol. 15, Art. no. 103, 2022.

Biographies

Kanisorn Sritapak is a Ph.D. student in the Department of Industrial Engineering at King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. He received his Bachelor's degree in Industrial Engineering from King Mongkut's University of Technology North Bangkok in 2018 and his Master's degree in Industrial Engineering from Burapha University in 2020. His research interests focus on stochastic optimization, metaheuristics, operations research, and machine learning.

Udom Janjarassuk is a lecturer in the Department of Industrial Engineering at King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. He received his Ph.D. degree in Industrial Engineering from Lehigh University, USA in 2009. His research interests include stochastic optimization, supply chain management, discrete event simulation, metaheuristics, and operations research.