Distributionally robust airline fleet assignment problem

Marco Graciotto Silva, Michael Poss

▶ **To cite this version:**

Marco Graciotto Silva, Michael Poss. Distributionally robust airline fleet assignment problem. INOC 2019 - 9th International Network Optimization Conference, Jun 2019, Avignon, France. pp.66-71, 10.5441/002/inoc.2019.13 . lirmm-02194250

HAL Id: lirmm-02194250

https://hal-lirmm.ccsd.cnrs.fr/lirmm-02194250

Submitted on 25 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Distributionally robust airline fleet assignment problem

Marco Silva
LIA, Université d'Avignon et des Pays du Vaucluse
Avignon
marco.costa-da-silva@univ-avignon.fr

Michael Poss
LIRMM, Université de Montpellier 2
Montpellier
michael.poss@lirmm.fr

ABSTRACT

In this work we consider the airline fleet assignment problem and we experiment with a robust solution where passenger demand is uncertain. To mitigate conservativeness of the classical robust optimization we consider a two-stage distributionally robust objective formulation. Our main contribution with respect to the airline fleet management problem literature lies in the modeling characteristics of our proposal.

We benchmark against current deterministic and robust fleet assignment formulations and verify solution performance results through simulation.

KEYWORDS

Distributionally robust optimization, Affine decision rules, Integer Programming, Airline fleet assignment

1 INTRODUCTION

Once an airline decides when and where to fly (flight legs) by developing its flight schedule, the next decision is determining the type of aircraft, or fleet, that should be used on each of the flight legs defined within the flight schedule. This process is called fleet assignment and its purpose is to assign fleet types to flight legs, subject to an available number of aircrafts and conservation of aircraft flow requirements, such as to maximize profits with respect to captured passenger demand. This decision needs to be made well in advance of departures when passenger demand is still highly uncertain. The factors that influence schedulers when assigning fleet types to various flights are: passenger demand, seating capacity, operational costs, and availability of maintenance at arrival and departure stations. One important requirement of the fleet assignment is that the aircraft must circulate in the network of flights. These so-called balance constraints are enforced by using time lines to model the activities of each fleet type. The period for which the assignment is done is normally one day for domestic flights.

Profit maximization is normally defined in terms of unconstrained revenue minus assignment cost. Unconstrained revenue of a flight leg is the maximum attainable revenue for that particular flight regardless of assigned capacity. Assignment cost, a function of the assigned fleet type, includes the flight operating cost, passenger carrying related cost and spill cost. Spill cost on a flight is the revenue lost when the assigned aircraft for that flight cannot accommodate every passenger. The result is that either the airline spills some passengers to other flights in its own network (in which case these passengers are recaptured by the airline), or they are spilled to other airlines.

In [12] the authors develop a two-stage stochastic programming model for integrated flight scheduling and fleet assignment where the fleet family assigned to each scheduled flight leg is decided at the first-stage. Then, the fleet type to assign to each flight leg is decided at the second-stage based on demand and fare realization. Sample average approximation (SAA) algorithm is then used to solve the problem and provide information on the quality of the solution.

In [9] the authors propose a new model based on itinerary grouping to mitigate the effect of demand uncertainty. Their itinerary group fleet assignment model deals with the difficulties caused by itinerary forecast by replacing them with aggregated demand forecasts. The authors affirm that an itinerary-based representation of demand (see Section 2 for details) has led to a high granularity of demand, making it hard to predict.

In this work, as an alternative to previous works presented, we propose a two-stage data-driven distributionally robust optimization model to address the question of airline robust planning for the fleet assignment problem. Our main contribution with respect to the airline fleet management problem literature lies in this novel modeling approach.

We adopt the concept of robust optimization as defined in [5] and [6] in that the demand uncertainty belongs to a known deterministic uncertainty set. In fact, we consider this uncertainty set as the support for the family of probability distributions associated with our random passenger demand parameter. We consider a data-driven approach by which this uncertainty set is constructed from available historical data. We assume that historical unconstrained (not subject to capacity issues) itinerary demand data is available and that we can use this historical data to predict future demand. By constructing the uncertainty set from historical data we are able to capture correlations between demands of different itineraries and thus mitigate the granularity demand effect as pointed out in [9].

On the other hand, we consider different modeling alternatives to mitigate conservatism of a robust approach. Since fleet assignment is a repetitive process, where fleet assignment decisions are made on daily basis, we mitigate the conservatism of the worst-case objective of classical robust optimization and consider a distributionally robust optimization approach on which we optimize the worst case expected performance on a set constituted by an infinite number of probability distributions, named ambiguity set (see [10] for main concepts). We also propose a two-stage model, as introduced in [4] where, although all the fleet assignments decisions are first stage, the calculation of lost revenue (spill) is only done after realization of uncertainty.

To facilitate handling large-scale fleet assignment problems, we propose the use of principal component analysis techniques to reduce dimension of the uncertainty set and the use of affine decision rules for our two-stage problem as approximations to improve time performance of our algorithms.

2 FLEET ASSIGNMENT FORMULATIONS

The fleet assignment model is typically formulated as a mixed-integer program. One can see the work in [16] for a survey of different modeling approaches for the problem.

© 2019 Copyright held by the owner/author(s). Published in Proceedings of the International Network Optimization Conference (INOC), June 12-14, 2019: Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.

In [1] the author first introduced for the fleet assignment problem a time-space network model to represent the availability of the fleet at each airport in the course of time. The proposed model resulted in a linear program that could either maximize profit or minimize operations cost.

In [11] the authors use the time-space network model and develop a large-scale integer program for fleet assignment. They propose several preprocessing techniques, namely node aggregation and isolated islands at stations, in order to reduce problem complexity.

In these two works demand is expressed for a specific flight leg and, therefore, these works do not capture demand dependencies between legs. This is because demand is defined for airline itineraries that can be comprised by multiple flight legs. Variations of demand in one itinerary flight leg will affect the others legs. This is called the network effect and it was taken into consideration in the model defined in [2]. There, the authors use the time-space network model and consider the effect of recapturing, where passengers spills from one itinerary can be redirected to alternative itineraries. In their model demand is deterministic.

The above model is reference for our work, with the difference that we do not consider recapturing. We replicate here the itinerary based formulation as presented in [9] where the authors also explicitly deal with itinerary fare classes to better capture the revenue dimension by favoring higher classes instead of considering all the fare classes at the same level. We present notations and formulation used.

Sets

P : the set of itinerary fare classes, indexed by p
A : the set of airports, indexed by o
L : the set of flight legs, indexed by i
K : the set of fleet types, indexed by k
T : the sorted set of all relevant event times (leg departures or aircraft availability) at all airports, indexed by t
$CL(k)$: the set of flight legs that pass the count time when flown by fleet type k
$I(k, o, t)$: the set of inbound flight legs to node (k, o, t)
$O(k, o, t)$: the set of outbound flight legs from node (k, o, t)

Decision variables

t_p : the number of passengers requesting itinerary fare class p and spilled by the model because of the capacity limit.
f_{ki} : binary variable equal to 1 if fleet type k is assigned to flight leg i, 0 otherwise.
y_{kot^+} : the number of fleet type k that are on the ground at airport o immediately after time t.
y_{kot^-} : the number of fleet type k that are on the ground at airport o immediately before time t. If t' is the time of the first event occurring after t, then $y_{kot} = y_{kot'^-}$

Data

$SEATS_k$: the number of seats available on aircraft of fleet type k.
c_{ki} : the cost of operating leg i with fleet type k.
N_k : the number of aircraft in fleet type k.
D_p : the unconstrained demand for itinerary fare class p.
$fare_p$: the fare class for itinerary p.
δ_i^p : a binary flag equal to 1 if itinerary fare class includes flight leg i, 0 otherwise.
$count\ time$: the time at which a snapshot of fleet utilization is taken to ensure consistency with the available fleet.
t_m : the last event before the $count\ time$, $t_m = count\ time^-$.

$(IFAM)$

$$\min \quad \sum_{i\in L, k\in K} c_{ki} f_{ki} + \sum_{p\in P} fare_p\, t_p \tag{1}$$

$$\text{s.t.} \quad \sum_{k\in K} f_{ki} = 1 \qquad \forall i \in L \tag{2}$$

$$\sum_{i\in I(k,o,t)} f_{ki} + y_{kot^-} = \sum_{i\in O(k,o,t)} f_{ki} + y_{kot^+}, \quad \forall k \in K, o \in A, t \in T \tag{3}$$

$$\sum_{o\in A} y_{kot_m} + \sum_{i\in CL(k)} f_{ki} \le N_k \qquad \forall k \in K \tag{4}$$

$$\sum_{p\in P} \delta_i^p D_p - \sum_{p\in P} \delta_i^p t^p \le \sum_{k\in K} f_{ki}\, SEATS_k \qquad \forall i \in L \tag{5}$$

$$t_p \le D_p \qquad \forall p \in P \tag{6}$$

$$f_{ki} \in \{0,1\}, y_{kot} \in \{0,1\}, t_p \ge 0, \quad \forall p \in P, k \in K, i \in L, o \in A, t \in T$$

The objective function (1) minimizes the total cost of operations plus the cost related to spilled itinerary fare class demand. This minimization is equivalent to profit maximization. Constraints (2) are the leg coverage constraints. Each flight leg has to be operated by exactly one aircraft type. The flow conservation constraint related to each single event is ensured through constraints (3). The limited size of each fleet is respected through constraints (4). The count time can be seen as a fixed time where a cut is applied on the network to ensure that the total aircraft of each fleet type k on the ground at all airports plus those flying at the time must not exceed the total aircraft N_k available for type k. The capacity constraints (5) ensure that satisfied demand fits with the number of seats available on any given leg. Last, constraints (6) ensure that spill does not exceed unconstrained demand for any given itinerary fare class.

We now propose a two-stage distributionally robust optimization formulation derived from $IFAM$ formulation to incorporate the random nature of passenger demand vector D. Distributionally robust optimization is an emerging and effective method to address the inexactness of probability distributions of uncertain parameters.

We formulate our problem assuming that passenger demand spill decision variable is a second-stage variable. This way passenger demand spill is only defined after realization of uncertainty and we represent this dependency defining it as a function map $t_p(D)$. We also assume the uncertainty of vector D is represented through a probability distribution $\mathbb{P}$ that belongs to a family of distributions $\mathcal{D}$.

We present the formulation developed.

(*DIFAM*)

$$\min \quad \sum_{i\in L, k\in K} c_{ki} f_{ki} + \sup_{\mathbb{P}\in\mathcal{D}} \mathbb{E}_{\mathbb{P}}[Q(f,D)] \tag{7}$$

s.t.

$$\sum_{k\in K} f_{ki} = 1 \qquad \forall i \in L \tag{8}$$

$$\sum_{i\in I(k,o,t)} f_{ki} + y_{kot^-} = \sum_{i\in O(k,o,t)} f_{ki} + y_{kot^+}, \quad \forall k \in K, o \in A, t \in T \tag{9}$$

$$\sum_{o\in A} y_{kot_m} + \sum_{i\in CL(k)} f_{ki} \le N_k \qquad \forall k \in K \tag{10}$$

$$f_{ki} \in \{0,1\}, y_{kot} \in \{0,1\}, \quad \forall i \in L, k \in K, o \in A, t \in T$$

where

$Q(f,D) =$

$$\min \quad \sum_{p\in P} fare_p\, t_p(D) \tag{11}$$

$$\sum_{p\in P} \delta_i^p D_p - \sum_{p\in P} \delta_i^p t^p(D) \le \sum_{k\in K} f_{ki} SEATS_k \qquad \forall i \in L \tag{12}$$

$$t_p(D) \le D_p \qquad \forall p \in P \tag{13}$$

$$t_p \ge 0, \forall p \in P$$

The cost $\sum_{i\in L, k\in K} c_{ki} f_{ki}$ incurred during the first stage is deterministic. In progressing to the second-stage, the random passenger demand vector D is realized. We can then determine the cost incurred at the second-stage. For a given first stage fleet type assignment decision, f, and a realization of the random passenger demand vector, D, we evaluate the second-stage cost via the linear optimization problem, $Q(f,D)$. Since the fleet type assignment is a repetitive daily process and the true probability distribution of D is unknown and belong to a family of distributions set $\mathcal{D}$ we are interested in the worst case expectation $\sup_{\mathbb{P}\in\mathcal{D}} \mathbb{E}_{\mathbb{P}}[Q(f,D)]$. Note that it is a relatively complete recourse problem because any first-stage solution leads to a feasible second-stage solution.

In order to be able to deal with large scale problems, our two-stage distributionally robust optimization formulation must admit a tractable reformulation. The reformulation is closely related to the choices of ambiguity set that we make. On the other hand these choices must correctly reflect properties of historical data available.

In the next section we show that defining ambiguity sets by linear relationships of uncertainty parameters and approximating second-stage variables as affine functions of uncertainty parameters yields a tractable problem. We will also use techniques of uncertainty dimensionality reduction as a further compromise between optimality and tractability.

3 TWO-STAGE DISTRIBUTIONAL REFORMULATION

3.1 Dimensionality reduction

In real case examples, the dimension of the random passenger demand vector D can be in the range of thousands of itineraries. This can impact performance of the formulation *DIFAM*. Employing dimensionality reduction techniques to reduce the number of random variables under consideration can improve performance of our formulations.

Here we assume there is a set $\mathcal{W}'$ of N demand data samples available, $\mathcal{W}' = \{D^{(i)}\}_{i=1}^N$, based on historical data, and use this set to calculate the mean vector $\bar{D}$, variance vector $\hat{D}$ and covariance matrix $cov(D^{(i)})$.

A linear technique for dimensionality reduction, principal component analysis, performs a mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. Intuitively, we change the system of coordinates and define this system by new vectors Y, but we select only some of them, therefore reducing dimension of the system. The new system of coordinates, vectors $\{Y^c\}_{c=1}^C$, are in fact normalized eigenvectors of the covariance matrix $cov(D^{(i)})$, where c is the index of the selected eigenvectors. We refer to [17] for more details on principal component analysis (PCA).

We execute a procedure to express the passenger demand vectors, $D^{(i)}$, in the new system of coordinates, but before we normalize the vectors $D^{(i)}$, using $\bar{D}$ and $\hat{D}$. Therefore we define $D^{(i)'} = (D^{(i)} - \bar{D})./\hat{D}$, where ./ is a component wise division of vectors.

We then compute the coordinates, $X_c^{(i)}$, in the system of coordinates of the principal components vectors, $\{Y^c\}_{c=1}^C$, where the principal component vector has dimension $|P|$. The value of $X_c^{(i)}$ results from the expression:

$$X_c^{(i)} =< D^{(i)'}, Y^c >, \tag{14}$$

where $<,>$ is a dot product.

In the following we need the random vector to be nonnegative, which may not be the case of components $X_c^{(i)}$. Hence, we introduce a new random vector $\xi^{(i)}$ where each component will vary in the nonnegative interval $[0,1]$. Each component of $\xi^{(i)}$ is defined as

$$\xi_c^{(i)} = (X_c^{(i)} - \max_i (X_c^{(i)}))/(\min_i (X_c^{(i)}) - \max_i (X_c^{(i)})), \tag{15}$$

where $\max_i (X_c^{(i)})$, $\min_i (X_c^{(i)})$ are, respectively, the maximum and minimum projection component values along each vector Y^c considering all instances, $i \in \{1,\ldots,N\}$. $X_c^{(i)}$ varies in the interval $[\min_i (X_c^{(i)}), \max_i (X_c^{(i)})]$ and as consequence $\xi_c^{(i)}$ will vary in the interval $[0,1]$.

Using the above definition of $\xi^{(i)}$, we can define the components of each demand vector $D^{(i)}$ as

$$D_p^{(i)} = D1_p + \sum_{c=1}^{C} D2_{pc} \xi_c^{(i)}, \tag{16}$$

where

$$D1_p = \bar{D}_p + \hat{D}_p \sum_{c=1}^{C} \min_i (X_c^{(i)}) Y_p^c, \tag{17}$$

$$D2_{pc} = \hat{D}_p (\max_i (X_c^{(i)}) - \min_i (X_c^{(i)})) Y_p^c \tag{18}$$

This is an important step in order to guarantee positive definite matrices in the algorithm developed in Section 4 for our distributionally robust ambiguity set.

3.2 Ambiguity set and first-order deviation moment functions

The tractability of a distributionally robust linear optimization problem is dependent on the choice of the ambiguity set. Several ambiguity sets have been proposed in the literature. In particular, moment-based uncertainty sets assume that all distributions in the distribution family share the same moment information. By leveraging conic duality many distributionally robust optimization problems with moment-based ambiguity sets can, in general, be reformulated equivalently as convex problems. Although these problems can be solved theoretically in polynomial time, they are not efficient for large-scale instances.

In [8], a moment-based second-order conic representable ambiguity set, $\mathcal{D}$, is defined as

$$\mathcal{D} = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathbb{R}^{|P|}) \;\middle|\; \begin{array}{c} D \in \mathbb{R}^{|P|} \\ \mathbb{E}_{\mathbb{P}}[GD] = \mu \\ \mathbb{E}_{\mathbb{P}}[g_i(D)] \leq \gamma_i \quad \forall i \in I \\ \mathbb{P}(D \in U) = 1 \end{array} \right\}.$$

We assume random passenger demand vector D, but the same results can be derived substituting for dimensional reduced random vector ξ derived in the previous section. $\mathcal{P}_0(\mathbb{R}^{|P|})$ represents the set of all probability distributions in $\mathbb{R}^{|P|}$ and new parameters are defined as $G \in \mathbb{R}^{n_1 \times |P|}$, $\mu \in \mathbb{R}^{n_1}$, $\gamma \in \mathbb{R}^{|I|}$, SOC (second-order conic) representable support set $U \in \mathbb{R}^{|P|}$ and SOC representable functions $g_i \in \mathbb{R}^{|P| \times 1}$.

$\mathcal{D}$ only contains valid distributions supported over the support set U and moment information of uncertainties are characterized via functions g_i. The equality expectation expression allow the modeler to specify the mean values of D.

The authors of [8] further reformulate the ambiguity set $\mathcal{D}$ as a projection of an extended ambiguity set $\bar{\mathcal{D}}$ by introducing an I-dimensional auxiliary random vector u in

$$\bar{\mathcal{D}} = \left\{ \mathbb{Q} \in \mathcal{P}_0(\mathbb{R}^{|P|} \times \mathbb{R}^{|I|}) \;\middle|\; \begin{array}{c} (D,u) \in \mathbb{R}^p \times \mathbb{R}^{|I|} \\ \mathbb{E}_{\mathbb{Q}}[GD] = \mu \\ \mathbb{E}_{\mathbb{Q}}[u] \leq \gamma_i \quad \forall i \in I \\ \mathbb{P}((D,u) \in \bar{U}) = 1 \end{array} \right\}$$

where $\bar{U}$ is the lifted support set defined as

$$\bar{U} = \left\{ (D,u) \in \mathbb{R}^{|P|} \times \mathbb{R}^{|I|} \;\middle|\; \begin{array}{c} D \in U \\ g_i(D) \leq u_i \quad \forall i \in I \end{array} \right\}$$

They observe that the lifted ambiguity set has only linear expectation constraints and show that the adaptive distributionally robust optimization problem can be reformulated as a classical robust optimization problem with uncertainty set $\bar{U}$.

To be able to reformulate adequately our fleet assignment problem *DIFAM* we must then define an ambiguity set $\mathcal{D}$ that will lead to a polyhedron lifted support set $\bar{U}$.

In [15] the authors define first-order deviation moment-based functions $g_i(.)$ that are second-order conic representable as piecewise linear functions

$$g_i(D) = \max\{h_i^T D - q_i, 0\} \quad \forall i \in I.$$

They can be understood as the first-order deviation of uncertain parameters along a certain projection h_i truncated at q_i. We apply these moment-based functions to our problem and also assume that the support set U is a polyhedron. We then define our ambiguity set $\mathcal{D}$ as

$$\mathcal{D} = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathbb{R}^{|P|}) \;\middle|\; \begin{array}{c} D \in \mathbb{R}^{|P|} \\ \mathbb{E}_{\mathbb{P}}[\max\{h_i^T D - q_i, 0\}] \leq \gamma_i \quad \forall i \in I \\ \mathbb{P}(D \in U) = 1 \end{array} \right\}$$

and the lifted support set $\bar{U}$ will be a polyhedron given as

$$\bar{U} = \left\{ (D,u) \in \mathbb{R}^{|P|} \times \mathbb{R}^{|I|} \;\middle|\; \begin{array}{c} D \in U \\ 0 \leq u_i \quad \forall i \in I \\ h_i^T D - q_i \leq u_i \quad \forall i \in I \end{array} \right\}$$

3.3 Affine decision rules

With the above definition of lifted support set we can apply, to our *DIFAM* formulation, the reformulation proposed by [8] for the adaptive distributionally robust optimization problem, approximating second-stage variables t_p as affine functions of the lifted support set parameters (D,u), $t_p(D,u) = t_p^0 + \sum_{i \in P} t_{pi}^1 D_i + \sum_{i \in I} t_{pi}^2 u_i$.

This reformulation is based on the dualization of the inner problem of *DIFAM*, $\sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[Q(f,D)]$, and by introducing Lagrangian multipliers r and β to it (alternatively see [15] for a summarized proof of this reformulation). This leads to the following classical robust optimization problem:

(*RRIFAM*)

$$\min \quad \sum_{i \in L, k \in K} c_{ki} f_{ki} + r + \sum_{i \in I} \gamma_i \beta_i$$

s.t.

$$r + \sum u_i \beta_i \geq \sum_{p \in P} fare_p t_p(D,u), \quad \forall (D,u) \in \bar{U}$$

$$\sum_{p \in P} \delta_i^p D_p - \sum_{p \in P} \delta_i^p t^p(D,u) \leq \sum_{k \in K} f_{ki} SEATS_k, \quad \forall i \in L, \forall (D,u) \in \bar{U}$$

$$t_p(D,u) \leq D_p, \quad \forall (D,u) \in \bar{U}$$

$$\sum_{k \in K} f_{ki} = 1, \quad \forall i \in L$$

$$\sum_{i \in I(k,o,t)} f_{ki} + y_{kot^-} = \sum_{i \in O(k,o,t)} f_{ki} + y_{kot^+}, \quad \forall k \in K, o \in A, t \in T$$

$$\sum_{o \in A} y_{kot_m} + \sum_{i \in CL(k)} f_{ki} \leq N_k, \quad \forall k \in K$$

$$t_p(D,u) = t_p^0 + \sum_{i \in P} t_{pi}^1 D_i + \sum_{i \in I} t_{pi}^2 u_i, \quad \forall p \in P$$

$$r \in \mathbb{R}, \beta_i \geq 0, f_{ki} \in \{0,1\}, y_{kot} \in \{0,1\}, t_p \geq 0, \quad \forall p \in P, k \in K, i \in L, o \in A, t \in T$$

4 DATA-DRIVEN AMBIGUITY SET

A desirable ambiguity set should flexibly adapt to the intrinsic structure behind real data, thereby well characterizing $\mathbb{P}$ and attempting to reduce natural conservatism of robust solutions. In face of complicated distributional geometry, making prior

assumptions on the form of probability distribution or using classical uncertainty sets to describe their support have limited modeling power. With this in mind we adopt a data-driven methodology to construct and define parameters of the support set and moment-based functions associated with our ambiguity set.

4.1 Support Set

In what follows, we use the technical approach of [14] to construct a support set U from data samples of the random variable ξ. We assume there is a set $\mathcal{W}$ of N data samples available, $\mathcal{W} = \{\xi^{(i)}\}_{i=1}^{N}$, and this set is constructed from the sample of demand vectors, $D^{(i)}{}_{i=1}^{N}$, as explained in Subsection 3.1.

In [14] the authors propose piecewise linear kernel-based support vector clustering (SVC) as a machine learning technique tailored to data-driven robust optimization. They explore the SVC's secondary effect that evolves the data samples inside a sphere in a high-dimensional space [3]. They use this sphere to characterize the uncertainty set. This mapping of data points to a high-dimensional space is done by means of a kernel function. Using well known techniques of machine learning they define a linear kernel that, in turn, is used to define a polyhedral region evolving the data in the original space.

Using these techniques, we define our data-driven support set U as the region inside or in the borders of this sphere. This sphere is given by the expression

$$U = \{\xi \mid K(\xi,\xi) - 2\sum_{i=1}^{N} \alpha_i K(\xi,\xi^{(i)}) + \\ + \sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i\alpha_k K(\xi^{(i)},\xi^{(k)}) \leq R^2\}$$

Parameters α and R are derived by applying Lagrangian relaxation to the original formulation and the linear kernel is given by

$$K(\xi^{(i)},\xi^{(j)}) = \sum_{k=1}^{N} l_k - |\xi^{(i)} - \xi^{(j)}|_1,$$

where $l_k = \max_{1\leq i\leq N} \xi_k^{(i)} - \min_{1\leq i\leq N} \xi_k^{(i)}$. We refer to [3] for details.

4.2 Moment-based functions

For moment-based functions we adopt the work of [15] where the authors define a two-step procedure for determining parameters h_i and q_i of our piecewise linear functions in order to capture meaningful information from available data.

The directions h_i are based on principal component analysis (PCA) such that the data space becomes decorrelated along each direction and the information overlap between different directions is slight. Since our random vector ξ already comprises decorrelated components we adopt vector h_i as standard unit vectors e_i.

After that, several truncation points $\{q_i\}$ are set along each direction h_i. For each direction h_i we choose $2J+1$ well-distributed truncation points. The first truncation point is set as the mean value $\bar{\xi}_i$ and the remaining $2J$ ones around the mean $\bar{\xi}_i$ symmetrically based on a fixed step-size given as the variance $\hat{\xi}_i$ along the i-th direction.

In this way, we will have $C(2J+1)$ piecewise functions $g_i(.)$ in total in the ambiguity set.

Intuitively, the parameter J can be deemed as the "size" of the ambiguity set, which can be manipulated to adjust the conservatism of the model. The more truncation points we have, the more statistical information will be incorporated, which leads to a smaller ambiguity set as well as a less conservative solution.

After determining the value of h_i and q_i, the next step is to estimate the parameters γ_i empirically based on N available data samples:

$$\gamma_i = \frac{1}{N}\sum_{j=1}^{N} \max(h_i^T \xi^{(j)} - q_i, 0)$$

Intuitively, with the values of size parameters γ_i increasing, the DRO model becomes more conservative.

5 IMPLEMENTATION AND RESULTS

5.1 Implementation details

We report on experiments conducted with the formulations proposed for the airline fleet assignment problem. Our objective is to verify the performance of each solution in a long run operation since fleet assignment is a daily repetitive process.

For our purposes, we create a small-sized hub-and-spoke airline instance, in which a unique major airport serves as a central point for coordinating flights to and from other airports. This way all our itineraries are composed of a maximum of two flight legs. We consider a structure of 9 airports, 3 fleet types and 24 daily itineraries based on three fare classes. A flight schedule with 21 flight legs is created and they are used to compose the daily itineraries.

We test this operation under four different problem formulations: *IFAM*, *RRIFAM*, as already presented in this study and two other formulations *RIFAM* and *RIFAM2*. Formulation *RIFAM* is a standard two-stage robust formulation where the objective is given as the worst case performance and dimensionality reduction is performed the same way as for *RRIFAM*. Formulation *RIFAM2* is the same as *RIFAM* where no dimensionality reduction is performed.

We randomly generate a set of 400 demand vectors. They are designed in a way that many itinerary demands are highly correlated.

We use 100 demand vectors as historical training data to create the ambiguity set of formulation *RRIFAM* and the 300 others to simulate the airline operating period.

We use naive approaches to determine demand vectors for formulations *IFAM*, *RIFAM* and *RIFAM2*. For formulation *IFAM* we consider three demand scenarios of low, medium and high total demand and consider the average of these three scenarios as input to our *IFAM* formulated problem. For formulation *RIFAM* and *RIFAM2* we consider maximum and minimum demand values for each leg and consider a box uncertainty set where each demand component varies within this interval.

With the solution of formulations *IFAM*,*RIFAM* and *RRIFAM* we simulate an airline operating period of 300 days and calculate an objective of total operating costs plus total loss revenue. We compare simulation results of the three formulations where our focus is on analyzing objective value and time performance.

Conservatism regulation parameters of our ambiguity set are fixed as $v = 0.6$ and $J = 0$. With $v = 0.6$, 100% of demand vectors were considered inside or in the border of the support set (no outliers). We calculate parameter C so that the sum of variances in the direction of each principal component considered sums

	IFAM	*RIFAM*	*RIFAM2*	*RRIFAM*
Objective value	335747	638817	651081	488135
Total Time (s)	4.5	239.77	10950.47	9041.14
Number of variables	789	982	1366	1181
Number of constraints	450	-	-	-
Number of iterations	-	32	49	81
Simulation Total cost*	1.38e8	1.35e8 (2.2%)	-	1.32e8 (4.3%)

*In parenthesis percentage gain when compared to worst result

Table 1: Implementation and performance comparison of fleet assignment formulations

up to a minimum of 90% of the sum of variances considering all principal components. For the instance we created, $C = 8$ of 24.

To solve formulations *RIFAM*, *RIFAM*2 and *RRIFAM* we use a master and adversarial problem approach where, at each iteration, we use the adversarial problems to search for a demand scenario instance that invalidates the master problem solution. See [7] for more details on this solution approach.

Algorithms were coded in Julia [13] using JuMP and Cplex 12.7. All algorithms were run in an Intel CORE i7 CPU 3770 machine.

5.2 Comparative performance of the formulations

Table 1 presents the results of the implementation and solution for the four different formulations. The relation between objective values are as expected since formulation *IFAM* is optimizing against a specific demand scenario, formulations *RIFAM* and *RIFAM*2 are optimizing against a worst case scenario and formulation *RRIFAM* is optimizing an expected performance (worst-case). *RIFAM* is designed to be a lower bound of *RIFAM*2 since it considers less constraints (restricted uncertainty set), but the results show that *RIFAM* is a reasonable approximation of *RIFAM*2. Since we use affine decision rules, formulations *RIFAM*, *RIFAM*2 and *RRIFAM* are themselves upper bound approximations of the true optimal worst-case or worst-case expected performance. Since we use auxiliary variables to compose affine decision rules for formulation *RRIFAM*, it leads to more flexible results than affine decision rules use original demand uncertainty.

The total time performance result is in direct link with the number of variables of each formulation, although the number of iterations for each of the robust formulations varies. In terms of time performance, dimensionality reduction has been effective to reduce total time. On the other hand, since the size of our airline instance is small, additional measures should be put in place to be able to deal with real large airline instances.

The simulation results are also as expected since the formulation *RRIFAM*, in the long run, leads to the less costly total solution. We note that there are no guarantees, in terms of the mathematical model proposed, on how formulations *IFAM* and *RIFAM* would perform in the long simulation run. We also note that formulation *RRIFAM* is an approximation of the true optimal result. Even though we would expect that, in the long simulation run, result of worst-case expected performance of formulation *RRIFAM* would out perform the two other formulations, and that is the case.

6 CONCLUSION

Initial computational results have shown that our proposed model can improve over other more traditional approaches. A further study can analyze the quality of the approximations performed, using real life data and comparing with data-driven stochastic optimization approximation algorithms.

REFERENCES

[1] Abara, J. Applying integer linear programming to the fleet assignment problem. *Interfaces 19*, 4 (1989), 20–28.

[2] Barnhart, C., Kniker, T. S., and Lohatepanont, M. Itinerary-based airline fleet assignment. *Transportation Science 36*, 2 (2002), 199–217.

[3] Ben-Hur, A., Horn, D., Siegelmann, H. T., and Vapnik, V. Support vector clustering. *J. Mach. Learn. Res. 2* (2002), 125–137.

[4] Ben-Tal, A., Goryashko, A., Guslitzer, E., and Nemirovski, A. Adjustable robust solutions of uncertain linear programs. *Mathematical Programming 99*, 2 (2004), 351–376.

[5] Ben-Tal, A., and Nemirovski, A. Robust convex optimization. *Math. Oper. Res. 23*, 4 (1998), 769–805.

[6] Ben-Tal, A., and Nemirovski, A. Robust solutions of uncertain linear programs. *Operations Research Letters 25*, 1 (1999), 1–13.

[7] Bertsimas, D., Dunning, I., and Lubin, M. Reformulation versus cutting-planes for robust optimization. *Computational Management Science 13*, 2 (Apr 2016), 195–217.

[8] Bertsimas, D., Sim, M., and Zhang, M. Adaptive distributionally robust optimization. *Management Science* (2018), online.

[9] Boudia, M., Delahaye, T., Gabteni, S., and Acuna-Agost, R. Novel approach to deal with demand volatility on fleet assignment models. *Journal of the Operational Research Society 69*, 6 (2018), 895–904.

[10] Delage, E., and Ye, Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research 58*, 3 (2010), 595–612.

[11] Hane, C. A., Barnhart, C., Johnson, E. L., Marsten, R. E., Nemhauser, G. L., and Sigismondi, G. The fleet assignment problem: Solving a large-scale integer program. *Math. Program. 70* (1995), 211–232.

[12] Kenan, N., Jebali, A., and Diabat, A. An integrated flight scheduling and fleet assignment problem under uncertainty. *Computers and Operations Research 100* (2018), 333 – 342.

[13] Lubin, M., and Dunning, I. Computing in Operations Research using Julia. *CoRR abs/1312.1431* (2013).

[14] Shang, C., Huang, X., and You, F. Data-driven robust optimization based on kernel learning. *Computers and Chemical Engineering 106* (2017), 464 – 479.

[15] Shang, C., and You, F. Distributionally robust optimization for planning and scheduling under uncertainty. *Computers and Chemical Engineering 110* (2018), 53 – 68.

[16] Sherali, H. D., Bish, E. K., and Zhu, X. Airline fleet assignment concepts, models, and algorithms. *European Journal of Operational Research 172* (2006), 1–30.

[17] WOLD, S., ESBENSEN, K., and GELADI, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems 2* (1987), 37–52.