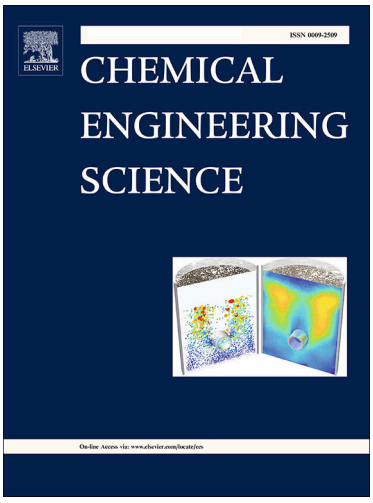
Journal Pre-proofs



Data Driven Robust Optimization for Handling Uncertainty in Supply Chain Planning Models

Kapil M. Gumte, Priyanka Devi Pantula, Srinivas Soumitri Miriyala, Kishalay Mitra

PII: S0009-2509(21)00454-1
DOI: https://doi.org/10.1016/j.ces.2021.116889
Reference: CES 116889

To appear in: *Chemical Engineering Science*

Received Date: 16 December 2020
Revised Date: 1 May 2021
Accepted Date: 18 June 2021

Please cite this article as: K.M. Gumte, P. Devi Pantula, S. Soumitri Miriyala, K. Mitra, Data Driven Robust Optimization for Handling Uncertainty in Supply Chain Planning Models, *Chemical Engineering Science* (2021), doi: https://doi.org/10.1016/j.ces.2021.116889

Data Driven Robust Optimization for Handling Uncertainty in Supply Chain Planning Models

Kapil M Gumte, Priyanka Devi Pantula, Srinivas Soumitri Miriyala and Kishalay Mitra*

Global Optimization and Knowledge Unearthing Laboratory,
Department of Chemical Engineering, Indian Institute of Technology Hyderabad,
Kandi, Sangareddy, Telangana 502285, INDIA

*corresponding author: Kishalay Mitra

email: kishalay@che.iith.ac.in

Abstract

While addressing supply chain planning under uncertainty, Robust Optimization (RO) is regarded as an efficient and tractable method. As RO involves calculation of several statistical moments or maximum / minimum values involving the objective functions under realizations of these uncertain parameters, the accuracy of this method significantly depends on the efficient techniques to sample the uncertainty parameter space with limited amount of data. Conventional sampling techniques, e.g. box/budget/ellipsoidal, work by sampling the uncertain parameter space inefficiently, often leading to inaccuracies in such estimations. This paper proposes a methodology to amalgamate machine learning and data analytics with RO, thereby making it data-driven. A novel neuro fuzzy clustering mechanism is implemented to cluster the uncertain space such that the exact regions of uncertainty are optimally identified. Subsequently, local density based boundary point detection and Delaunay triangulation based boundary construction enable intelligent Sobol based sampling to sample the uncertain parameter space more accurately. The proposed technique is utilized to explore the merits of RO towards addressing the uncertainty issues of product demand, machine uptime and production cost associated with a multiproduct, and multisite supply chain planning model. The uncertainty in supply chain model is thoroughly analysed by carefully constructing examples and its case studies leading to large scale mixed integer linear and nonlinear programming problems which were efficiently solved in GAMS® framework. Demonstration of efficacy of the proposed method over the box, budget and ellipsoidal sampling method through comprehensive analysis adds to other highlights of the current work.

Keywords: Uncertainty Modelling; Supply chain Management; Data driven Robust Optimization; Neuro Fuzzy Clustering; Multi-Layered Perceptron

1. Introduction

Supply Chain (SC) Planning aims at predicting optimal future requirements through effective coordination among key business units and successful integration of activities undertaken by the enterprises, to balance supply and demand over a time horizon (Simchi-Levi et al., 2004). Several issues in today's fierce competition, such as shrinking resources, rising cost, short product life, customer's changing preference with demand variability, technology obsoleteness and market globalization are causing threat to many companies leading them to invest in SC. It becomes more realistic to consider the presence of uncertainty during planning a SC owing to the volatile market conditions, where enterprises must meet customer satisfaction under such changing environments (Santoso et al., 2005). Mathematically, this can be expressed as an optimization problem i.e. $\min_{\mathbf{x}} \{[f(\mathbf{x},\mathbf{u})]:g(\mathbf{x},\mathbf{u}) \geq 0\}$, where f represents objective function and g represents the set of constraints and both these objective and constraints can be functions of **u**, showing uncertain parameter and **x**, showing the decision variable. Examples of uncertain parameters in SC include price, cost of raw material, interest rate, currency exchange rates, penalties, demands, machine uptime, safety stock level at inventories, delivery time between echelons, rate of production and process conversion etc.(Reid and Sanders, 2019). Conventionally, uncertainty in industrial supply chains is handled by overestimation or over-design of the capacities so that disturbances due to uncertainties can be absorbed (Chernobai et al., 2006). Another popular approach in industry is to make use of nominal values of the uncertain parameters and solve the deterministic formulation(Long et al., 2012). However, the former leads to a very costly design whereas the latter, though relatively agile, either tends to miss opportunity or over produces during uncertain parameter values higher and lower than the assumed nominal value, respectively. Most popular software providing solutions to supply chain problems neglect the effect of uncertainties in parameters for the ease in analysis of the results and solve a deterministic model. Hence, there is a need for development and usage of efficient uncertainty handling methods while solving SC planning problems.

Several uncertainty handling methods e.g. Stochastic Programming (SP) (Guillén et al., 2005), fuzzy programming (FP) (Mitra, 2009), Chance Constrained Programing (CCP) (Mitra, 2009), Robust Optimization (RO) (Vallerio et al., 2016), etc., are in popular use in academics and research to perform optimization under uncertainty. Scenario based stochastic programming approach has been utilized to model Supply chains having discrete as well as continuous uncertain parameters with known probability distribution (Hammami et al., 2014). In one variant of this algorithm, namely scenario based two stage stochastic programming (TSSP), decision variables are divided into two sets: variables that are independent of uncertain parameters ("here and now") and the variables that are dependent on the uncertain parameters ("wait and see"). In the first stage, "here and now" variables are decided before the realization of the uncertain parameters (Guillén et al., 2005; Shapiro, 2011). To deal with infeasibility due to stochastic nature of uncertain parameters, the "wait-and-see" variables are selected in recourse manner. The combined effect of the first stage costs and the expected value of stochastic second stage cost is minimized to find the "here and now" variables. Recourse function can include uncertainty as binary, integer, non-linear variable or parameter with multiple time periods and hence planning horizon variations due to stochastic parameter can be resolved (Georgiadis et al., 2011). Novel versions of established algorithms such as Bender's decomposition (Keyvanshokooh et al., 2016), L-shaped algorithm (Rajgopal et al., 2011), Dantzig-Wolfe decomposition approach (Dantzig, 1998), fix-and-relax coordination (Abdelaziz et al., 2007) and LR algorithm (Aghezzaf, 2005) have been developed to handle difficult instances of stochastic programming problems. The limitation to this approach is the exponential increase in problem size with the increase in number of uncertain parameters and their assumed scenarios of realizations, leading to immense computational expense and suboptimal solution within the given time frame. Even the decomposition of the problem into multiple stages might be quite difficult at times.

When relaxation of one or more constraints are allowed due to the presence of uncertainty, constraints having uncertain parameters can be modified defining some probability of constraint satisfaction associated

with them and thereby introducing reliability of the obtained solutions (Govindan et al., 2017). Here, the original chance constraints, having uncertain parameters, are transformed into their deterministic equivalent forms using probability concepts. CCP can be joint or individual based on the nature of correlation exists among the uncertain parameters. The deterministic equivalent form can be non-linear in nature due to the involvement of mean and variance terms of the uncertain parameters, where the property of convexity can be restricted (Mitra et al., 2008). The size of the equivalent deterministic problem is generally manageable even in the presence of a large number of uncertain parameters unless the probability of constraint satisfaction is set to be very high (i.e. very close to unity). However, going by the definition of a robust solution, which is a fixed decision variable vector that should remain feasible irrespective of the realization of the uncertainty in the parameters (Rajgopal et al., 2011), the solutions obtained by CCP may not be robust always. So, robust optimization is an alternate way to handle optimization under uncertainty (OUU) problems and might be extremely important under situations where the cases of constraint violation are highly restricted.

In RO formulation, generally, the stochastic nature of the parameters is made available in terms of data collected over a broad period of time and the stochastic optimization problem is converted into an equivalent deterministic problem, known as the robust counterpart (RC), where the stochastic objective functions and constraints are computed for various realizations of uncertain parameters. As this needs sampling from the uncertain parameter space, a significant RO based research was focused on how to sample the uncertain space more effectively. The earliest attempts by Soyster provided a computationally tractable approach with guaranteed feasibility, where the uncertain space is approximated by the box defined by the bounds of the uncertain set (Soyster, 1973). To improve upon the nature of over-conservative decisions in the box approach, ellipsoidal uncertainty sets have been introduced (El Ghaoui et al., 1998), that simplifies the robust counterpart model into a conic quadratic problem with linear constraints. Further improvements include the combination of interval, ellipsoidal and adjustable polyhedral uncertainty sets that can help a decision maker to obtain solutions better than the box uncertainty (Bertsimas and Sim, 2004; Gregory et al., 2011). Data driven RO based supply chain network design (SCND) for large scale waste water sludge to biodiesel conversion is developed by Mohseni and Pishvee (Mohseni and Pishvaee, 2020) for case study based in Iran. Utilizing support vector clustering (SVC), the uncertainty sets are constructed in this work that yielded more realistic results compared to that of the conventional uncertainty set. SVC has been used in another network planning model (Shang et al., 2017) based on piece wise linear kernel functions. Here, the geometry of uncertain data was captured by solving quadratic programming and the resultant convex uncertainty set has been shown to result in solutions with less conservatism. Instead of SVC, Ning and You (Ning and You, 2018) applied principal component analysis (PCA) for the uncertain data to extract distributional information of uncertainties using kernel density estimation techniques. Applicability of this algorithm was checked against applications involving model predictive control, batch production scheduling and process network planning. A Bayesian non parametric Dirichlet process mixture model (Ning and You, 2017) combined with variational inference algorithm has been developed through four level optimization framework. The conservatism is reduced by accounting the data's correlation, asymmetry and multimodality. To solve this multi-level model, column and constraint generation algorithms are proposed and applied to batch process scheduling and SC network planning.

Thus, from the aforementioned literature study, the specific knowledge gap can be identified as follows:

- The combination of machine learning concepts, RO for uncertainty handling and application to supply chain is rare to find.
- Moreover, the existing uncertainty sets such as box, budget, ellipsoidal, etc. have fixed geometric shapes and may not be flexible enough for handling sparse and discontinuous uncertain data. Further, the prevailing studies based on reducing the conservatism of RO solutions, are specific to the case studies considered and may or may not work for all the supply chain models.

- The techniques that are used to analyze the uncertain data (such as SVC, PCA, etc.) are applicable for handling only non-overlapping data sets, which is not always the case and most of these techniques by themselves suffer with some major drawbacks such as fixing the cluster number beforehand, less interpretable features of given data and so on, which make the overall algorithm less efficient. Particularly, in supply chain models, there are high chances that the uncertain data is over-lapping, owing to the varying cost components, sudden change in demand or other external factors.
- From the studies done till now, it is evident that there is a significant need of accurate data based sampling strategies that can provide true representation of uncertain space, which is generic in nature and devoid of heuristics to maximize the efficiency.
- Moreover, for application in Supply Chain Models (SCMs) which host significantly large number of uncertain parameters, these methods should work without constraints on dimensionality of the problem.

The current work addresses the issues mentioned above and fills the knowledge gap by presenting a techno-economic SC model with better uncertainty handling capability via supervised and unsupervised machine learning concepts. The contribution and novelties of the work are listed below:

- Utilizing the worst-case scenario out of several realizations of uncertain parameters that are obtained by efficiently sampling the uncertain parameter space using a new methodology called Data-Driven RO (DDRO), the overall formulation provides an envelope of resilient and feasible SC operation under uncertainty.
- The DDRO algorithm uses a new parameter free Fuzzy C-Means clustering algorithm along-with RO, which amalgamates the ideas of machine learning and RO to solve Supply Chain Planning problems under Uncertainty (see section 3.2 for detailed technique). In DDRO algorithm, compact and flexible uncertainty sets are constructed for sampling. The proposed technique is utilized to explore the merits of RO towards addressing the uncertainty issues of product demand, machine uptime and production cost associated with a multiproduct and multisite supply chain planning model. Comprehensive comparison among the box, budget and ellipsoidal sampling method and the proposed technique adds further value to the proposed work.
- Even if the uncertain data is sparse and discrete, the proposed data driven technique fills those gaps by generating the data points within the engulfed uncertain boundaries providing high accuracy.
- As the number of samples generated from the uncertainty set is one of the important factors that impacts the quality of the solution in RO, the effect of sample size in the uncertain parameter space has been studied systematically and compared with conventional budget, box and ellipsoidal uncertainty set based RO to find the variation in cost for the SC.

In this paper, section 2 describes the midterm planning model of McDonald and Karimi (McDonald and Karimi, 1997), which is then followed by the detailed explanation of the technique and algorithm used for the data driven robust optimization in section 3. The results and discussion section describes the considered examples and its case studies and presents elaborate analysis of effect of data-driven sampling on supply chains under uncertainty in section 4. Finally, the work is summarized and concluding remarks are enlisted in section 5.

2. Formulation and Model used

2.1 Midterm Supply Chain Model and Optimization problem formulation

Supply Chain networks generally comprise several entities beginning with raw material supplier, production or manufacturing facilities, inventories or warehouse with distribution centers and customers or end users. Given the architecture of the network among these entities, a planning model is supposed to estimate the variables such as raw material procurement, downstream mass supply for inventories and

customer, inventory needed to mitigate stock-outs, transport to be used etc. over a planning horizon. The model has two Supply Chain (SC) layers e.g. Manufacturing site and Customer Market. This example of SC (Fig. 1) manufactures 34 products p_1 to p_{34}. Products p_1 - p_{23} are produced at facility site s_1 having demands only from market 1, whereas the products p_{24} - p_{34} are produced at facility site s_2 having demands from market 2. Each facility has single unit for manufacturing with separate source of raw material suppliers and delivers products at two different markets as shown in Fig. 1. The inventory layer is not separate and is combined with the manufacturing unit. Out of 23 products from facility s_1, 11 products namely p_1, p_4, p_6, p_8, p_{10}, p_{12}, p_{14}, p_{16}, p_{18}, p_{20} and p_{22} act as resource materials for products p_{24} - p_{34} respectively. Products p_1 - p_{23} from facility s_1 form 11 sets of product families F_1 - F_{11} as shown in Fig. 1. The planning horizon is considered for one year with 12 periods, with each period representing one month.

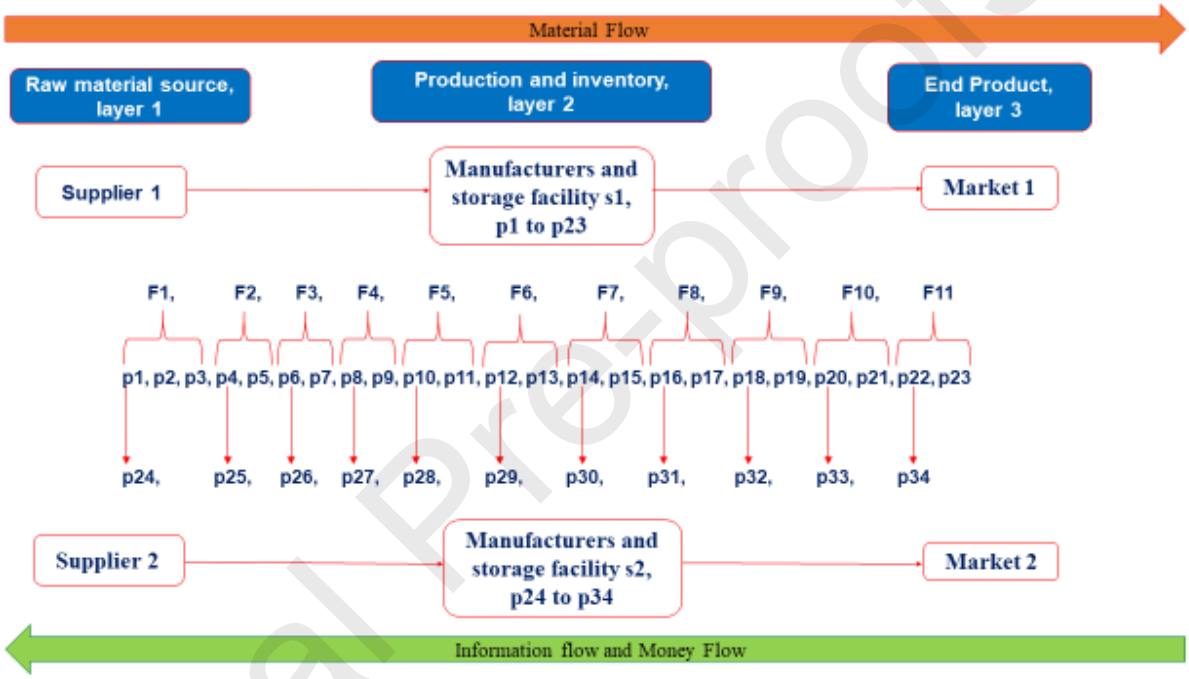


Fig. 1. Supply chain diagram with raw material, intermediate and finished product flow

Table 1. Notations of the parameters present in Planning model.

S. No	Symbol	Name	Definition
1	$FC_{f,u,l,t}$	Fixed Production Cost	Fixed cost of producing unit quantity of product family f using machine u at site l during time period t.
2	$VC_{p,u,l}$	Variable Production Cost	Variable cost of producing product p manufactured by machine u at site l.
3	$RC_{p,l}$	Raw material Cost	Unit cost of Raw material p consumed at site l.
4	$HC_{p,l,t}$	Inventory Holding Cost	Inventory Cost associated with storing unit quantity of Product p at site l during time period t.
5	$TC_{l,c}$	Transportation Cost	Cost associated with transporting the products from a site l to market / customer c.
6	$TC_{l,l'}$	Transportation Cost	Cost associated with transporting the products from a site l to another site l′.
7	$PC_{p.l}$	Penalty	Penalty for dipping below safety target of product p at site l.

8	$\mu_{p,c}$	Revenue	Revenue per unit product p sold to customer c.
9	$R_{p,u,l,t}$	Rate of production	Effective rate of producing product P on machine u at site l during time period t.
10	$MT_{u,l,t}$	Machine Uptime	Time for which machine u can be made available at site l during time period t.
11	$D_{p,c,t}$	Product Demand	Demand for product p at market / customer c during time period t.
12	$I^{L}_{p,l,t}$	Safety stock target	Safety stock target for product p at site l during time period t.
13	$MRL_{f,u,l,t}$	Minimum Run length	Minimum Time for which machine u may be run at site l to produce product family f during time period t.

Table 2. Notations in Planning model which serve as decision variables in MILP formulation.

S. No	Symbol	Name	Type	Definition
1	$B_{f,u,l,t}$	Binary Variable	Integer	{0, 1} i.e. Whether unit quantity of product family f using machine u at site l during time period t is produced or not.
2	$P_{p,u,l,t}$ / $P_{p',u,l,t}$	Production Variable	Real	Amount of final/ intermediate product p manufactured by machine u at site l during time period t.
3	$C_{p,l,t}$	Consumption Variable	Real	Amount of Intermediate Product / Raw material p consumed at site l during time period t.
4	$I_{p,l,t}$	Inventory Variable	Real	Amount of Product p stored at site l during time period t.
5	$S_{p,l,c,t}$	Supply Variable	Real	Amount of Product p supplied from a site l to market / customer c during time period t.
6	$C^{IP}_{p,l,l',t}$	Consumption Variable	Real	Amount of Intermediate Product p which was brought from site l' and consumed at site l during time period t.
7	$I^{\Delta}_{p,l,t}$	Inventory Variable	Real	Amount of Product p which needs to be added to Inventory so as to reach the safety stock target $I^{L}_{p,l,t}$ at site l during time period t.
8	$I^{-}_{p,c,t}$	Inventory Variable	Real	Amount of product p by which the agglomerated supply from all sites at the inventory missed the demand in market c during time period t.
9	$T_{p,u,l,t}$	Run Length Variable	Real	Time for which machine u was run at site l to produce product p during time period t.
10	$T_{f,u,l,t}$	Run Length Variable	Real	Time for which machine u was run at site l to produce family f during time period t.

The planning model formulates a linear cost function involving the costs of raw material consumption, production, inventory and transportation along with loss functions such as the revenue loss due to missing demand. The cost function needs to be minimized with respect to several constraints in form of linear inequalities, representing the limitations of supply chains in real world. The notations used are defined in Tables 1 and 2 while the Supply Chain Model (SCM) is described by equations 1- 19. The indices used in the SCM are defined as follows: p – product, f – family of products, u – machine, l – site / site, c – customer / market and t – time period. P signifies the product set such that, $P = \{P^{RM} \cup P^{IP} \cup P^{FP}\}$, where RM is raw materials, IP is intermediate products and FP is finished products. $\Phi_{p,f}$ is cross set indicating product p is member of family f. $\beta_{p',p,l}$ is yield adjusted amount of raw or intermediate product p that must be consumed to produce a unit of intermediate or finished product p' at site l.

Cost function:

$$\sum_{f,u,l,t} FC_{f,u,l,t} B_{f,u,l,t} + \sum_{p,u,l,t} VC_{p,u,l} P_{p,u,l,t} + \sum_{p,l,t} RC_{p,l} C_{p,l,t} + \sum_{p,l,t} HC_{p,l,t} I_{p,l,t} + \sum_{p,l,c,t} TC_{l,c} S_{p,l,c,t} + \sum_{p,l,l',t} TC_{l,l'} C^{IP}_{p,l,l',t} + \sum_{p,l,t} PC_{p,l} I^{\Delta}_{p,l,t} + \sum_{p,c,t} \mu_{p,c} I^{-}_{p,c,t} \quad (1)$$

Eq 1 represents the cost function to be minimized, which is the sum of fixed cost, variable cost, raw material cost, inventory holding cost, transportation cost, penalty cost and revenue lost respectively.

Manufacturing constraints:

$$P_{p,u,l,t} = R_{p,u,l,t} T_{p,u,l,t} \quad \forall\, p \in P \backslash P^{RM} \quad (2)$$

$$\sum_{p} T_{p,u,l,t} - MT_{u,l,t} \leq 0 \quad (3)$$

$$\sum_{f} T_{f,u,l,t} - MT_{u,l,t} \leq 0 \quad (4)$$

$$T_{f,u,l,t} = \sum_{p \in \Phi_{p,f}} T_{p,u,l,t} \quad (5)$$

$$C_{p,l,t} = \sum_{p' \ni \beta_{p',p,l} \neq 0} \beta_{p',p,l} \sum_{u} P_{p',u,l,t} \quad \forall\, p \in P \backslash P^{FP} \quad (6)$$

$$C_{p,l,t} = \sum_{l'} C^{IP}_{p,l,l',t} \quad \forall\, p \in P^{IP} \quad (7)$$

Eq 2 calculates the production as the product of production rate and machine run time. The total run time of machine will always be less than machine available time as given by Eqs 3 and 4. As per Eq 5, the run time for family of product can be represented by sum of run time for single products. Eq 6 models the consumption of raw or intermediate material using the bills of material. Raw materials from external supplier are taken for consumption and are assumed to be available based on demand. Intermediate product consumed at site l supply must come from same site or another site l` in the same time period as represented by Eq 7.

Supply chain constraints:

$$I_{p,l,t} = I_{p,l,t-1} + \sum_{u} P_{p,u,l,t} - \sum_{l'} C^{IP}_{p,l,l',t} - \sum_{c} S_{p,l,c,t} \quad \forall\, p = P \backslash P^{RM} \quad (8)$$

$$I^{-}_{p,c,t} \geq I^{-}_{p,c,t-1} + D_{p,c,t} - \sum_{l} S_{p,l,c,t} \quad \forall\, p \in I^{FP} \quad (9)$$

$$\sum_{l,t' \leq t} S_{p,l,c,t'} \leq \sum_{t' \leq t} D_{p,c,t'} \quad \forall\, t \in T \quad (10)$$

$$I^{\Delta}_{p,l,t} \geq I^{L}_{p,l,t} - I_{p,l,t} \quad \forall\, p \in I^{FP} \quad (11)$$

Eq 8 shows the basic material flow balance where the inventory at current time period is sum of inventory at the previous time period plus the production happened in the current time period minus outflow of intermediates to other plants minus shipments of finished product to the customers at the same time period. Eq 9 shows the cumulative customer shortfalls between demand and supply, where shortfall from previous time period is carried to the next time period. Cumulative demand for the current time period can be fulfilled by the supply from previous time periods also as per Eq 10. Eq 11 gives the inventory constraint, indicating that current inventory at production site $I_{p,l,t}$ should always by equal to or less than the inventory safety level $I^{L}_{p,l,t}$, such that safety stock shortage value $I^{\Delta}_{p,l,t}$ should be positive or zero.

Lower bound constraints:

$$P_{p,u,l,t}, T_{p,u,l,t}, T_{f,u,l,t}, C_{p,l,t}, S_{p,l,c,t}, I_{p,l,t}, I^{-}_{p,c,t}, C^{IP}_{p,l,l',t}, I^{\Delta}_{p,l,t} \geq 0 \quad (12)$$

Eq 12 clearly indicates the decision variables to be either positive or zero, due to model physicality in real time.

Upper bound constraints:

$$P_{p,u,l,t} \leq R_{p,u,l,t} MT_{u,l,t} \quad (13)$$

$$T_{p,u,l,t} - MT_{u,l,t} \leq 0 \quad (14)$$

$$T_{f,u,l,t} - MT_{u,l,t} B_{f,u,l,t} \leq 0 \quad (15)$$

$$T_{f,u,l,t} - MRL_{f,u,l,t} B_{f,u,l,t} \geq 0 \quad (16)$$

$$I^{\Delta}_{p,l,t} \leq I^{L}_{p,l,t} \quad (17)$$

$$S_{p,l,c,t} \leq \sum_{t' \leq t} D_{p,c,t'} \quad (18)$$

$$I^{-}_{p,c,t} \leq \sum_{t' \leq t} D_{p,c,t'} \quad (19)$$

The quantity produced will always be less than machine available time (which is always more than the actual machine run time as Eq 14 and 15) multiplied by the production rate as given by Eq 13. Eq 16 shows that the time run by the machine should be greater that minimum run time for family of products. Eq 17 states that the inventory safety level should be greater than safety stock shortage. As given by Eq 18 and 19 for model feasibility, the demand should be greater than supply and demand missed.

In the aforementioned deterministic model, uncertainty is introduced at two levels by considering the stochastic nature of parameters present in constraint functions and objective function. The uncertainty in product demand influences the constraint equations 9, 10, 18 and 19, while the uncertainty in machine uptime parameter influences the constraint equations 3, 4, 13, 14 and 15. Finally, the uncertainty in objective function (equation 1) is introduced by considering the variable production cost to be stochastic. In the present work, data driven RO (DDRO) technique is adopted to solve the optimization problem under uncertainty. In the next section, details on DDRO is presented. The data used for solving the mid-term SC model can be found in the supplementary file attached.

3. Data-Driven Robust Optimization (DDRO)

3.1 Overview of Robust optimization formulation

A generic form of stochastic optimization can be presented by Eq 20, where f is the objective function and g represents the set of constraints and both of these objective and constraints can be functions of **u** and **x,** where **u** corresponds to the vector of uncertain parameters (bounded), **x** denotes the decision variable set (bounded). This is opposed to the deterministic formulation where variability in the uncertain vector **u** is not considered and assumed the vector **u** as constant. In RO, the presence of uncertain vector **u** in objective and constraint functions is treated in two stages. First, the constraints are handled by defining a robust feasible solution. A solution, which remains feasible under all realizations of uncertain vector **u**, is called as the robust feasible solution. Next, among the robust feasible solutions, one can find a solution, called the worst-case, where the maximum value of the objective function under several realizations of uncertain parameters is minimized as presented by the Eq 21. This is called the equivalent deterministic robust counterpart (RC), where supremum among feasible set is calculated first in the uncertain space and then minimized across the decision variable space (Gorissen et al., 2015).

$$\min_{\mathbf{x}} \{[f(\mathbf{x},\mathbf{u})]: g(\mathbf{x},\mathbf{u}) \geq 0\} \quad [20]$$

$$\min_{\mathbf{x}} \left\{ \sup_{\mathbf{u}} [f(\mathbf{x},\mathbf{u}): g(\mathbf{x},\mathbf{u}) \geq 0] \right\} \quad [21]$$

The supremum in Eq. 21 is calculated using the samples drawn from the uncertainty set u, which contains all possible realizations of uncertain parameters.

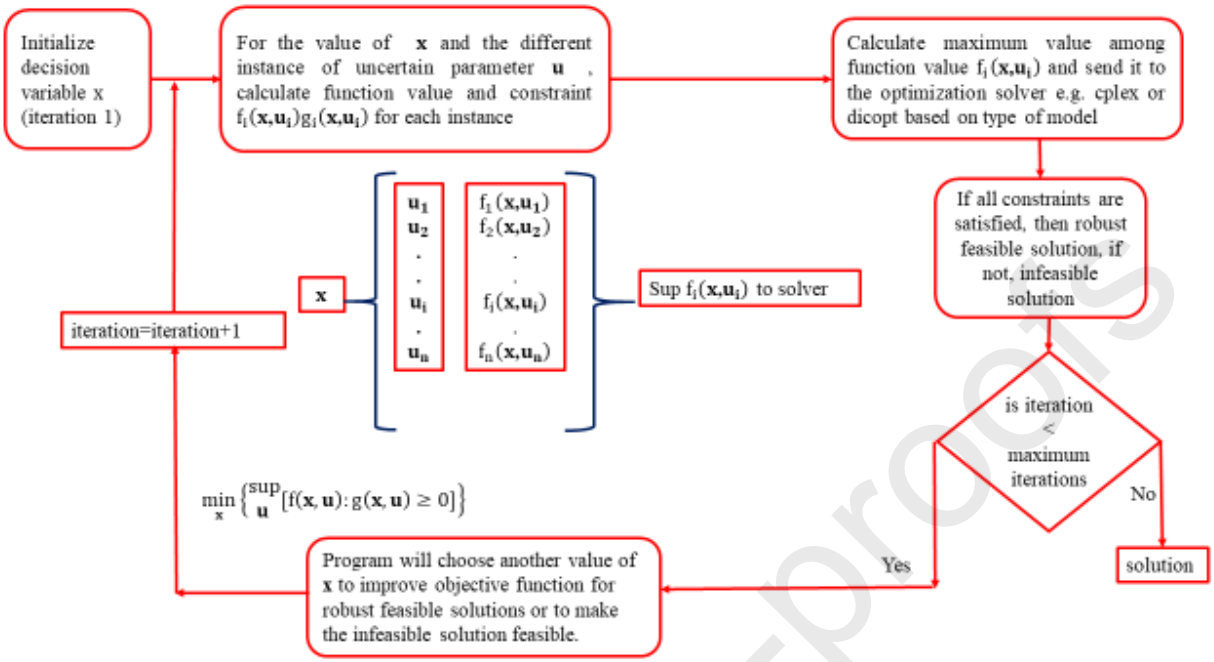


Fig 2. Schematic of solution procedure for Robust Counterpart formulation (worst-case: Eq. 21)

The algorithm of the worst-case RO formulation has been shown in Fig 2. First the value of **x** is initialized randomly. Now, keeping the decision variables **x** fixed, various instances of **u** are considered from the uncertainty set to calculate the constraint and objective function values. If all the constraints are satisfied, then from the set of robust feasible solutions, the maximum or the worst objective function value is chosen. Next, an optimization solver (e.g. CPLEX® or DICOPT®) improves the overall objective with respect to the decision variable set **x** over several iterations till the termination criteria (e.g. maximum number of iterations) are satisfied. Even after trying different values of **x**, if constraints are found violated, then solution is termed as infeasible. In this study, the uncertainty is assumed in one of the cost components e.g. production cost (Eq 1), which makes the objective function stochastic in nature. Other uncertain parameters considered in this study include demand (Eq 9, 10, 18, 19) and machine uptime (Eq 3, 4, 13, 14, 15).

Analogously, the best-case RO formulation corresponding to Eq. (20) can be defined as shown in Eq. 22, where the infimum of the robust feasible set is computed using several instantiations of uncertain parameters obtained from the uncertainty set.

$$\min_{\mathbf{x}} \left\{ \inf_{\mathbf{u}} [f(\mathbf{x},\mathbf{u}): g(\mathbf{x},\mathbf{u}) \geq 0] \right\} \quad [22]$$

For calculating the respective supremum and infimum values efficiently in Eq. 21 and Eq. 22, the cardinality of the uncertainty set needs to be sufficiently large, which demonstrates the importance of chosen uncertainty set and the ability to generate samples efficiently from the uncertain parameter space. Box, budget, ellipsoidal and polyhedral sets are commonly deployed uncertainty sets in literature. However, none of them ensures complete removal of regions, where the original uncertain data is not present. Such type of inaccurate sampling from uncertain space leads to loss of accuracy and deviation in solution of Eq 21 and 22, thereby generating over conservative results, especially when the data given is less in number and scattered in the entire uncertain space (Gorissen et al., 2015). Moreover, the given uncertain data may not necessarily follow any well-behaved statistical distribution and the data set might be discontinuous in nature. Another practical problem when this data set is coming from real life, mostly, the number of data

points present in this set is relatively less in number. Owing to these drawbacks and issues existing with the existing uncertainty sets, there is a need to design a new algorithm, which has the capability of efficient identification of the right uncertain parameter set and sampling accurately from the identified uncertain parameter space. In this regard, a novel framework called Data-driven RO (DDRO), which is proposed for handling the optimization under uncertainty in SC models is presented in the next section.

3.2 Data-Driven RO (DDRO)

This section explains the novel methodology for constructing flexible and compact uncertainty set followed by efficient sampling from the engulfed uncertain parameter space as described below.

A. The sparse, scattered and discrete data is clustered using an effective unsupervised machine learning algorithm, called Neuro Fuzzy C-means clustering (NFCM) (Pantula et al., 2020). By mapping the input data to the membership function using artificial neural network (ANN), NFCM converts large scale optimization problem in Fuzzy C-means clustering (FCM) to small scale. This enables the usage of global optimization algorithms such as, genetic algorithm (GA) for clustering the data efficiently. Further, the NFCM algorithm enables the estimation of optimal number of clusters by optimizing an internal cluster validation index (Pantula et al., 2020).
B. After the data in the uncertain space is clustered, the following steps are performed for each of the cluster.
   1. The local density (LD) of all the points is calculated by identifying the number of points that lie within a threshold radius.
   2. The boundary points are detected using LD such that the points having relatively less LD lie in the outer space of cluster.
   3. Next, the boundary points are linked using Delaunay triangulation (Fortune, 1992) for creating a continuous boundary of the cluster.
   4. Subsequently, a hypercube is generated using the maximum and minimum points along each dimension of the cluster boundary.
   5. New sample points are generated within the hypercube by means of Sobol sampling (Sobol, et al., 2011) for ensuring uniform sampling within the uncertain space.
   6. The Sobol points that are inside the boundary are preserved while those lying outside the boundary are eliminated.
C. The samples thus generated using B1 – B6 for each cluster thus forms the new uncertain data set, which is used in worst and best-case RO formulations as mentioned in Section 3.1.

Therefore, in case of DDRO, the unnecessary regions of sampling are eliminated intelligently by combining unsupervised machine learning (i.e., NFCM) with density-based boundary point detection and Delaunay triangulations such that the unnecessary regions of sampling are eliminated to a maximum possible extent.

The pictorial representation of data-driven, box, budgeted (with budget $= 1$) and ellipsoidal uncertainty sets is shown in Fig. 3. The black colored points in Fig. 3 represent the given data for two uncertain parameters ($u_1$and u_2) and the dotted lines signify the boundary of the uncertainty set within which the sampling has to be performed. It can be observed that in case of box, budgeted and ellipsoidal uncertainty sets (Fig. 3 (a), (b) and (c)), the regions of sample picking include some unnecessary regions, that is, empty regions inside the uncertainty set (dotted lines). On the other hand, the uncertainty set designed in DDRO (Fig. 3 (d)) ensures drawing of samples only from the regions where the given uncertain data is present.

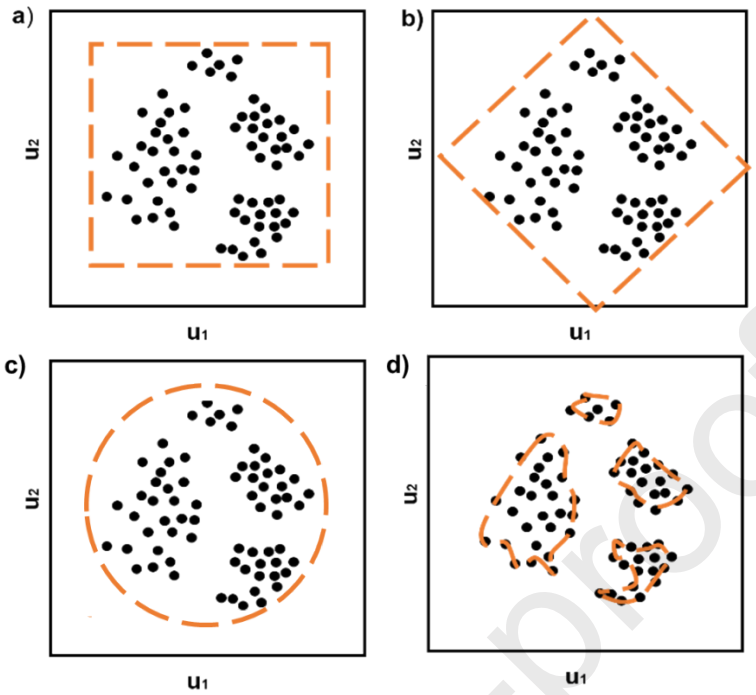


Fig. 3. Comparison of box, budgeted and DDRO uncertainty sets where black colored points represent the given data for 2 uncertain parameters and the dotted lines denote the boundary of the uncertainty set within which sampling would be performed (a) Box uncertainty set (b) Budgeted uncertainty set (Budget = 1) (c) Ellipsoidal uncertainty set (d) DDRO uncertainty set.

4 Results & Discussion

This section presents several case studies. The first case (case 1) is an instance of a multi-production site, multi-market supply chain planning model, where demands of products produced from one of the markets are assumed to be uncertain. Considering the deterministic demand values as their means, a ± 20% deviation in demand values is assumed to generate the ranges of the uncertain parameters and within this hyperspace, the uncertain data has been created synthetically in 3 different clusters using combination of several Gaussian distributions. This represents that the data is not present everywhere in the uncertain space i.e. it is scattered and limited in number in the uncertain space symbolizing the nature of realistic data. Accurate transcription of the uncertain data space is going to be the key differentiating factor among several uncertainty sets (box, budget, ellipsoidal and the proposed approach) used. Similar approach has been followed for generating uncertain data in other cases as well. In the next case (case 2), demands of products produced from both the markets are subjected to uncertainty. The extent of uncertainty has been further intensified in case 3, where along with demand, machine uptimes at both the sites are considered uncertain. Lastly, case 4 oversees the effect of uncertainty in objective function by considering the production cost as uncertain variable along with demand and machine uptime. In all these cases, the number of data points used for sampling the uncertain space has been varied from 500 – 10000 to show the effect of sampling on the final results whereas the ideal case used for comparison has been generated using 20000 data points only inside the clusters. Each case study is elaborated in the following sections.

4.1 Example 1-Case 1

First, the case of the supply chain facing uncertainty in demands from the products p_1 - p_{23} of only market 1 is considered. This brings in modifications in Eqs 9, 10, 18 and 19 as several instances of uncertain

parameter i.e. demand realizations are to be incorporated in them. After modifying the above equations, new Eqs 28, 29, 30 and 31 are obtained, where the subscript *k* represents the instances of uncertain realizations for demand data. Representing a scenario of cyclical demand, the demands values of all products as reported in the original work (McDonald and Karimi, 1997) are modified by 300% for the 6th and 12th time periods and by 20% for the rest of the time periods. This kind of sudden hike in demand might be helpful in simulating Bullwhip effect of supply chains, which is about storing inventory up the supply chain to handle sudden hike in demand in future. Keeping the demand values of products p_{24} - p_{34} as deterministic, only the demand values of products p_1 - p_{23} are perturbed as mentioned in the beginning of this section. Also, the machine uptime is kept 70% of the deterministic values presented in the original work. Since the demand parameter in every month is considered uncertain, 12 sets (12 months) of 23 dimensional (23 products) uncertain parameters were considered. The formulated MILP problem consisting of Eqs 1 to 8, 28, 29, 11 to17, 30 and 32 was solved in GAMS® using CPLEX® solver and the objective function values are shown in Table 3 for varying number of sampling points *k* in the uncertain parameter space. The problem has 132 binary variables, 8163154 single equations and 23298505 non zero elements. The computational time taken for clustering using NFCM algorithm is 0.98 seconds (for a fixed architecture of Artificial Neural Network). For the entire proposed uncertainty set in Data-driven RO, the computational time is 3.4 seconds (along-with clustering) while that of box, budget and ellipsoidal uncertainty sets are 1.64, 2.1 and 2.26 seconds respectively. The times reported here correspond to generation of fixed number of sample points. However, the computational time differs by very minimal amount (~milliseconds) on variation of sample sizes from 500 to 10000. Thus, it can be observed that the difference in computational times is not quite large on comparing the existing uncertainty sets with the proposed uncertainty set.

$$I^{-}_{p,c,t} \geq I^{-}_{p,c,t-1} + D_{p,c,t,k} - \sum_{l} S_{p,l,c,t} \quad \forall\, p \in I^{FP} \qquad [28]$$

$$\sum_{l,t' \leq t} S_{p,l,c,t'} \leq \sum_{t' \leq t} D_{p,c,t',k} \quad \forall\, t \in T \qquad [29]$$

$$S_{p,l,c,t} \leq \sum_{t' \leq t} D_{p,c,t',k} \qquad [30]$$

$$I^{-}_{p,c,t} \leq \sum_{t' \leq t} D_{p,c,t',k} \qquad [31]$$

Table 3. Comparison of cost function values obtained through box, budget, ellipsoidal and proposed method with respect to ideal value for case study 2

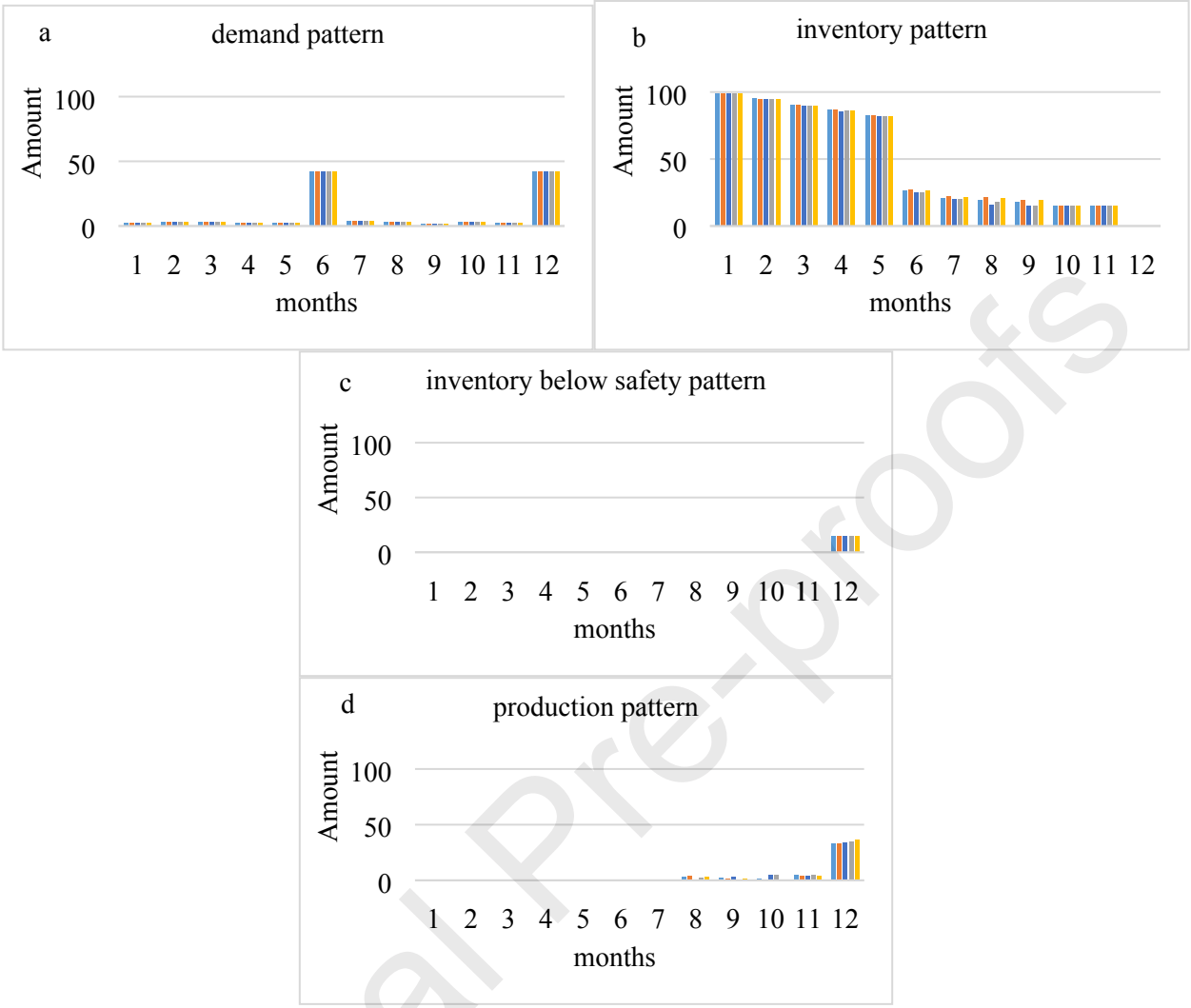
Data points	Box	box absolute deviation	Budget	Budget absolute deviation	Ellipsoidal	Ellipsoidal absolute deviation	DDRO	DDRO absolute deviation	Ideal value
500.00	19159.67	196.97	19157.75	195.05	19138.65	175.95	18398.17	564.53	18962.70
1000.00	19241.95	279.25	19222.81	260.11	19197.49	234.79	18550.08	412.62	
3000.00	19524.00	561.30	19537.67	574.97	19485.11	522.41	18731.62	231.08	
5000.00	19626.57	663.87	19651.25	688.55	19598.32	635.62	18779.87	182.83	
10000.00	19667.15	704.45	19701.25	738.55	19628.68	665.98	18822.74	139.96	

When the objective values for box, budget, ellipsoidal and proposed DDRO are compared for various number of uncertain data points i.e. from 500 to 10000 in Table 3, it is observed that initially DDRO is under-estimating the ideal value and subsequently, it approaches the ideal value as the number of data points increases. However, other approaches, like box, budget and ellipsoidal uncertainty set based RO, over-estimate the desired ideal solution irrespective of the sample size considered. The reason underlying this phenomenon lies is the way uncertain data realizations are being sampled from these sets. In case of box, budget and ellipsoidal uncertainty sets, the sampled data points either consider unnecessary regions or miss

out the regions, where the original or given uncertain data exists. Specifically, it is observed that the existing uncertainty sets have less ability to deal with non-uniform, uneven sparse uncertain data, which in turn increases the respective solution variance and deviation from ideal values. On the contrary, in case of DDRO, the Neuro Fuzzy C-means clustering algorithm used, ensures better identification of the regions to be sampled through clustering the given uncertain data efficiently, along-with estimation of cluster number. Further, owing to the law of large numbers, with the increase in data points, the accuracy of RO solution was found to be improving on implementation of DDRO.

Also the point to be notice here that the RO problem solved in the paper is of minimization worst case in nature, where worst case scenarios results are shown in table and figure. Here, first the supremum value of objective function is selected from various uncertain instances and then sent to the optimization solver for minimization. For the box, budget and ellipsoidal sets, the variance is already increasing with data points and effect of finding the supremum among the various instances further aggravates the situation providing over estimated values.

The patterns of decision variables in response to the variations in demands can be seen in Fig. 4. Out of 23 products from site s_1, product p_1 is chosen to show the trend, when the uncertain parameter space of 23 dimensions is sampled with 500 points. It can be observed from Fig. 4 that apart from being the finished product, p_1 is also the raw material for producing the product p_{24} at site s_2. In this case study, the demand of products produced at $s_2$is considered to be certain. Responding to the unusual surge in demands at 6^{th} and 12^{th} month, no production is observed till the 7^{th} month. The trend of production continues till the 12^{th} month where a surge in production is observed again. The reason for no production at the initial time periods can be attributed to the fact that the production cost is substantially higher than the inventory cost. Because of this reason, the optimizer opted to clear the initial inventory first rather than going for fresh production. However, since there is a penalty associated with the inventory going below safety level, optimal plan recommended production in time period between 7^{th} to 11^{th} month such that the safety level of inventory is maintained. Despite the fact that penalty is added for the 12^{th} month for not meeting the safety level of inventory, the optimal plan suggested clearing the 12^{th} month inventory and surged production only to meet the unusually high demand in 12^{th} month.



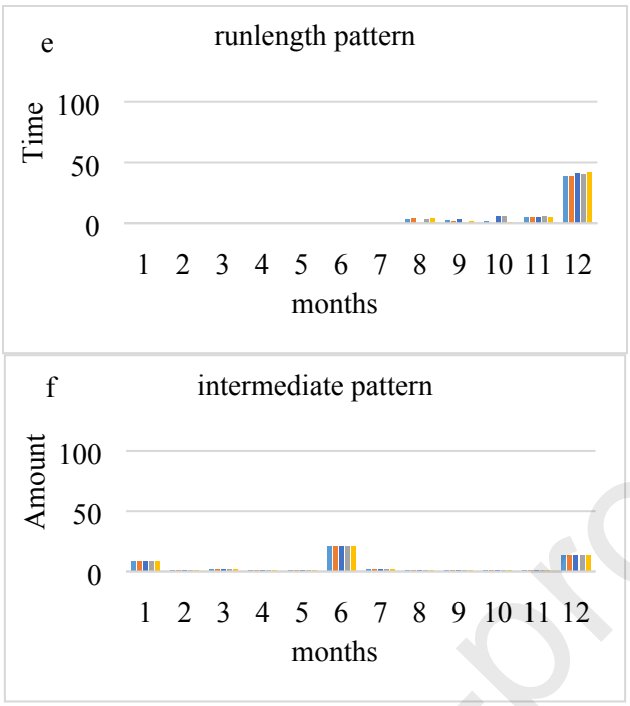


Fig. 4. a) Uncertain demand with 300% hike at 6th and 12th time periods and 20% at other time periods. The effect of uncertain demand parameter is shown on 5 sets of decision variables (moving left to right) – b) inventory c) inventory below safety d) production e) run-length f) intermediate pattern over horizon for product p_1, for 500 data points in Case study 1. The blue, orange, indigo, grey and yellow bars represent the results for box, budget, ellipsoidal, proposed methodology and ideal solution respectively.

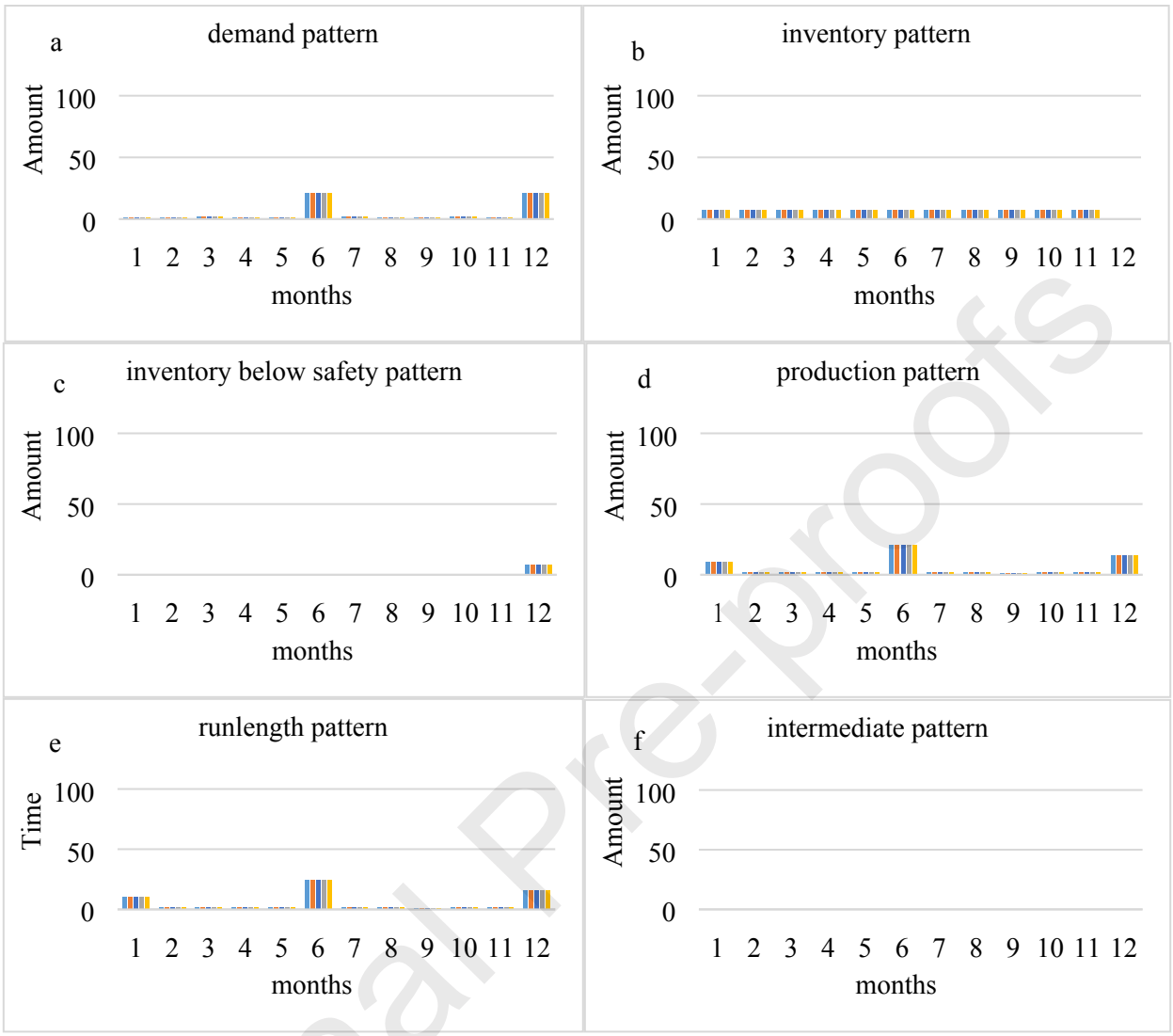
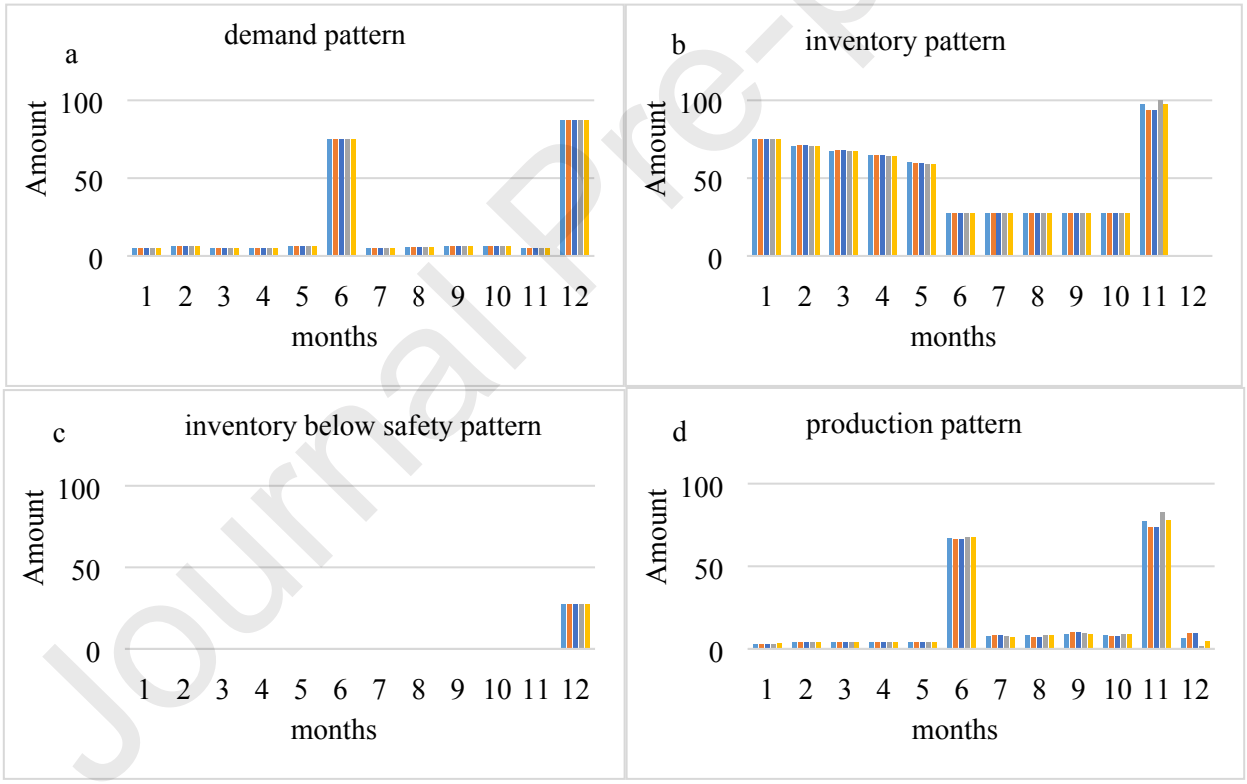


Fig. 5. a) Uncertain demand with 300% hike at 6th and 12th time period and 20% at other time periods. The effect of uncertain demand parameter is shown on 5 sets of decision variables (moving left to right) – b) inventory c) inventory below safety d) production e) run-length f) intermediate pattern over horizon for p_{24}, for 500 data points in Case study 1. The blue, orange, indigo, grey and yellow bars represent the results for box, budget, ellipsoidal, proposed methodology and ideal solution respectively.

Fig. 5 shows the trends for product p_{24} for 500 data points. The effect of demand increase at the 6th and 12th time period can be observed clearly. Inventory patterns for product p_1 and p_{24} are quite different. Since the cost for inventory going below safety level for p_{24} is higher than that of p_1 and storage cost of p_{24}, the strategy adopted by the optimizer in this case is to maintain the inventory at the safety level starting from beginning till 11th month and then let the inventory be released completely to avoid higher production cost at the 12th time period. To meet the surge in demand of p_{24} at 6th time period, the inventory is insufficient, and this additional load is taken up by production which is visible by the corresponding production profile. Rest of the profiles are following the expected patterns expressed in the constraint equations and as described for the product p_1.

Fig. 6 presents the supply chain predictions for product p_{18} when 500 points were sampled in the uncertain space. A noteworthy point here is the difference between the trends of box, budget, ellipsoidal and the proposed uncertainty handling algorithm.

The anomalous trend in bars is clearly visible in inventory, production and run length patterns at times t_{11} and t_{12} in Fig 6. If one observes closely for inventory pattern, production pattern and run length pattern in Fig 6, the blue bar (box set) seems to be closer to ideal solution at 11^{th} month time period. The main reason for this is consideration of less number of data points i.e. 500. For less number of data points, box, budget, ellipsoidal were giving less deviation from the ideal values, but the results become more reliable and prominent only on consideration of more instantiations of uncertain parameters, as per the law of large numbers in probability theory, i.e. approximately 10000 sample points in this case study (see Table 3). With such a high number of uncertain samples, it was observed that DDRO generated better solutions (closer to that of ideal solution) as compared to other techniques. The anomalous trend is further justified when it is observed in Fig. 7, which plots the trends for p_{18} for 3000 sample points, where the solution for box deviates more with respect to ideal value. Also from Fig 6 and Fig 7, the proposed DDRO bar shows values close to ideal, which indicates the working efficiency of proposed DDRO even when the sample points are less (500 with respect to 3000 data points) and is able to manage with spare data points.



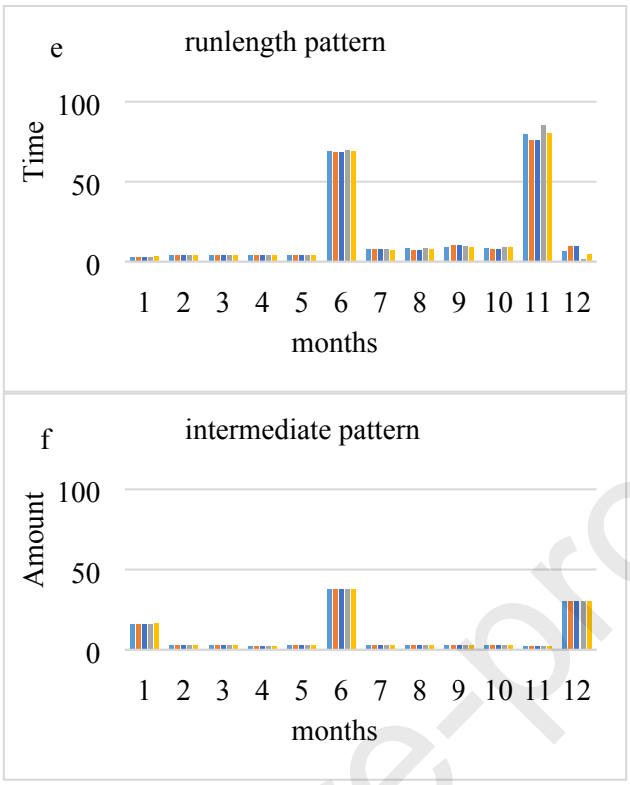


Fig. 6. a) Uncertain demand with 300% hike at 6th and 12th time periods and 20% at other time periods. The effect of uncertain demand parameter is shown on 5 sets of decision variables (moving left to right) – b) inventory, c) inventory below safety, d) production, e) run-length and f) intermediate pattern over horizon for product p_{18}, for 500 data points in Case study 1. The blue, orange, indigo, grey and yellow bars represent the results for box, budget, ellipsoidal, proposed methodology and ideal solution respectively.

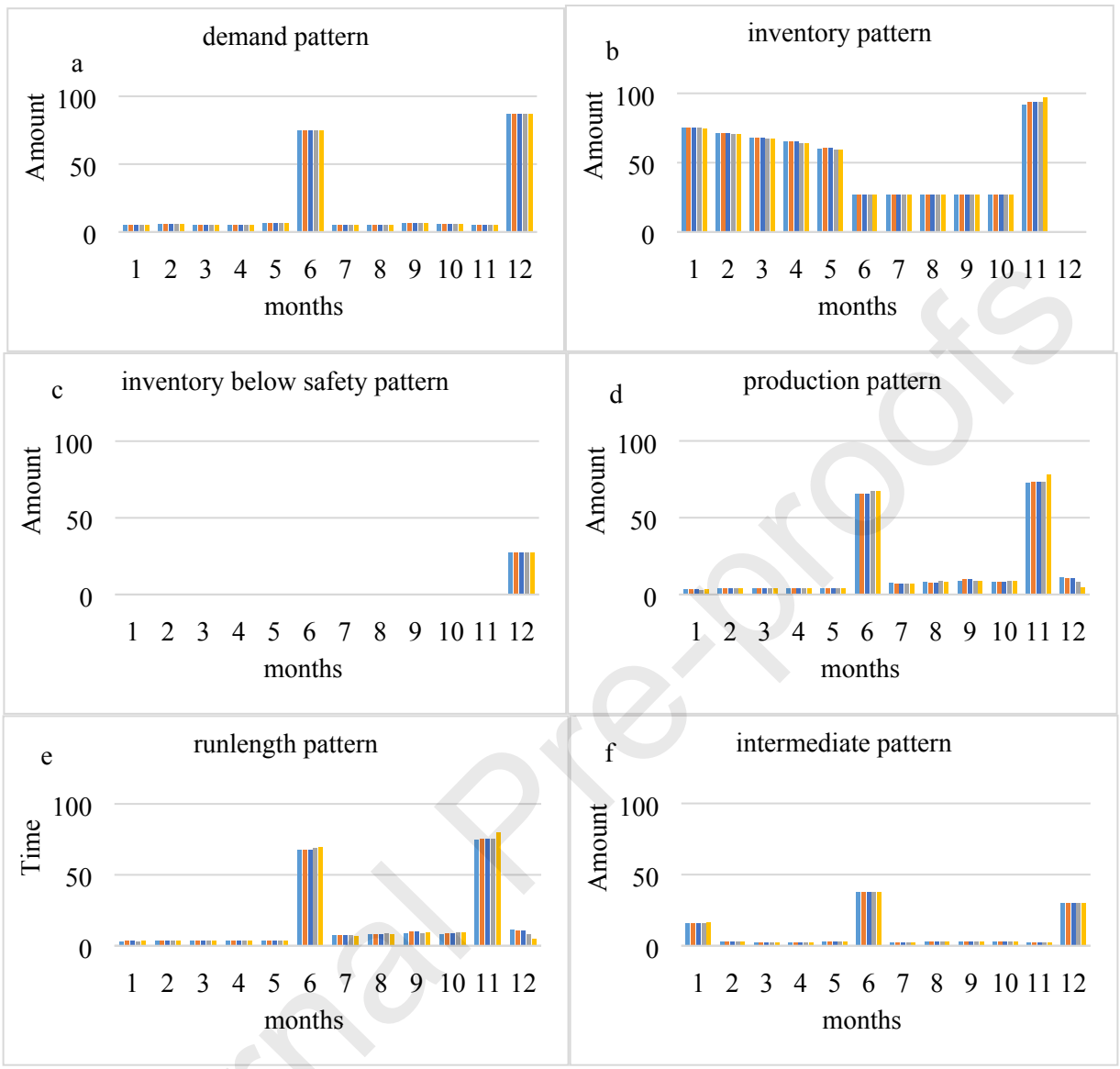


Fig. 7. a) Uncertain demand with 300% hike at 6th and 12th time periods and 20% at other time periods. The effect of uncertain demand parameter is shown on 5 sets of decision variables (moving left to right) – b) inventory, c) inventory below safety, d) production, e) run-length and f) intermediate pattern over horizon for product p_{18}, for 3000 data points in Case study 1. The blue, orange, indigo, grey and yellow bars represent the results for box, budget, ellipsoidal, proposed methodology and ideal solution respectively.

4.2 Example 1- Case 2

This is the second case study for SC uncertainty, where uncertainty is inculcated in product demand for all the products from p_1 to p_{34}, for both sites as given by Eqs 28, 29, 30 and 32. The uncertain space now is of 34 dimensions (34 products) spanned over 12 months. The MILP problem was solved in GAMS® using CPLEX® solver and the cost function values are shown in Table 4 for varying number of sample points *k* in the uncertain parameter space. The problem has 132 binary variables, 8163154 single equations and 23298505 non zero elements and takes 46.594 sec of execution time to find the optimized solution. One

can observe from Table 4 that with increase in uncertain parameters from 23 to 34 in the system compared to case study 1 (Table 3), the objective value increases showing its effect on worst-case analysis.

Table 4. Comparison of cost function values obtained through box, budget, ellipsoidal and proposed method compared to the ideal value for case study 3

Data points	Box	Box absolute deviation	Budget	Budget absolute deviation	Ellipsoidal	Ellipsoidal absolute deviation	DDRO	DDRO absolute deviation	Ideal value
500.00	23489.96	202.57	23447.80	160.41	23421.96	134.58	22577.46	709.92	23287.38
1000.00	23568.84	281.46	23580.97	293.58	23489.75	202.37	22789.78	497.61	
3000.00	23927.12	639.74	23934.14	646.76	23889.47	602.09	22991.68	295.70	
5000.00	23976.00	688.62	23991.79	704.41	23907.98	620.60	23064.59	222.79	
10000.00	23992.44	705.06	24021.86	734.48	24002.44	715.06	23487.89	200.51	

From Table 4, one can observe that the proposed DDRO objective values are moving closer to ideal values with increase in number of data points. However, for box, budget and ellipsoidal sets, the objective values are moving away from ideal values. The reason for this trend is similar to the previous case study 1.

Fig. 8 presents the predictions for product p_{18} for 3000 data points. Behavior of most of the products remains the same except few, which are affected by the uncertainty at the site 2. Due to surge in demand at 6^{th} month, and insufficient inventory, the production also jumps maintaining inventory at safety level to avoid penalty. Run-length pattern follows trend similar to production. At 12^{th} month, when demand is the highest, all the inventory is used, and to meet the left over demand, the productions at 11^{th} and 12^{th} month are used. In order to avoid the situation, where supply might not meet the market demand, the production might happen ahead in time at 11^{th} in anticipation of a sudden surge in demand at 12^{th} time period. One can see inventory below safety for 12^{th} month similar to previous cases. It is interesting to note here that when the sample size is increased from 500 to 3000, proposed methodology corrects its solution and emulates very closely the trend predicted by the ideal solution, whereas the solution obtained by box, budget and ellipsoidal sets-based RO, deviates from the ideal values. Compared to the blue (box) and orange (budget) bars, the closeness of grey bars (proposed method) to yellow bar (ideal) can be easily seen for inventory, production, run length, supply and intermediate patterns.

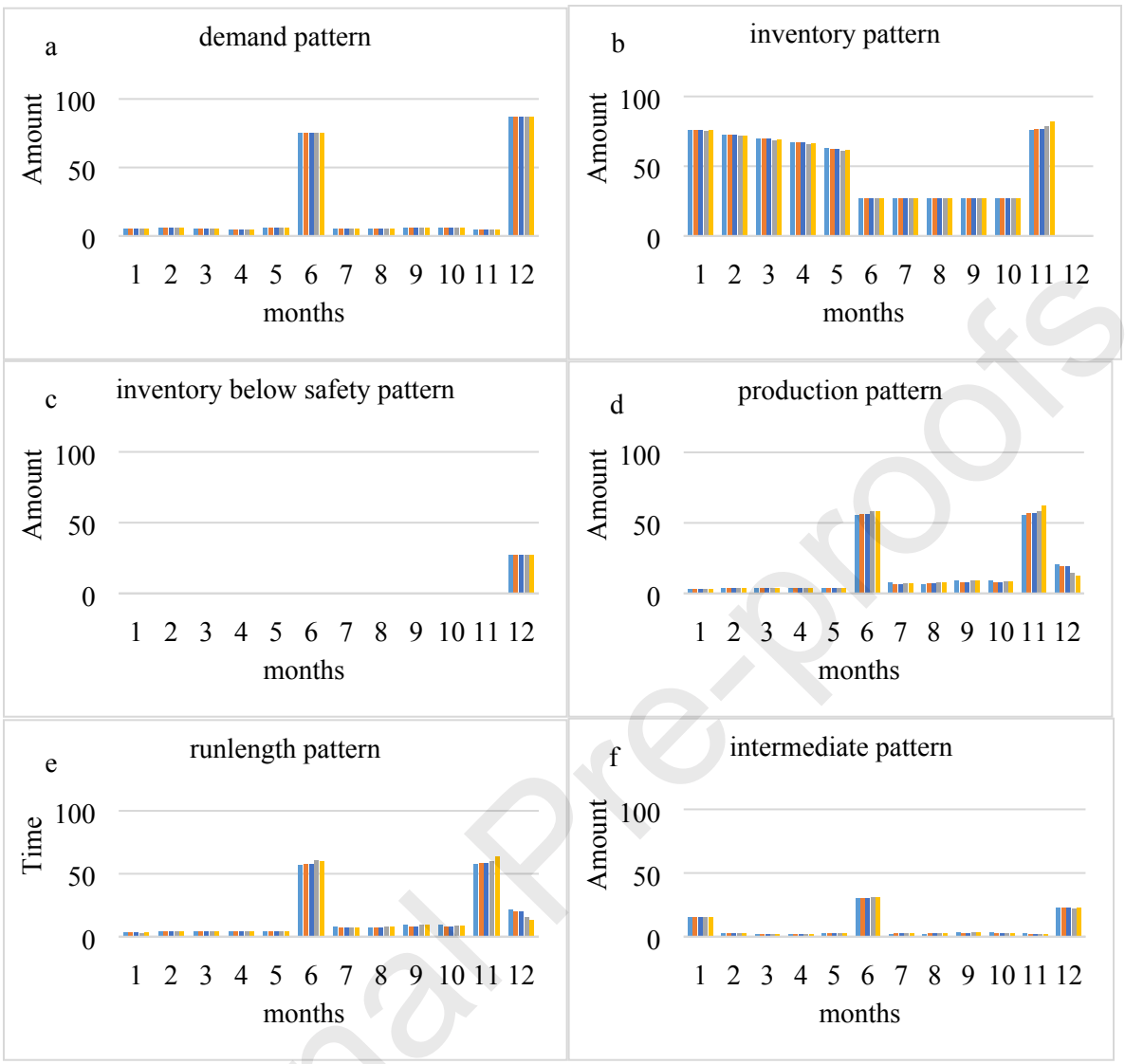


Fig. 8. a) Uncertain demand with 300% hike at 6th and 12th time periods and 20% at other time periods. The effect of uncertain demand parameter is shown on 5 sets of decision variables (moving left to right) – b) inventory, c) inventory below safety, d) production, e) run-length and f) intermediate pattern over horizon for product p_{18}, for 3000 data points in Case study 2. The blue, orange, indigo, grey and yellow bars represent the results for box, budget, ellipsoidal, proposed methodology and ideal solution respectively.

4.3 Example 1-Case 3

In this case, machine uptime is also considered uncertain along with the demand uncertainties. There are two machines, one at each manufacturing site. Along with the 34 products of the previous case (case 2), 2 more uncertain parameters one each for two of the machine uptimes give a total of 36 uncertain parameters spanned over 12 months. Eqs 32 to 36 are added in the deck of equations by modifying the Eqs 3, 4, 13, 14 and 15 with sub script *k* indicating instances of uncertain parameter realizations. The MILP problem formed is solved in GAMS® using CPLEX® solver. Further, the problem size is now increased to 132 binary variables, 13082170 single equations and 31396885 non zero elements and is executed in 53.11sec to find the optimized solution.

$$\sum_{p} T_{p,u,l,t} - MT_{u,l,t,k} \leq 0 \quad (32)$$

$$\sum_{f} T_{f,u,l,t} - MT_{u,l,t,k} \leq 0 \quad (33)$$

$$P_{p,u,l,t} \leq R_{p,u,l,t} MT_{u,l,t,k} \quad (34)$$

$$T_{p,u,l,t} - MT_{u,l,t,k} \leq 0 \quad (35)$$

$$T_{f,u,l,t} - MT_{u,l,t,k} B_{f,u,l,t} \leq 0 \quad (36)$$

Table 5. Comparison of cost function values obtained through box, budget, ellipsoidal and proposed method compared to ideal value for case study 4

Data points	Box	Box absolute deviation	Budget	Budget absolute deviation	Ellipsoidal	Ellipsoidal absolute deviation	DDRO	DDRO absolute deviation	Ideal value
500	23611.2	210.3946	23578.5	177.6496	23411.8	10.91358	22855.4	545.39	23400.84
1000	23664.5	263.6146	23647.9	247.0326	23598.9	198.0126	23071.9	328.96	
3000	24054.8	653.9336	24001.2	600.3786	23955.8	554.9816	23075.4	325.47	
5000	24249.09	848.2526	24109.75	708.9136	24002.42	601.5816	23168.92	231.92	
10000	24275.09	874.2526	24129.53	728.6906	24001.16	600.3196	23198.78	202.06	

The results are presented in Table 5. The differences among box, budget, ellipsoidal and the proposed algorithm for sampling uncertain space are clearly visible as the objective values obtained using the proposed sampling is found relatively closer to the ideal values. The proposed method is shown to underestimate the objective function in a minimization setup whereas the other methods are consistently shown to overestimate the solution. The reason for this trend remains the same as explained in the previous case study 1.

The patterns of decision variables in response to the variations in machine uptime and demands combined can be seen in Fig. 9. This figure shows the effect of adding machine uptime uncertainty with respect to the previous case study where there is no machine uptime consideration. The product p_{18} is an intermediate product. Hence the inventory pattern becomes valuable feature for final product formation from p_{18}. In response to the demand increase at 6th time period, inventories till 5th time period are used to meet the demand and hence no production can be seen till 4th time period as inventory cost is less compared to the production cost. Optimizer suggested some production at 5th time period, in case inventory is insufficient. Further, to meet the demand surge at 12th month, inventory till 11th month is used completely along with additional production at the same time period. No inventory and production at the 12th month have been observed to avoid unnecessary extra inventory and production cost. The inventory pattern for 5th time period shows the effectiveness of proposed method (grey bar) compared to budget (red bar) and box (blue bar) as grey bar is almost equal to yellow bar (ideal solution). Higher values appearing at the 5th and 11th month for the budget (orange bar) and the box (blue bar) approach indicate lack of accuracy in those methods for sampling the uncertain parameter space. These differences in production and inventory patterns visible in Fig. 9 compared to Fig. 8 are due to consideration of additional uncertainty in machine uptime. Inventory below safety can be seen at 12th time period as it was in the previous cases, but for 6th to 10th time periods, values are shown prominently large for the box approach. This action increases the penalty cost resulting in higher objective function values for this approach compared to the budget and the proposed DDRO method.

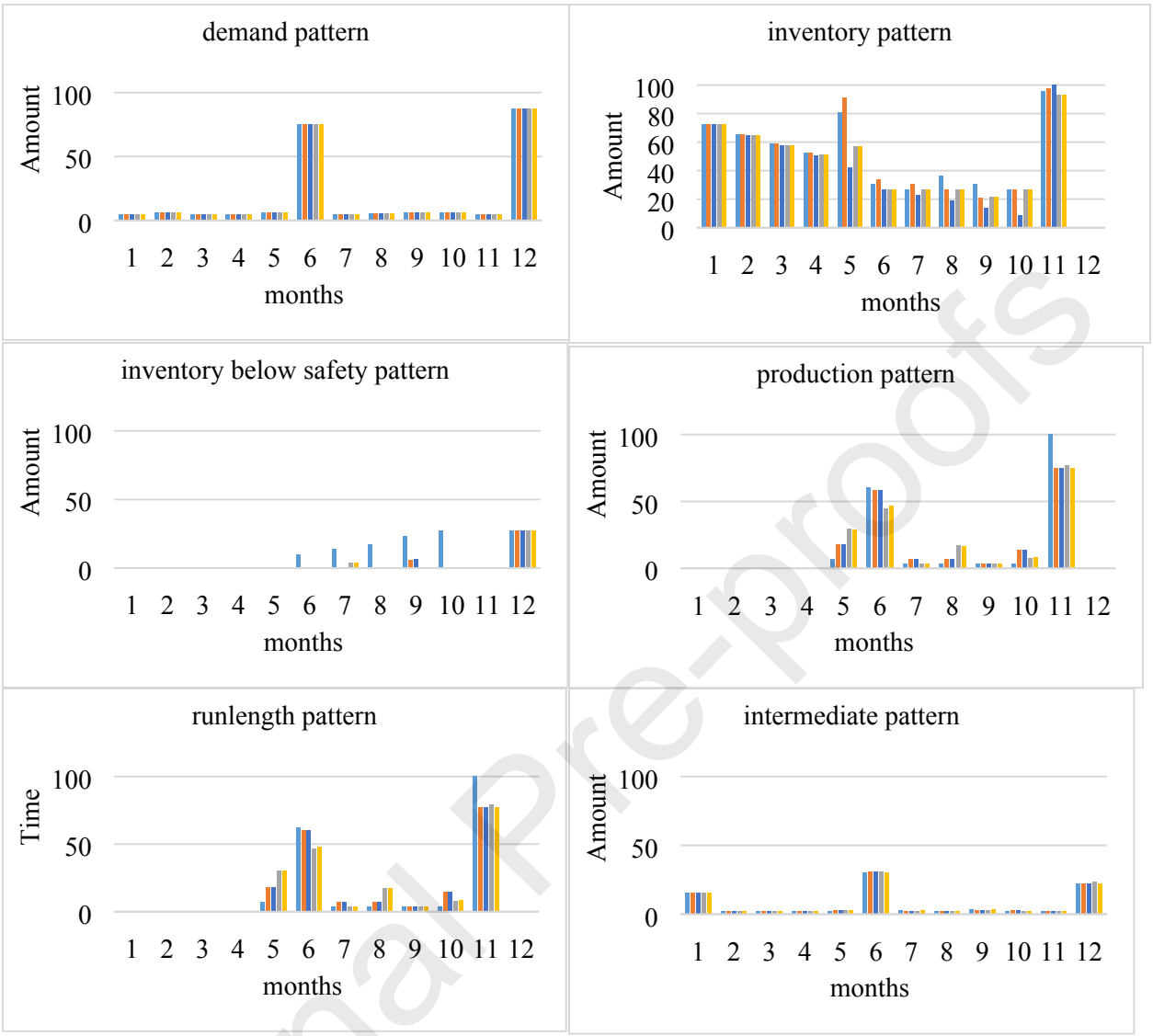


Fig. 9. a) Uncertain demand with 300% hike at 6th and 12th time periods and 20% at other time periods. The effect of uncertain demand parameter is shown on 5 sets of decision variables (moving left to right) – b) inventory, c) inventory below safety, d) production, e) run-length and f) intermediate pattern over horizon for product p_{18}, for 3000 data points in Case study 3. The blue, orange, indigo, grey and yellow bars represent the results for box, budget, ellipsoidal, proposed methodology and ideal solution respectively.

4.4 Example 1- Case 4

The final case study involves uncertainty in objective function by considering the production cost $VC_{p,u,l,k}$ as uncertain (see Eq 37 obtained from Eq 1 with *k* instances of uncertain data points). Product demand and machine uptime uncertainties are also included in the constrain functions. Hence, 34 uncertain parameter values are added due to uncertainty in production cost for all products along with 34 uncertain parameters from demand and 2 uncertain parameters for machine uptime at two production sites making a total of 70 uncertain parameters. There are 132 binary variables, 13082170 single equations and 31396885 non-zero elements in this formulation and it takes 54.67 sec to find the optimal solution. The entire model follows MILP formulation, but due to the usage of inbuilt functions *smin* for the worst-case implementation under

GAMS® environment, non-linearity gets introduced during implementation and this is solved using DICOPT® solver in GAMS®.

$$\sum_{f,u,l,t} FC_{f,u,l,t} B_{f,u,l,t} + \sum_{p,u,l,t} VC_{p,u,l,k} P_{p,u,l,t} + \sum_{p,l,t} RC_{p,l} C_{p,l,t} + \sum_{p,l,t} HC_{p,l,t} I_{p,l,t} + \sum_{p,l,c,t} TC_{l,c} S_{p,l,c,t} + \sum_{p,l,l',t} TC_{l,l'} C^{IP}_{p,l,l',t} + \sum_{p,l,t} PC_{p,l} I^{\Delta}_{p,l,t} + \sum_{p,c,t} \mu_{p,c} I^{-}_{p,c,t} \quad [37]$$

Table 6. Comparison of cost function values obtained through box, budget, ellipsoidal and proposed method compared to ideal value for case study 5

Data points	Box	Box absolute deviation	Budget	Budget absolute deviation	Ellipsoidal	Ellipsoidal absolute deviation	DDRO	DDRO absolute deviation	Ideal value
500	21390.0	190.4482	21354.8	155.3212	21247.2	47.65423	20460.5	739.0068	21199.5
1000	21485.7	286.2452	21502.8	303.3442	21356.4	156.8512	20668.3	531.2178	
3000	21860.9	661.4022	21858.4	658.8652	21741.2	541.7012	20907.3	292.2298	
5000	22042.86	843.3612	22017.47	817.9682	21894.46	694.9532	20982.37	217.1328	
10000	22079.74	880.2372	22047.18	847.6732	22004.26	804.7552	21009.86	189.6428	

Table 6 indicates that the cost function values obtained through the DDRO method are closer to the ideal values compared to those obtained by the box and the budget approach showing a similar trend as obtained in the previous case studies. The proposed algorithm gives value closer to the ideal value with increase in the number of data points used to sample the uncertain space. The values of cost function are increased gradually from case 1 to case 3 (see Table 3, 4 and 5), but decreased for case 4 due to the consideration of uncertainty in the production cost. The figures involving the SC parameters have similar patterns as those presented earlier.

4.4.1 Effect of uncertain data points and parameters on computation time.

From Table 7 one can observe that across the columns, as number of uncertain parameters are increased for each case study, the computation time also increase. Also, across the row, as number of data points increases from 500 to 10000 the computational time increases. Hence both i.e. number of data points and number of uncertain parameter have effect on computational time.

Table 7. Computational time for simulation runs for DDRO

	Case 1 demand uncertainty at market 1, p1 to p23 (sec)	Case 2 demand uncertainty at both markets, p1 to p34 (sec)	Case 3 demand + machine uptime uncertainty (sec)	Case 4 demand + machine uptime + production cost uncertainty (sec)
500	5.30	4.94	4.41	9.19
1000	7.88	7.78	9.75	13.77
3000	23.45	24.52	29.59	29.61
5000	42.88	46.59	53.11	54.67
10000	97.80	100.09	124.78	131.39

4.5 Example 2

Next, a continuous variable optimization under uncertainty problem expressed in terms of a set of algebraic equations (Eq 38 to 41) has been used to study the effect of proposed DDRO technique with respect to the box, budget and ellipsoidal sets. Here, there are three decision variables x_1,x_2 and x_3 in the decision variable set **x** and the uncertain parameter set **u** is composed of u_1 and u_2 with 20% variance. The minimization objective with constraint equations are solved using CPLEX® solver in GAMS®. The Table 8 shows the objective values for different uncertainty sets and various sampling points from 500 to 10000. It can be observed that for the proposed DDRO, the objective values are moving close to ideal value with increasing number of data points, whereas for box, budget, ellipsoidal sets, the values are deviating away. The reason for such behavior is same as explained in example 1 case study 2.

Cost function : $5x_1 + 3x_2 + 4x_3$ [38]

$$(1 + u_1 + 2u_2)x_1 + (1 - 2u_1 + u_2)x_2 + (2 + 2u_1)x_3 \leq 18 \quad [39]$$

$$(u_1 + u_2)x_1 + (1 - 2u_1)x_2 + (1 - 2u_1 - u_2)x_3 \leq 16 \quad [40]$$

$$-1 \leq u_1,u_2 \leq 1 \quad [41]$$

$$-19 \leq x_1 \leq 12 \quad [42]$$

$$-16 \leq x_2 \leq 0 \quad [43]$$

$$-3 \leq x_3 \leq 0 \quad [44]$$

Table 8. Comparison of cost function values obtained through box, budget, ellipsoidal and proposed method with respect to ideal value for case study example 2

Data points	Box	Box absolute deviation	Budget	Budget absolute deviation	Ellipsoidal	Ellipsoidal absolute deviation	DDRO	DDRO absolute deviation	Ideal value
500	-1428.25	89.598	-1469.87	47.973	-1498.17	19.678	-2161.27	643.421	-1517.85
1000	-1400.88	116.962	-1414.38	103.465	-1420.38	97.465	-1821.74	303.89	
3000	-1329.18	188.662	-1358.75	159.095	-1372.37	145.475	-1739.45	221.605	
5000	-1319.63	198.218	-1339.76	178.083	-1363.42	154.425	-1647.86	130.018	
10000	-1314.47	203.38	-1328.21	189.635	-1348.78	169.065	-1597.85	80.009	

4.5 Example 3

Here, a supply chain warehouse inventory problem is solved (Dantzig, 2016), where items are stocked to sell at a later date. The uncertainty **u** is kept in initial stock units $inistock_t$ which can vary between zero to hundred units. The warehouse can store maximum hundred units in each quarter indicating profit is to be gained when buying at low price and selling at higher price at appropriate time in four quarters annually. The objective function here is to minimize the total $cost$, by identifying the decision variable **x** i.e. number of buying buy_t, stocking $stock_t$ and selling $sell_t$ units at each quarter. The storage cost $storecost$ is kept \$1 per quarter per unit, with selling price $price_t$ of \$ 10, 12, 8, 9 in each quarter per

unit. The problem is solved in GAMS® using CPLEX® solver. From Table 9, it is observed that as the data points increases from 500 to 10000, the DDRO gives values closer to ideal values compared to the box, budget and ellipsoidal methods. The nature of solution obtained is similar to previous examples.

$$stock_t = stock_{t-1} + buy_t - sell_t + inistock_t \quad [45]$$

$$cost = \sum_t price_t(buy_t - sell_t) + stock_t\, storecost \quad [46]$$

$$0 \leq inistock_t, stock_t \leq 100 \quad [47]$$

$$0 \leq buy_t,\ sell_t \leq 1000 \quad [48]$$

Table 9. Comparison of cost function values obtained through box, budget, ellipsoidal and proposed method with respect to ideal value for case study example 3

Data points	Box	Box absolute deviation	Budget	Budget absolute deviation	Ellipsoidal	Ellipsoidal absolute deviation	DDRO	DDRO absolute deviation	Ideal value
500	-288.20	12.61	-289.36	11.45	-290.36	10.45	-315.39	14.58	-300.81
1000	-285.69	15.12	-285.94	14.87	-286.76	14.05	-313.56	12.75	
3000	-284.47	16.34	-284.73	16.08	-284.82	15.99	-312.13	11.33	
5000	-280.98	19.83	-281.54	19.27	-281.84	18.97	-309.95	9.14	
10000	-278.62	22.19	-278.85	21.96	-279.83	20.98	-306.68	5.87	

5 Conclusion

In this work, the mid-term supply chain planning problem under various SC parameter uncertainty has been solved using data driven robust optimization. The slot-based planning model of McDonald and Karimi (McDonald and Karimi, 1997) is adopted for constructing various uncertain scenarios and analyzing the effect of uncertain parameters on the planning model. The model has three echelons i.e. supplier, manufacturer and the market. Few products from one manufacturing unit acts as intermediate material for product generation at another manufacturing site. Supplier has two nodes supplying two manufacturing units and these two manufacturing units finally satisfy demand of two different market units respectively. Overall aim of the MILP SC model is to reduce the overall cost of operation of the SC with satisfaction of production, logistics, inventory and safety stock constraints. The uncertainty in SC parameters was introduced in demand, machine uptime and production cost to study the effect of stochasticity on the SC performance.

The proposed uncertainty handling method utilizes the power of machine learning algorithms to build data driven RO technique for identifying and sampling the uncertain parameter space more accurately to generate superior performance compared to the conservative results of conventional approaches. The data points in the uncertain space are first clustered (via NFCM based ANN+FCM) and their boundary points are marked using local density (LD) measure. The obtained boundary points are joined using Delaunay triangulation for creating outer envelope of each cluster, within which the Sobol sampling is used for generation of uncertain parameter realizations. The verification of proposed DDRO technique has been done, not just on supply chain model of (McDonald and Karimi, 1997), but also on other examples such as one from continuous decision variable domain and another supply chain warehouse inventory model (Dantzig, 2016). The examples clearly demonstrate the importance of efficient sampling in the uncertain

parameter space when the data points are less in number and scattered in nature, which might be difficult to approximate using known statistical distributions. Across all examples, it is observed that the increase in number of uncertain sampling data points (500 to 10000) improves closeness of objective function values to the ideal value for proposed data driven RO as compared to the box, budget and ellipsoidal methods due to the accurate transcription of uncertain parameter space by the proposed method. The method is also shown to be scalable for large number of uncertain parameters.

References

Abdelaziz, F. Ben, Aouni, B., Fayedh, R. El, 2007. Multi-objective stochastic programming for portfolio selection. Eur. J. Oper. Res. 177, 1811–1823. https://doi.org/10.1016/j.ejor.2005.10.021

Aghezzaf, E., 2005. Capacity planning and warehouse location in supply chains with uncertain demands. J. Oper. Res. Soc. 56, 453–462. https://doi.org/10.1057/palgrave.jors.2601834

Bertsimas, D., Sim, M., 2004. The price of robustness. Oper. Res. 52, 35–53. https://doi.org/10.1287/opre.1030.0065

Chernobai, A., Menn, C., T., S.R., C.Trûck, Moscadelli, M., 2006. Treatment of incomplete data in the field of operational risk: The effects on parameter estimates, EL and UL figures. The Advanced Measurement Approach to Operational Risk, pp. 145–468.

Dantzig, G., 2016. Linear Programming and Extensions, in: Linear Programming and Extensions. Princeton university press.

Dantzig, G.B., 1998. Linear programming and extensions. Princeton university press.

El Ghaoui, L., Oustry, F., Lebret, H., 1998. Robust solutions to uncertain semidefinite programs. SIAM J. Optim. 9, 33–52. https://doi.org/10.1137/S1052623496305717

Fortune, S., 1992. Voronoi diagrams and Delaunay triangulations. Comput. Euclidean Geom. 193–233. https://doi.org/10.1142/9789814355858_0006

Georgiadis, M.C., Tsiakis, P., Longinidis, P., Sofioglou, M.K., 2011. Optimal design of supply chain networks under uncertain transient demand variations. Omega 39, 254–272. https://doi.org/10.1016/j.omega.2010.07.002

Gorissen, B.L., Yanikoğlu, I., den Hertog, D., 2015. A practical guide to robust optimization. Omega (United Kingdom) 53, 124–137. https://doi.org/10.1016/j.omega.2014.12.006

Govindan, K., Fattahi, M., Keyvanshokooh, E., 2017. Supply chain network design under uncertainty: A comprehensive review and future research directions. Eur. J. Oper. Res. 263, 108–141. https://doi.org/10.1016/j.ejor.2017.04.009

Gregory, C., Darby-Dowman, K., Mitra, G., 2011. Robust optimization and portfolio selection: The cost of robustness. Eur. J. Oper. Res. 212, 417–428. https://doi.org/10.1016/j.ejor.2011.02.015

Guillén, G., Mele, F.D., Bagajewicz, M.J., Espuña, A., Puigjaner, L., 2005. Multiobjective supply chain design under uncertainty. Chem. Eng. Sci. 60, 1535–1553. https://doi.org/10.1016/j.ces.2004.10.023

Hammami, R., Temponi, C., Frein, Y., 2014. A scenario-based stochastic model for supplier selection in global context with multiple buyers, currency fluctuation uncertainties, and price discounts. Eur. J. Oper. Res. 233, 159–170. https://doi.org/10.1016/j.ejor.2013.08.020

Keyvanshokooh, E., Ryan, S.M., Kabir, E., 2016. Hybrid robust and stochastic optimization for closed-

loop supply chain network design using accelerated Benders decomposition. Eur. J. Oper. Res. 249, 76–92. https://doi.org/10.1016/j.ejor.2015.08.028

Long, Y., Lee, L.H., Chew, E.P., 2012. The sample average approximation method for empty container repositioning with uncertainties. Eur. J. Oper. Res. 222, 65–75. https://doi.org/10.1016/j.ejor.2012.04.018

McDonald, C.M., Karimi, I.A., 1997. Planning and Scheduling of Parallel Semicontinuous Processes. 1. Production Planning. Ind. Eng. Chem. Res. 36, 2691–2700. https://doi.org/10.1021/ie960901+

Mitra, K., 2009. Multiobjective optimization of an industrial grinding operation under uncertainty. Chem. Eng. Sci. 64, 5043–5056. https://doi.org/10.1016/j.ces.2009.08.012

Mitra, K., Gudi, R., Patwardhan, S., Sardar, G., 2008. Supply chain planning under uncertainty: A chance constrained programming approach. IFAC Proc. Vol. 17, 5501–5511. https://doi.org/10.3182/20080706-5-KR-1001.0511

Mohseni, S., Pishvaee, M.S., 2020. Data-driven robust optimization for wastewater sludge-to-biodiesel supply chain design. Comput. Ind. Eng. 139, 105944. https://doi.org/10.1016/j.cie.2019.07.001

Ning, C., You, F., 2018. Data-driven decision making under uncertainty integrating robust optimization with principal component analysis and kernel smoothing methods. Comput. Chem. Eng. 112, 190–210. https://doi.org/10.1016/j.compchemeng.2018.02.007

Ning, C., You, F., 2017. Data-driven adaptive nested robust optimization: general modeling framework and efficient computational algorithm for decision making under uncertainty. AIChE J. 63, 3817–3970. https://doi.org/10.1002/aic

Pantula, P.D., Miriyala, S.S., Mitra, K., 2020. An Evolutionary Neuro-Fuzzy C-means Clustering Technique. Eng. Appl. Artif. Intell. 89, 103435. https://doi.org/10.1016/j.engappai.2019.103435

Rajgopal, J., Wang, Z., Schaefer, A.J., Prokopyev, O.A., 2011. Integrated design and operation of remnant inventory supply chains under uncertainty. Eur. J. Oper. Res. 214, 358–364. https://doi.org/10.1016/j.ejor.2011.04.039

Reid, R.D., Sanders, N.R., 2019. Operations management: an integrated approach. John Wiley & Sons.

Santoso, T., Ahmed, S., Goetschalckx, M., Shapiro, A., 2005. A stochastic programming approach for supply chain network design under uncertainty. Eur. J. Oper. Res. 167, 96–115. https://doi.org/10.1016/j.ejor.2004.01.046

Shang, C., Huang, X., You, F., 2017. Data-driven robust optimization based on kernel learning. Comput. Chem. Eng. 106, 464–479. https://doi.org/10.1016/j.compchemeng.2017.07.004

Shapiro, A., 2011. Analysis of stochastic dual dynamic programming method. Eur. J. Oper. Res. 209, 63–72. https://doi.org/10.1016/j.ejor.2010.08.007

Simchi-Levi, D., Kaminsky, P., Simchi-Levi, E., 2004. Managing The Supply Chain: Definitive Guide. Tata McGraw-Hill Education.

Sobol', I.M., Asotsky, D., Kreinin, A., Kucherenko, S., 2011. Construction and Comparison of High-Dimensional Sobol' Generators. Wilmott 56, 64–79. https://doi.org/10.1002/wilm.10056

Soyster, A.L., 1973. Technical Note—Convex Programming with Set-Inclusive Constraints and Applications to Inexact Linear Programming. Oper. Res. 21, 1154–1157. https://doi.org/10.1287/opre.21.5.1154

Vallerio, M., Telen, D., Cabianca, L., Manenti, F., Van Impe, J., Logist, F., 2016. Robust multi-objective dynamic optimization of chemical processes using the Sigma Point method. Chem. Eng. Sci. 140, 201–216. https://doi.org/10.1016/j.ces.2015.09.012

Highlights

- Accurate transcription of uncertain parameter space using novel clustering algorithm
- More reliable estimation of statistical moments with less number of samples
- Machine learning based data driven robust optimization for sparse data sets
- Capability to handle large number of uncertain parameters

Kapil M Gumte: Methodology; Software; Formal Analysis; Investigation; Data curation; Roles/Writing – original draft; Visualization; Writing – review & editing

Priyanka Devi Pantula: Methodology; Data curation; Roles/Writing – original draft; Resources; Writing – review & editing

Miriyala Srinivas Soumitri: Investigation, Data curation, Resources, Writing – review & editing

Dr. Kishalay Mitra: Conceptualization; Validation; Resources; Writing – review & editing; Funding acquisition; Supervision; Project administration;

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: