**Les Cahiers du GERAD** ISSN: 0711–2440

# A stochastic approach to reoptimizing air cargo shipping plans

P. Zago, P. Munroe,
I. El Hallaoui, F. Soumis

G–2019–87

November 2019

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2019
– Bibliothèque et Archives Canada, 2019

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2019
– Library and Archives Canada, 2019

**GERAD** HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

**Tél. : 514 340-6053**
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

# A stochastic approach to reoptimizing air cargo shipping plans

**Paul Zago** [a,b]

**Patrick Munroe** [a,b]

**Issmail El Hallaoui** [a,b]

**François Soumis** [a,b]

[a] *GERAD, Montréal (Québec), Canada, H3T 2A7*

[b] *Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal (Québec) Canada, H3C 3A7*

paul.zago@polymtl.ca
patrick.munroe@gerad.ca
issmail.elhallaoui@gerad.ca
francois.soumis@gerad.ca

**November 2019**
**Les Cahiers du GERAD**
**G–2019–87**

**Abstract:** Overbooking is a common practice in the air cargo industry because booked and actual demands often differ greatly. As a consequence, in case of excessive overbooking, the overflowing commodities must be reassigned to other flights on the day of departure. We propose a stochastic approach to reoptimizing cargo shipping plans so that the expectation of cargo overflow is minimized. A probability distribution consisting of the mixture of a normal and a constant 0 distribution is introduced to model the show-up rate so that no-show is taken into account. From this mixture distribution, we derive an analytic expression for the expectation of overflow. An approximation of this expression is then used as the objective function of the formulation of an integer programming model to reoptimize an air cargo shipping plan. In a series of simulation experiments, we compare the resulting analytical model with an equivalent scenario-based stochastic linear program. It is found that the analytical model delivers results that are similar to the ones of the scenario-based stochastic program, but with a significant reduction in computation time.

**Keywords:** Transportation, stochastic programming, air cargo, overbooking

**Acknowledgments:** This project was funded by the Canadian Natural Science and Engineering Research Council (NSERC), Prompt, and Air Canada under grant RDCPJ 537140–18; and Institute for Data Valorisation (IVADO).

# 1 Introduction

Over the last decades, the air cargo industry has experienced a strong growth and is still expected to grow at an annual rate of 4.3 % through 2037 (Boeing (2016)) because of globalization. Moreover, despite representing less than 1% of world trade shipments by volume, air cargo accounts for around 35% of them by value (Shepherd et al. (2016)).

Air passenger transportation has already been studied extensively by the scientific community. However, there exist many differences between the air passenger and cargo industries. As a result, the methods developed for planning passenger transportation can generally not be applied directly to the world of air cargo transportation. Kasilingam (1997a) discusses the main differences:

1. Flight capacity is unknown until a few hours before departure. Indeed, since cargo is often stored in the baggage hold of passenger airplanes, a part of the available space is already used by luggage.
2. Many dimensions must be taken into account. Because pieces of cargo have specific weights and volumes, capacity in terms of volume can be reached before weight capacity, and vice versa.
3. The number of available routes is much greater. Indeed, contrary to passengers, cargo will (in general) not suffer from detours or from a larger number of connections.
4. A large part of the capacity is kept for *allotments*, i.e., some clients have special contracts that allow them to use a reserved block of capacity over a specific period of time.

Another important difficulty related to planning air cargo transportation is that no penalty is generally incurred by a client that books a certain capacity but does not use it all. This has for effect that clients often book more capacity than actually required or simply do not show up at the airport with their goods on the day of departure. This last phenomenon is called a *no-show*. Moreover, some of the demands correspond to allotments with *freight forwarders*, i.e., companies that organize the shipment for individuals or other companies. At the time of signing the contract, forwarders usually do not know exactly how much space they will need for their clients. As a result, the capacity booked might differ greatly from what is used. The (weight) percentage of the demand that is actually brought to the airport on the day of departure is called the *show-up rate*. In order to compensate for this potential waste of capacity, airfreight carriers resort to *overbooking*, i.e., they sell more capacity than is actually available.

When a client places an order for delivery with an airfreight carrier, it is often an employee of this airline that decides, along with the client, by which *route*, i.e., series of flights, the commodities will travel from the origin to the destination. Taken together, these route assignments form what we call a *shipping plan*. Given the presence of no-shows and overbooking, the way that orders are assigned to routes in the shipping plan can have important effects on air cargo operations. For example, a shipping plan that does not take into account uncertainty on the quantity of goods could result in an overflow of cargo for some planes and in an under-utilization of the cargo capacity for others. When an overflow occurs, the commodities in excess have to be reassigned to other flights, and this procedure often takes place only a few hours before departure, which usually results in a loss of efficiency.

In order to promote a better utilization of cargo capacity, we study the question of rebalancing a transportation network by reoptimizing the shipping plan under uncertainty about the quantity of cargo. More precisely, given a *flight schedule*, i.e., a list of all the flights available (including for each flight their departure date, origin, destination, volume and weight capacity) and a shipping plan (including the weight and volume of the individual demands), we propose to generate an improved shipping plan by reassigning in advance some of the orders to other routes in a way that minimizes the total expectation of overflow. Besides preventing that overflows arise too often, a better distribution of cargo among the different flights would also allow cargo airlines to accept more orders by creating space on the busiest flights.

## 1.1 Literature survey

We present here an overview of the literature that addresses the problem of reoptimizing a shipping plan. Other related approaches that are orthogonal to ours and could be used to complement it are included below as well. For an extensive literature survey on air cargo operations, the reader is referred to Feng et al. (2015).

### 1.1.1 Reoptimization of the shipping plan

**Deterministic** In Nahum et al. (2019), the authors discuss deterministic optimization models that aim to minimize the total transportation cost while satisfying time and capacity constraints. Moreover, their models suppose that goods can be transported by full cargo aircraft, that transport exclusively cargo, as well as in the baggage hold of passenger aircraft. However, overbooking is not taken into consideration, and capacity limits are treated as hard constraints.

**Stochastic demand** The problem of reoptimizing an existing shipping plan in order to load as much priority cargo as possible is studied in Peng et al. (2019). The authors developed a data-driven air cargo redistribution model based on a combination of dynamic and linear programming that reassigns the commodities that could not be loaded to other flights of the same airline segment. It was found that for a given origin/destination pair, some flights are full, whereas others are not used at all. The model they developed allows for more homogeneous flight loads, but treats airline segments independently. Steadie Seifi (2017) addresses a multimodal transportation problem for perishable products in which empty loading units have to be managed. The goal is to allocate and reposition the empty loading units while finding a trade-off between the operation costs and product freshness. In order to do so, a two-stage scenario-based stochastic model embedded within a rolling horizon framework is developed. Since the model they end up with is very large, they cannot solve them using exact algorithms, but resort instead to an adaptive large neighbourhood search heuristic.

**Stochastic time** Patomtummakan and Nananukul (2017) present a decision support system that suggests routes and that allows the user to accept or reject clients' demands. Moreover, in their system, an optimization problem that produces an improved shipping plan by reassigning orders is solved daily. It takes direct flights into account, but not routes possibly consisting of many flights. Their model assumes deterministic demand, but stochastic arrival times. The problem of optimizing cargo routing has also been investigated from the point of view of a freight forwarder by Azadian et al. (2012). They developed a dynamic programming model in which weight and volume are considered as deterministic quantities, but that takes into account stochastic departure delays (due, e.g., to weather or holidays). Since they treat the problem from the standpoint of forwarders, their goal is not to reduce overflow or to make a better utilization of the flight capacity.

### 1.1.2 Other approaches

A frequent approach to improving air cargo planning and operations consists in tackling the problem from the point of view of revenue management. Since the spot market price is not fixed over the sales period, rejecting orders in anticipation of a future increase in price might lead to higher profits than accepting all of them right away at a lower price and thus using up capacity too early. Many papers are then dedicated to this question (see, e.g., Han et al. (2010), Levina et al. (2011), and Barz and Gartner (2016)). Moreover, since allotments take up a considerable part of cargo capacity, numerous efforts are directed at determining how much space should be allocated to them (see, e.g., Amaruchkul and Lorchirachoonkul (2011), Zhang et al. (2010), and Levin et al. (2012)). In another common approach, attempts are made at determining the best overbooking rate for a given flight, i.e., the space that should be sold in excess of the flight capacity so as to reduce the different costs resulting from an overflow of goods or a waste of capacity (see, e.g., Kasilingam (1997b), Luo et al. (2009), Wang and Kao (2008), and Wannakrairot and Phumchusri (2016)).

### 1.2 Contributions and organization of the paper

In this paper, we propose a new approach to dealing with the uncertainty inherent to the air cargo industry, namely reoptimizing the shipping plan in a way that minimizes the expectation of overflow. The main contributions of this paper are as follows. 1) We model the show-up rate with a mixture distribution that takes no-shows into account. 2) Based on this probability distribution, we derive an analytic expression for the expectation of overflow. 3) We introduce an integer programming model which takes as objective function this analytic expression. This model solves the stochastic problem without resorting to the construction of scenarios. Only an estimation of the no-show rate, the average and the variance is required. Moreover, the whole transportation network is taken into account when reoptimizing the shipping plan and not only direct flights or airline segments, which potentially allows for a greater reduction of overflow. 4) We propose different approximations of the analytic expression (which is nonlinear and may contain a very large number of terms) that transform the model into an integer program solvable by a commercial solver. 5) We show with simulation experiments that the resulting model provides reductions of overflow that are similar to the ones that could be obtained with a standard (scenario-based) stochastic programming approach, but with calculation times much smaller. Since we do not use real demand data, these experiments are not suitable for evaluating actual cost savings that could be achieved by using our models. Our goal is rather to analyze computational time and precision of our different formulations, and thus to show that our new approach is ready to be tested with real data coming from the air cargo industry.

The rest of the paper is organized as follows. In Section 2, we model the weight distribution of demands, and we derive from it an analytic expression for the expectation of overflow. Moreover, different models of overflow minimization are presented there. In Section 3, we discuss some possible approximations of the analytic expression. In Section 4, we present the results of our simulation experiments. Finally, we conclude and summarize future work in Section 5.

## 2 Overflow minimization

In this section, we introduce our model to reoptimize shipping plans in order to minimize the expectation of overflow. First, we present in Section 2.1 a simpler version of the model, in which the weight of the demand is considered as a deterministic quantity. In Section 2.2, we introduce a probability distribution to model the show-up rate. The deterministic model is then modified to take into account stochasticity of the demand weight. The resulting stochastic model is tackled using a standard scenario-based approach in Section 2.3. Finally, we derive in Section 2.4 an analytic expression for the expectation of overflow, which is then used as the objective function of the stochastic model developed in Section 2.2.

### 2.1 Deterministic model

Let $D$ be the set of all cargo demands and $\Omega_d$ be the set of all the routes that can satisfy demand $d \in D$. Define

$$x_{d\omega} = \begin{cases} 1 & \text{if route } \omega \in \Omega_d \text{ is chosen to satisfy demand } d \in D, \\ 0 & \text{otherwise.} \end{cases}$$

Let $L$ be the set of all legs. Define

$$a_{\omega l} = \begin{cases} 1 & \text{if leg } l \in L \text{ is in route } \omega \in \Omega_d, \\ 0 & \text{otherwise.} \end{cases}$$

For each $l \in L$, let $u_l$ be the capacity of leg $l$ (in terms of weight). Let $m_d$ be the weight of the commodities associated with demand $d$. Define the weight overflow on the leg $l \in L$ by

$$O_l(x, m) := \max\left(0, \sum_{d \in D} \sum_{\omega \in \Omega_d} m_d a_{\omega l} x_{d\omega} - u_l\right), \qquad l \in L,$$

where $x := (x_{d\omega})_{d\in D,\omega\in\Omega}$ and $m := (m_d)_{d\in D}$. The total overflow for the entire transportation network is then given by

$$O(x,m) := \sum_{l\in L} O_l(x,m). \tag{1}$$

We are interested in minimizing the total overflow of the transportation network. In other words, we want to solve the following nonlinear program:

$$\min_x O(x,m) \tag{2a}$$

subject to

$$\sum_{\omega\in\Omega_d} x_{d\omega} = 1 \text{ for all } d\in D, \tag{2b}$$

$$x_{d\omega} \in \{0,1\} \text{ for all } d\in D \text{ and for all } \omega\in\Omega. \tag{2c}$$

The objective function (2a) is not linear, but can easily be linearized by introducing one slack variable $Z_l$ for each leg $l \in L$ and the corresponding appropriate constraints.

For logistical and administrative reasons, it may be advisable to limit the number of demands that can be reassigned to a new route during the optimization. Indeed, an important adjustment of the workforce at the airport might be required if the shipping plan has changed substantially. Moreover, since clients must be informed whenever their order is reassigned, unnecessary reassignments might lead to additional administrative costs. To limit the number of demand reassignments, the following constraint can be added to the model:

$$\sum_{d\in D} \sum_{\substack{\omega\in\Omega_d,\\ \omega\neq\omega_d}} x_{d\omega} \leq n_r, \tag{3}$$

where $\omega_d$ is the route to which the demand $d$ was assigned initially (e.g., during the booking) and $n_r$ is the maximal number of demand reassignments.

## 2.2 Stochastic model

Since the quantity of commodities arriving at the airport on the day of departure can differ greatly from the one booked, the weight of a commodity should be considered as a random variable. Popescu et al. (2006) argue that the show-up rate does not follow any of the usual parametric distributions (e.g., exponential, beta, normal, etc.), but rather a nonparametric discrete distribution, the density of which can be estimated using a histogram estimator. This can be explained, at least to some extent, by the fact that there is an important probability of no-show, which cannot be taken directly into account by the usual parametric distributions. In order to capture this phenomenon, we make the assumption that for a given demand, there can be a no-show with probability $0 \leq 1 - p \leq 1$.

Moreover, we will suppose that when the client shows up at the airport (with probability $p$), the demand weight is not exactly the one booked, but follows a normal distribution. This assumption is made in view of the central limit theorem, which implies that when the number of demands on a flight is large, a normal law should be a good approximation for the total weight, regardless of the weight distribution of the individual demands. Moreover, in the case of an order placed by a freight forwarder, since the commodities correspond to an aggregation of several demands, its weight distribution is already expected to resemble a normal distribution. For these reasons, modeling the weight of the demands that showed up at the airport with the help of a normal law is not expected to introduce large errors on the total freight distribution on each aircraft. This error will be discussed when presenting the experimental results in Section 4.

Given the aforementioned assumptions, we will model the weight of a commodity as a random variable following a probability distribution that is a mixture of a normal distribution and a constant random variable with value 0. Its probability density function is then given by

$$f_{m_d} = (1-p)f_{\mathbf{0}} + pf_{X_d}, \tag{4}$$

where $f_{\mathbf{0}}$ and $f_{X_d}$ correspond to the probability density functions of $\mathbf{0}$, i.e., the constant random variable taking the value 0, and of $X_d \sim \mathcal{N}(\mu_d, \sigma_d)$ respectively. Note that when one takes the *normalized weight* $\frac{m_d}{\mu_d}$ instead of $m_d$ (so that $X_d \sim \mathcal{N}(1, \sigma_d/\mu_d)$), then (4) corresponds to the density function of the *distribution of the show-up rate.*

Instead of minimizing the overflow, we minimize the expectation of overflow:

$$V(x) := \mathbb{E}_m \left( \sum_{l \in L} \max\left(0, g_l(x,m) - u_l\right) \right),$$

where $m := (m_d)_{d \in D}$, and $g_l(x,m) := \sum_{d \in D} \sum_{\omega \in \Omega_d} m_d a_{\omega l} x_{d\omega}$ is the weight of the commodity assigned to the leg $l \in L$. A stochastic version of the overflow minimization model (2a)–(2c) is obtained by replacing the objective function with

$$\min_x V(x) \tag{5}$$

## 2.3 Sample average approximation (SAA)

The expectation can be approximated using the Monte Carlo method as follows:

$$\mathbb{E}_m \left( \sum_{l \in L} \max\left(0, g_l(x,m) - u_l\right) \right) \approx \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{l \in L} \max\left(0, g_l(x,m^i) - u_l\right) \right),$$

where $(m^i)_{i=1}^n$ is a sequence of equal-probability realizations of the (multivariate) random variable $m$. In the rest of the paper, those realizations will be called *scenarios.*

Since the function $f$ contains a max function, the resulting model is not linear. The max function can be linearized as follows:

$$\min \frac{1}{n} \sum_{i=1}^{n} \sum_{l \in L} Z_l^i \tag{6a}$$

subject to

$$Z_l^i \geq 0 \text{ for all } l \in L, \text{for all } i = 1, \ldots, n, \tag{6b}$$

$$Z_l^i \geq \sum_{d \in D} \sum_{\omega \in \Omega_d} m_d^i a_{\omega l} x_{d\omega} - u_l \text{ for all } l \in L, \text{for all } i = 1, \ldots, n, \tag{6c}$$

$$\sum_{\omega \in \Omega_d} x_{d\omega} = 1 \text{ for all } d \in D, \tag{6d}$$

$$x_{d\omega} \in \{0,1\} \text{ for all } d \in D, \text{ for all } \omega \in \Omega. \tag{6e}$$

This stochastic integer programming model will be referred to as the *SAA model* in the remainder of the paper.

## 2.4 An analytic expression for the expectation of overflow

This subsection is dedicated to deriving an analytic expression for $V(x)$.

**Theorem 1** *Let $D_l \subset D$ be the set of all the demands that could potentially travel by the leg $l \in L$, and let $n_l := |D_l|$, i.e., the cardinality of $D_l$. Let $\Sigma_1, \cdots, \Sigma_{2^{n_l}}$ be an enumeration of all the $2^{n_l}$ subsets of $D_l$. Then, the expectation of overflow is given by*

$$V(x) = \sum_{l \in L} \sum_{i=1}^{2^{n_l}} \frac{P_{\Sigma_i}}{2} \left[ (\mu_i(x) - u_l)\left(1 - \operatorname{erf}\left(z_i(x, u_l)\right)\right) + \sigma_i(x) e^{-z_i^2(x,u_l)} \right] \tag{7}$$

*where*

$$P_{\Sigma_i} := p^{|\Sigma_i|}(1-p)^{n_l - |\Sigma_i|}, \quad \operatorname{erf}(x) := \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2}\, dt, \quad z_i(x, \xi) := \frac{\xi - \mu_i(x)}{\sqrt{2}\sigma_i(x)}$$
$$\mu_i(x) := \sum_{d \in \Sigma_i} c_{dl}(x) \mu_d, \quad \sigma_i^2(x) := \sum_{d \in \Sigma_i} c_{dl}^2(x) \sigma_d^2, \quad \textit{and} \quad c_{dl}(x) := \sum_{\omega \in \Omega_d} a_{\omega l} x_{d\omega}.$$

**Proof.** Since

$$\mathbb{E}_m \left( \sum_{l \in L} \max\left(0, g_l(x,m) - u_l\right) \right) = \sum_{l \in L} \mathbb{E}_m \left( \max\left(0, g_l(x,m) - u_l\right) \right), \tag{8}$$

we can consider independently the expectation of overflow on each flight leg, i.e.,

$$V_l(x) := \mathbb{E}_m \left( \max\left(0, g_l(x,m) - u_l\right) \right).$$

By definition of the expectation,

$$V_l(x) = \int_{g_l(x,m) > u_l} \left( g_l(x,m) - u_l \right) dP(m) = \int_{u_l}^{\infty} \left( \xi - u_l \right) f_{g_l}(\xi) d\xi,$$

where $f_{g_l}$ is the probability density of the random variable $g_l(x,m)$.

In order to obtain an analytic expression for the expectation of overflow, the density $f_{g_l}$ has first to be determined. Note that given a vector $x$, the random variable $g_l(x,m)$ corresponds to a linear combination of the random variables $m_d$. Indeed, $g_l(x,m) = \sum_{d \in D} c_{dl}(x) m_d$.

Let $Y_l$ be the random variable defined by $Y_l = i$ if $D_l \setminus \Sigma_i$ are all no-shows. Since the events $Y_l = i$ correspond to the success of $|\Sigma_i|$ independent Bernoulli trials of parameter $p$, it follows that

$$P(Y_l = i) = p^{|\Sigma_i|}(1-p)^{n_l - |\Sigma_i|}. \tag{9}$$

By the Law of Total Expectation,

$$\begin{aligned} V_l(x) &= \mathbb{E}_{Y_l} \left[ \mathbb{E}_m \left( \max\left(0, g_l(x,m) - u_l\right) | Y_l \right) \right] \\ &= \sum_{i=1}^{2^{n_l}} V_l^i(x) P(Y_l = i) = \sum_{i=1}^{2^{n_l}} p^{|\Sigma_i|}(1-p)^{n_l - |\Sigma_i|} V_l^i(x), \end{aligned} \tag{10}$$

where $V_l^i(x) := \mathbb{E}_m \left( \max\left(0, g_l(x,m) - u_l\right) | Y_l = i \right)$. Moreover,

$$\begin{aligned} V_l^i(x) &= \int_{g_l(x,m) > u_l} \left( g_l(x,m) - u_l \right) dP(m | Y_l = i) \\ &= \int_{u_l}^{\infty} \left( \xi - u_l \right) f_{g_l | Y_l = i}(\xi) d\xi, \end{aligned} \tag{11}$$

where $f_{g_l | Y_l = i}$ is the conditional density function of $g_l(x,m)$ given $Y_l = i$. When $Y_l = i$, the function $g_l(x,m)$ corresponds to a linear combination of $X_d$, i.e., of normally distributed random variables. Therefore, $g_l(x,m)$ is also a normally distributed random variable, and its probability density is

$$f_{g_l | Y_l = i}(\xi) = \frac{1}{\sqrt{2\pi}\sigma_i(x)} \exp\left\{ -z_x^2(\xi) \right\}. \tag{12}$$

If we replace this expression of $f_{g_l|Y_l=i}(\xi)$ in (11) and then integrate by parts, we obtain

$$\begin{aligned} V_l^i(x) &= \frac{1}{\sqrt{2\pi}\sigma_i(x)} \int_{u_l}^{\infty} (\xi - u_l) \exp\left\{-z_x^2(\xi)\right\} d\xi \\ &= \frac{1}{\sqrt{2\pi}\sigma_i(x)} \left[ \sqrt{\frac{\pi}{2}}\sigma_i(x)(\mu_i(x) - u_l)\mathrm{erf}\left(z_i(x,\xi)\right) - \sigma_i^2(x) \exp\left\{-z_i^2(x,\xi)\right\} \right]_{\xi=u_l}^{\infty} \\ &= \frac{1}{2} \left[ (\mu_i(x) - u_l)\left(1 - \mathrm{erf}\left(z_i(x,u_l)\right)\right) + \sigma_i(x) \exp\left\{-z_i^2(x,u_l)\right\} \right]. \end{aligned} \tag{13}$$

By combining (8), (10), and (13), we obtain (7), which concludes the proof of the theorem. □

**Remark 1** *For the sake of simplicity, we assumed in the proof of Theorem 1 that all the demands have the same no-show rate $1-p$. However, this assumption could be lifted by considering for each demand $d \in D$ a different no-show rate $1-p_d$. Equation (9) should then be replaced by*

$$P(Y_l = i) = \left( \prod_{d \in \Sigma_i} p_d \right) \left( \prod_{d \in D_l \setminus \Sigma_i} (1-p_d)^{n_l - |\Sigma_i|} \right),$$

*and the rest of the proof would remain the same.*

From a practical point of view, it might be useful to express the expectation of overflow in terms of the (weight) load factor rather than in terms of absolute weight. This is the result of the following corollary of Theorem 1.

**Corollary 1** *Let $\tilde{\mu}_i(x) := \frac{\mu_i(x)}{u_l}$ (i.e., the average of the weight load factor), and $\tilde{\sigma}_i(x) := \frac{\sigma_i(x)}{u_l}$ (i.e., the standard deviation of the weight load factor). Then*

$$V(x) = \sum_{l \in L} u_l \sum_{i=1}^{2^{n_l}} p^{|\Sigma_i|}(1-p)^{n_l - |\Sigma_i|} E(\tilde{\mu}_i(x), \tilde{\sigma}_i^2(x)), \tag{14}$$

*where*

$$E(u,v) := \frac{1}{2}\left[ (u-1)\left(1 - \mathrm{erf}\left(-\frac{u-1}{\sqrt{2v}}\right)\right) + \sqrt{v}\exp\left\{-\frac{(u-1)^2}{2v}\right\} \right]. \tag{15}$$

**Proof.** By using the definitions of $\tilde{\mu}_i(x)$ and $\tilde{\sigma}_i(x)$ in (13), we obtain

$$V_l^i(x) = \frac{u_l}{2}\left[ (\tilde{\mu}_i(x) - 1)\left(1 - \mathrm{erf}\left(-\tilde{z}_i(x,\xi)\right)\right) + \tilde{\sigma}_i(x)\exp\left\{-\tilde{z}_i^2(x,\xi)\right\} \right],$$

where $\tilde{z}_i(x,\xi) := \left(\frac{\tilde{\mu}_i(x)-1}{\sqrt{2}\tilde{\sigma}_i(x)}\right)$. Hence, $V_l^i(x) = u_l E(\tilde{\mu}_i(x), \tilde{\sigma}_i^2(x))$, and the statement of the corollary follows. □

If we use the analytic expression for $V(x)$ provided by Corollary 1 as the objective function (5) of the stochastic model described in Section 2.2, we obtain the following (nonlinear) integer programming model, that will be referred to as the *analytic model*:

$$\min_x \sum_{l \in L} u_l \sum_{i=1}^{2^{n_l}} p^{|\Sigma_i|}(1-p)^{n_l - |\Sigma_i|} E(\tilde{\mu}_i(x), \tilde{\sigma}_i^2(x)) \tag{16a}$$

subject to

$$\sum_{\omega \in \Omega_d} x_{d\omega} = 1 \text{ for all } d \in D, \tag{16b}$$

$$x_{d\omega} \in \{0,1\} \text{ for all } d \in D \text{ and for all } \omega \in \Omega. \tag{16c}$$

This nonlinear integer program can be solved if its objective function is convex. This is a consequence of the following proposition.

**Proposition 1** *The function $x \mapsto E(\tilde{\mu}_i(x), \tilde{\sigma}_i^2(x))$ is convex.*

**Proof.** Define $G(x,m) := \max(0, g_l(x,m) - u_l)$. Since $g_l$ is a linear function of $x$, the function $x \mapsto G(x,m)$ is convex. By definition of the expectation,

$$\begin{aligned} E(\tilde{\mu}_i(x), \tilde{\sigma}_i^2(x)) &= \frac{1}{u_l}\mathbb{E}_m\left(\max(0, g_l(x,m) - u_l)\,|Y_l = i\right) \\ &= \int_{\mathbb{R}^{|D|}} \frac{G(x,m)}{u_l}\, dP(m|Y_l = i). \end{aligned}$$

The function $x \mapsto E(\tilde{\mu}_i(x), \tilde{\sigma}_i^2(x))$ then corresponds to the integral of a convex function, which itself is convex. □

# 3 Approximation of the analytic expression for the expectation of overflow

The analytic expression proven in Theorem 1 has two main drawbacks. First, since it is neither linear nor quadratic, standard commercial solvers would not be able to solve the analytic model (16a)–(16c) directly. This issue is addressed in Section 3.1. Second, since the number of terms in the inner summation of (7) increases exponentially with the number of demands that can potentially travel on a leg, the total number of terms in the objective function of the resulting analytic model will most likely be too large to be tractable, when reoptimizing shipping plans of a real transportation network.

## 3.1 Approximation of the function $E$

The integer model (16a)–(16c) could be solved directly by a commercial solver, if we first approximate the function $E$ (see Corollary 1) by piecewise linear or quadratic functions. Some ways of approximating it are described below.

Before proceeding with specific approximation methods, note that even though $\sigma_i^2(x)$ is a quadratic function in the variable $x$, it can be considered as a linear function of $x$ since $x_{d\omega} \in \{0,1\}$, which implies that $x_{d\omega}^2 = x_{d\omega}$. Similarly, $a_{\omega l} \in \{0,1\}$ implies $a_{\omega l}^2 = a_{\omega l}$, and then $c_{dl}^2(x) = c_{dl}(x)$. Therefore,

$$\sigma_i^2(x) = \sum_{d \in \Sigma_i} c_{dl}^2(x)\sigma_d^2 = \sum_{d \in \Sigma_i} c_{dl}(x)\sigma_d^2, \tag{17}$$

which is a linear function in the variable $x$. A piecewise linear (resp. quadratic) approximation of $E$ will then lead to a piecewise linear (resp. quadratic) objective function. Moreover, keeping in mind that $\sigma_i^2(x)$ will be used in the linear relaxation of the analytic model, replacing $x_{d\omega}^2$ with $x_{d\omega}$ gives a better lower bound for the branch and bound algorithm. Indeed, since $x_{d\omega} > x_{d\omega}^2$ for $0 < x_{d\omega} < 1$, this replacement provides a convex upper bound on the original objective function, which corresponds to a better lower bound.

### 3.1.1 Surface approximation

By letting $u$ and $v$ varying over the possible values for $\tilde{\mu}_i(x)$ and $\tilde{\sigma}_i(x)$ respectively, the values of $E(u,v)$ form a convex surface. More precisely, the set

$$\mathcal{S} := \left\{(u,v,w) \in \mathbb{R}^3 : 0 \le u < \infty, 0 < v < \infty, w = E(u,v)\right\}$$

corresponds to a smooth convex surface in $\mathbb{R}^3$. This surface is represented in Figure 1.

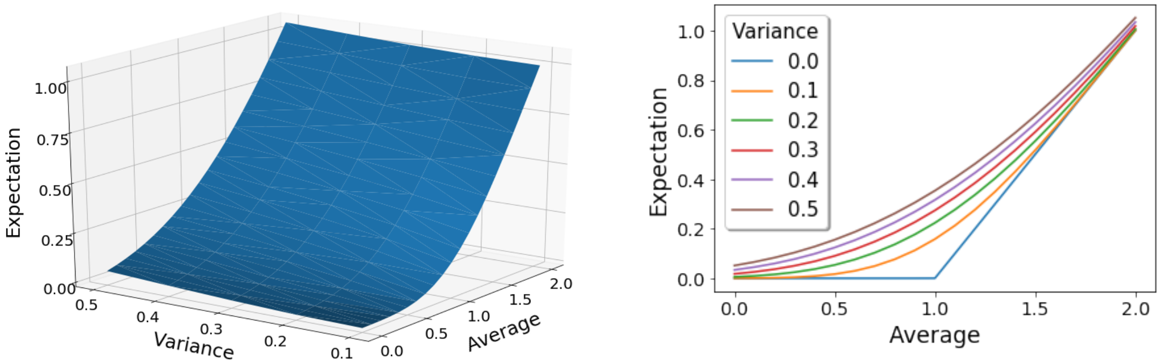

**Figure 1: On the left, the surface $\mathcal{S}$, and on the right, curves described by the function $w : u \mapsto E(u,v)$ for different fixed values of $v$. Here, $u$ corresponds to the "average", $v$ to the "variance", and $w$ to the "expectation".**

**Linear approximation** The surface $\mathcal{S}$ can be approximated using standard triangulation algorithms. Each triangle lives in a plane that can then be added as a constraint to the model by using a standard slack-variable method. Note that even though $\mathcal{S}$ is convex, the resulting triangulation is not necessarily convex. Therefore, in order to make sure that an optimal solution is found, additional precautions have to be taken when constructing the triangulation. Another solution would be to use a different convex linear approximation. For example, the surface could be approximated by a certain number of tangent planes, which then provides a convex lower bound on the objective function.

An important drawback of this approach is that a fair number of triangles should be used in order to obtain a good approximation of the surface. This results in a large number of new constraints and variables to be added to the model.

**Quadratic approximation** The surface $\mathcal{S}$ can also be approximated by a two-variable quadratic function. By (17), if $E(u,v)$ is approximated by a quadratic function, say $\tilde{E}(u,v)$, then $\tilde{E}(\tilde{\mu}_i(x), \tilde{\sigma}_i^2(x))$ is a quadratic function of $x$. It follows from Corollary 1 that the objective function is also quadratic. Since finding a single quadratic function that approximates well enough $E$ is not an easy task, quadratic approximation is expected, in general, to lead to poor results.

### 3.1.2 Curve approximation

If we were able to express $\tilde{\sigma}_i(x)$ as a function of $\tilde{\mu}_i(x)$, we could replace the surface $\mathcal{S}$ depending on the two parameters $\tilde{\mu}_i(x)$ and $\tilde{\sigma}_i(x)$ by a curve depending on the single parameter $\tilde{\mu}_i(x)$. Approximating the function $E$ with better precision would then be possible with a smaller model.

Since the standard deviation of the load factor should generally increase as the average of the load factor increases, finding a precise relation between $\tilde{\sigma}_i(x)$ and $\tilde{\mu}_i(x)$ does not appear too far-fetched at first glance. However, given that the specific values of the elements of the vector $x$ have an impact on the variance, it is not possible, in theory, to express $\tilde{\sigma}_i(x)$ directly as a function of $\tilde{\mu}_i(x)$.

On the other hand, from a more practical point of view, a rough estimate of $\tilde{\sigma}_i(x)$ expressed as a function of $\tilde{\mu}_i(x)$ might represent a good approximation of the actual value of $\tilde{\sigma}_i(x)$. The loss of precision incurred by this approximation is expected to be compensated by the better approximation of the function $E$.

A simple way to express $\tilde{\sigma}_i(x)$ as a function of $\tilde{\mu}_i(x)$ is to interpolate $\tilde{\sigma}_i(x)$ from its maximal value on a given flight, i.e.,

$$\tilde{\sigma}_i^{\text{app}}(\tilde{\mu}_i(x)) := \frac{\tilde{\mu}_i(x)}{\tilde{\mu}_i^{\max}} \tilde{\sigma}_i^{\max}, \tag{18}$$

where $\tilde{\sigma}_i^{\max} := \max_x \tilde{\sigma}_i(x)$ and $\tilde{\mu}_i^{\max} := \max_x \tilde{\mu}_i(x)$. The surface $\mathcal{S}$ can then be replaced by a set of curves, one for each leg $l$ and each of the $2^{n_l}$ configurations $\Sigma_1, \ldots, \Sigma_{2^{n_l}}$:

$$\mathcal{C}_{l,i} := \left\{(s,w) \in \mathbb{R}^2 : 0 \leq s < \infty, w = E(s, \tilde{\sigma}_i^{\text{app}}(s))\right\}, \qquad l \in L,\ i = 1, \ldots, 2^{n_l}$$

An example of such curves is shown in Figure 2. The functions defining these curves are nonlinear, but can be interpolated by linear or quadratic functions. Since curves are one dimensional, they are easier to approximate than surfaces, and a smaller number of pieces are generally required to reach the same precision.

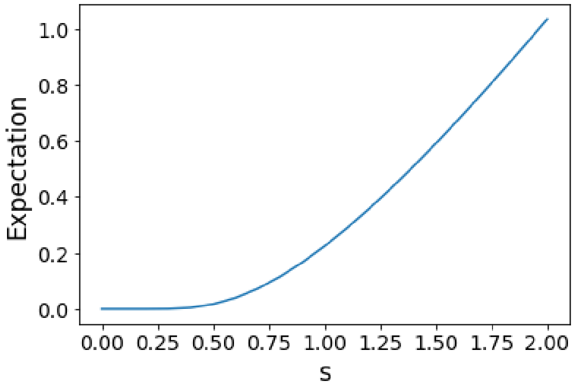

**Figure 2: Curve described by the function $s \mapsto E(s, \tilde{\sigma}_i^{\text{app}}(s))$ for $\tilde{\mu}_i^{\max} = 2$.**

For the computational experiments presented in Section 4, we use a linear interpolation of this curve approximation. More precisely, for each flight $l \in L$ and each subset $\Sigma_i^*$ of $D_l^*$, we first determine the value that $\tilde{\mu}_i(x)$ would take if all the goods were assigned to the flight $l$, i.e., $\tilde{\mu}_i^{\max}$. Using at most $n_{\text{pces}}$ pieces, we then interpolate linearly the function $s \mapsto E(s, \tilde{\sigma}_i^{\text{app}}(s))$. In order to have a finite number of linear pieces of same length, the interpolation is done on the finite interval $0 \leq s \leq s_{\max}$, for some $s_{\max} \in \mathbb{R}$ fixed beforehand. Given $n_{\text{pces}}$, the set of interpolation points $\mathcal{P}$ is chosen as follows:

$$\mathcal{P}(\tilde{\mu}_i^{\max}) := \begin{cases} \left\{0, \frac{1}{n_{\text{pces}}}, \frac{2}{n_{\text{pces}}}, \ldots, n^*\right\} & \text{if } 0 \leq \tilde{\mu}_i^{\max} < 1 \\ \left\{0, \frac{1}{n_{\text{pces}}}\tilde{\mu}_i^{\max}, \frac{2}{n_{\text{pces}}}\tilde{\mu}_i^{\max}, \ldots, \tilde{\mu}_i^{\max}\right\} & \text{if } 1 \leq \tilde{\mu}_i^{\max} < s_{\max} \\ \left\{0, \frac{1}{n_{\text{pces}}}s_{\max}, \frac{2}{n_{\text{pces}}}s_{\max}, \ldots, s_{\max}\right\} & \text{if } \tilde{\mu}_i^{\max} \geq s_{\max} \end{cases}$$

where $n^*$ is the smallest integer such that $\frac{n^*}{n_{\text{pces}}} > \tilde{\mu}_i^{\max}$. Note that with this choice, $|\mathcal{P}(\tilde{\mu}_i^{\max})| = n_{\text{pces}}$ for $\tilde{\mu}_i^{\max} \geq 1$ and $|\mathcal{P}(\tilde{\mu}_i^{\max})| \leq n_{\text{pces}}$ for $\tilde{\mu}_i^{\max} < 1$. Considering that a good approximation is only required when the average of the load factor is close to 100%, or in other words, when $s \approx 1$, the main idea behind the definition of $\mathcal{P}(\tilde{\mu}_i^{\max})$ is to restrict the number of interpolation points when $\tilde{\mu}_i^{\max}$ is small. For the experiments described in Section 4, we fixed $s_{\max} := 4$.

## 3.2 Aggregating the demands

Since the number of subsets of $D_l$ increases exponentially with $n_l$, the analytic expression for the expectation of overflow is inapplicable when the number of demands that can potentially travel by a leg is not relatively small. Indeed, the number of terms in the sum in the expression of the objective function (10) becomes quickly prohibitive when $|D_l|$ increases.

A solution to this problem would consist in conditioning on a fixed number $n_{\max}$ of events $Y_l = i$ in the equations that lead to (10). More precisely, if $|D_l| > n_{\max}$, let $D_l^*$ be a subset of $D_l$ that contains $n_{\max}$ demands. Considering that no-shows for orders of larger sizes have a greater impact on planning than no-shows for orders of smaller sizes, a simple and efficient way of selecting the demands that will be part of $D_l^*$ is to keep the $n_{\max}$ ones with the largest weights. Let $\Sigma_1^*, \cdots, \Sigma_{2^{n_{\max}}}^*$ be an enumeration of the $2^{n_{\max}}$ subsets of $D_l^*$. Let $Y_l^*$ be the random variable defined by $Y_l^* = i$ if $D_l^* \setminus \Sigma_i^*$ are all no-shows. By repeating the argument that led to (10), we obtain

$$V_l(x) = \sum_{i=1}^{2^{n_{\max}}} p^{|\Sigma_i^*|}(1-p)^{n_{\max}-|\Sigma_i^*|}\mathbb{E}_m\left(\max\left(0, g_l(x,m) - u_l\right)|Y_l^* = i\right).$$

However, when $Y_l^* = i$, the function $g_l(x,m)$ does not correspond to a linear combination of normally distributed random variables. Indeed, only the $n_{\max}$ terms associated with the demands in $D_l^*$ form a linear combination of normally distributed random variables. The other terms, associated with the demands in $D_l \setminus D_l^*$, correspond to random variables following the mixture distribution whose density is given by (4). The function $g_l(x,m)$ can then be decomposed as two sums, one corresponding to the "main" $n_{\max}$ demands, which are larger, and thus no-shows for them would have a greater impact on the total weight of the demands travelling by the leg $l$, and another corresponding to the demands exceeding $n_{\max}$:

$$g_l(x,m) = \underbrace{\sum_{d\in D_l^*} c_{dl}(x)m_d}_{S_{\text{main}}:=} + \underbrace{\sum_{d\in D_l\setminus D_l^*} c_{dl}(x)m_d}_{S_{\text{excess}}:=}.$$

Since the first sum is a linear combination of normally distributed random variables, the probability density of $S_{\text{main}}$ is given by (12), i.e.,

$$f_{S_{\text{main}}|Y_l=i}(\xi) = \frac{1}{\sqrt{2\pi}\bar{\sigma}_i(x)}\exp\left\{-\bar{z}_i(x,\xi)\right\},$$

where $\bar{z}_i(x,\xi) := \frac{\xi-\bar{\mu}_i(x)}{\sqrt{2}\bar{\sigma}_i(x)}$,

$$\bar{\mu}_i(x) := \sum_{d\in\Sigma_i^*} c_{dl}(x)\mu_d \quad \text{and} \quad \bar{\sigma}_i^2(x) := \sum_{d\in\Sigma_i^*} c_{dl}^2(x)\sigma_d^2.$$

On the other hand, when $|D_l \setminus D_l^*|$ is large, it seems reasonable in view of the central limit theorem to approximate the probability density of $S_{\text{excess}}$ using a normal approximation, i.e.,

$$f_{S_{\text{excess}}}(\xi) \approx \frac{1}{\sqrt{2\pi}\sigma_{\text{excess}}(x)}\exp\left\{-z_{\text{excess}}^2(x,\xi)\right\},$$

where $z_{\text{excess}}(x,\xi) := \frac{\xi-\mu_{\text{excess}}(x)}{\sqrt{2}\sigma_{\text{excess}}(x)}$,

$$\begin{aligned}\mu_{\text{excess}}(x) &:= \sum_{d\in D_l\setminus D_l^*} c_{dl}(x)p\mu_d,\\ \sigma_{\text{excess}}^2(x) &:= \sum_{d\in D_l\setminus D_l^*} c_{dl}^2(x)(p\sigma_d^2 + p(1-p)\mu_d).\end{aligned}$$

Moreover, given that most of the demands contained in $D_l \setminus D_l^*$ will generally be much smaller than the demands in $D_l^*$, the value of the errors induced by this approximation should be relatively small.

Since $S_{\text{main}}$ is normally distributed and $S_{\text{excess}}$ can be approximated by a normally distributed random variable, their sum, i.e. $g_l$, can also be approximated by a normally distributed random variable as follows:

$$f_{g_l|Y_l=i}(\xi) \approx \frac{1}{\sqrt{2\pi}\sigma_i^*(x)}\exp\left\{-\frac{(\xi-\mu_i^*(x))^2}{2{\sigma_i^*}^2(x)}\right\}, \tag{19}$$

where $\mu_i^*(x) := \bar{\mu}_i(x) + \mu_{\text{excess}}(x)$ and ${\sigma_i^*}^2(x) := \bar{\sigma}_i^2(x) + \sigma_{\text{excess}}^2(x)$. Since the right-hand side of (19) has the same form as (12), the same argument that led to (14) can be applied to (19). Hence

$$V(x) \approx \sum_{l \in L} u_l \sum_{i=1}^{2^{n_{\max}}} p^{|\Sigma_i^*|}(1-p)^{n_{\max}-|\Sigma_i^*|} E(\tilde{\mu}_i^*, (\tilde{\sigma}_i^*)^2), \tag{20}$$

where $E$ is given by (15), $\tilde{\mu}_i^*(x) := \frac{\mu_i^*(x)}{u_l}$ and $\tilde{\sigma}_i^*(x) := \frac{\sigma_i^*(x)}{u_l}$.

By applying some of the approximations of $E$ presented in Section 3.1, the analytic model (16a)–(16c) can then be solved using a commercial solver.

# 4 Computational experiments

In this section, we present the results of computational experiments carried out with the three models developed in the previous sections: the deterministic model (see Section 2.1), the SAA model (see Section 2.3), and the analytic model (see Section 2.4) with the linear interpolation of the curve approximation described in Section 3.1.2. Our goal is twofold: first, to show that it is worthwhile using a stochastic approach to reduce overflow, and second, to evaluate the gain in computational time that it is possible to obtain with the analytic model as compared to the SAA model.

All experiments were run using a personal computer equipped with an Intel Core i7 processor (i7-6700 CPU @ 3.40GHz) and 32 GB of RAM. The models were all implemented with CPLEX 12.8.0.0 (and an optimality gap of 2 %) in C++11 and compiled with GCC 8.1.0 on Linux.

## 4.1 Generation of benchmark instances

In the experiments, we use the actual flight schedule (including flight capacity) of a passenger airline for the month of May 2017. We will reoptimize randomly generated shipping plans for the week from May 5th to May 11th. In order to construct those shipping plans, three instances A, B and C of 60 000 demands each were generated randomly so that the resulting instances have different maximum load factors of 1.25, 1.10 and 1.05 respectively. Each demand is randomly associated with a day between May 3rd and May 19th. It corresponds to the day at which the demand is expected to be brought to the airport by the client. Moreover, each demand has a 50 % chance of being a priority shipment, in which case the corresponding commodities have to leave the airport the same day. If not, then they are allowed to stay at the airport an extra day. Note that the period during which demands are created (i.e., May 3rd to May 19th) extends outside of the optimization week (i.e., May 5th to May 11th) in order to provide more realistic flight loads. Indeed, demands associated with the latter days of the week could be reassigned to routes that include flights on days after the optimization week.

Care was taken when fixing demand weights so that their distribution reflects what one can expect in practice, namely that relatively few demands have a very large weight. The graph on the left of Figure 3 provides a graphical representation of the general distribution of demand weights for each instance A, B, and C. The vertical axis corresponds to the actual weight divided by the mean weight of the instance (i.e., the sum of all the demand weights divided by the number of demands). The graph indicates that for a given point $(x, y)$ on the curve, $x$% of all the demands have a normalized weight less than $y$. Note that the curves for B and C overlap.

With each instance is then associated an initial shipping plan that was constructed by assigning randomly each demand to a series of flights taken from the set of all available routes. These available routes correspond to routes that the airline actually used to transport cargo in May 2017. The graph on the right of Figure 3 gives for each instance A, B, and C information about the distribution of demand weight among the flights. The vertical axis corresponds to the actual flight load factor (in terms of weight), i.e. for each flight the total weight of the commodities assigned to it divided by the capacity of the plane. The graph indicates that for a given point $(x, y)$ on the curve, $x$% of all the flights have a load factor less than $y$.

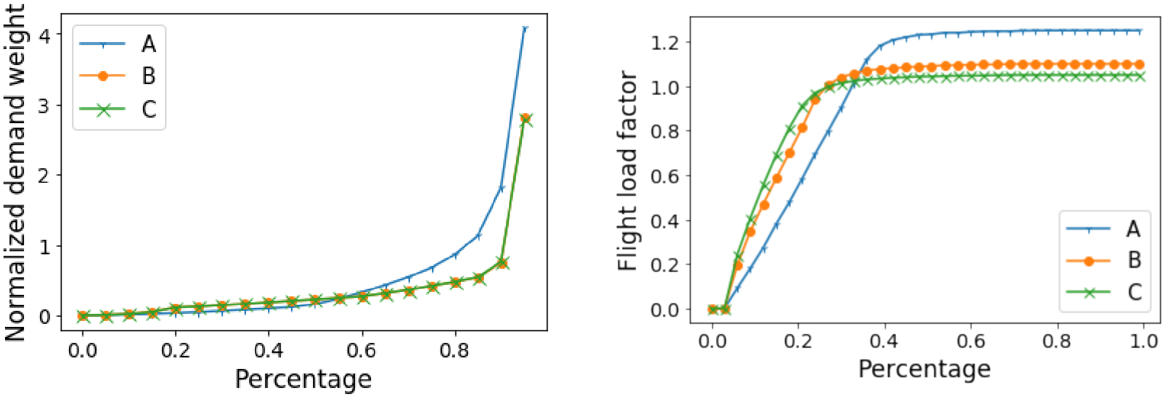

**Figure 3: On the left, the distribution of the demand weights normalized by their mean weight. On the right, the distribution of the flight load factor.**

It is worthwhile to mention that since the demands are generated randomly, the results we obtain should not be used to estimate potential economic gains from using our approach. On the other hand, we use a real flight schedule with actual routes, as well as a realistic number of demands. As a result, the solution times we obtain should be representative of what to expect with a real industrial dataset. Our partially simulated benchmark instances are then sufficient in order to assess the value of the approximations used in our approach.

## 4.2 Overflow reduction

With a shipping plan is associated an order-route assignment vector $x = (x_{d\omega})_{d\in D,\omega\in\Omega}$ (see Section 2.1). Whenever a shipping plan is optimized, a new order-route assignment vector is produced from the initial one. In order to determine the quality of the optimized shipping plan, we would like to evaluate how much cargo overflow can be reduced with the optimized shipping plan, as compared to the initial shipping plan. This would be possible only if we knew the actual weights of the demands that were brought to the airport on the day of departure. To this end, we simulate a weight vector sample, $\tilde{m} = (\tilde{m}_d)_{d\in D}$, corresponding to the actual weight of the demands. More precisely, for each demand, we pick up a show-up rate $\alpha_d$ randomly from the mixture distribution given by (4) (with respect to the normalized weight $\frac{m_d}{\mu_d}$), and we multiply it by the corresponding average weight, i.e., $\tilde{m}_d := \alpha_d\mu_d$.

Since our goal is not to assess the potential economic benefits of the reoptimization, the precise values of the parameters of the mixture distribution do not play an essential role in this research. We fixed them arbitrarily as follows. The no-show rate is chosen to be $1 - p = 0.3$ for all demands. The standard deviation of the weight of a demand that shows up at the airport is assumed to be proportional to its mean, namely, $\sigma_d = 0.3\mu_d$, where $\mu_d$ corresponds to the weight of the demand.

Let $x_{\text{initial}}$ and $x_{\text{optimized}}$ be the order-route assignment vectors of the initial and optimized shipping plans respectively. Given a set of $n_{\text{sample}}$ sample weight vectors $\tilde{m}_1, \tilde{m}_2, \ldots, \tilde{m}_{n_{\text{sample}}}$, we define the *(relative) (average) overflow reduction (OR)* as follows:

$$\text{OR} := \frac{1}{n_{\text{sample}}} \sum_{i=1}^{n_{\text{sample}}} \frac{O(x_{\text{initial}}, \tilde{m}_{\text{i}}) - O(x_{\text{optimized}}, \tilde{m}_i)}{O(x_{\text{initial}}, \tilde{m}_i)},$$

where $O(x, \tilde{m})$ is the total overflow of the network, given by (1). All the experiments below are done with $n_{\text{sample}} = 100$.

## 4.3 Parameters for the experiments

Before carrying out the experiments that would allow us to compare the models, we will proceed to a calibration of the SAA model and the analytic model. In other words, we will be looking for the values of the following parameters that are expected to offer the best trade-off between overflow reduction and computational time: the number of scenarios in the SAA model, the number of segments in the linear approximation of the analytic expression (i.e., the value of $n_{\max}$ as defined in Section 3.1), and the number of unaggregated demands in the analytic model (i.e., the value of $n_{\text{pces}}$ as defined in Section 3.2).

### 4.3.1 Number of scenarios

Keeping in mind that the SAA model will be used as our standard of comparison, it is of fundamental importance to choose a number of scenarios that will deliver a good overflow reduction while requiring a relatively small solution time. In order to determine this suitable number of scenarios, we proceeded as follows. Starting with 10 scenarios, we increased the number of scenarios by 10 until no improvement could be observed (given that the solver was given an optimality gap of 2 %). This experiment was carried out with instance B. The results are shown in Figure 4. Given that there is a large increase in computational time for more than 50 scenarios and essentially no improvement in terms of overflow reduction, the best trade-off between quality of solution and computational time seems to be reached with about 50 scenarios. Thus, the number of scenarios in all the remaining experiments was fixed to 50. Note that the jumps in the graphs of Figure 4 as well as of all the figures below are typical of hard integer programs.

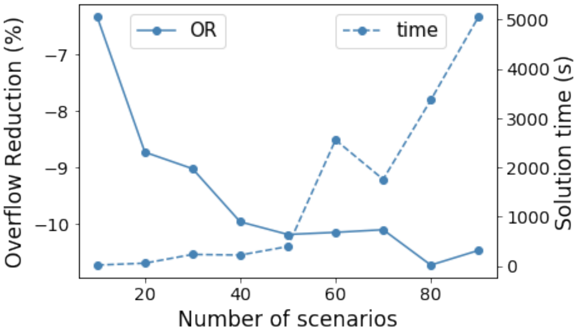

**Figure 4: Overflow reduction and solution time of the SAA model for different numbers of scenarios (with instance B).**

### 4.3.2 Number of unaggregated demands

As mentioned in Section 3.2, the number of terms in the objective function increases exponentially with the number of (unaggregated) demands that could potentially travel on a leg. Therefore, for the analytic model to be solvable in practice, the value of $n_{\max}$ should be kept relatively small. In order to determine a suitable value for $n_{\max}$, we ran an experiment with instance A in which we fixed $n_{\text{pces}} = 10$ and increased gradually the value of $n_{\max}$ starting from 0 (i.e., all the demands are aggregated). Figure 5 shows the results of this experiment. Comparing the overflow reduction and solution time for different maximal numbers of unaggregated demands, we came to the conclusion that $n_{\max} = 4$ offers the best trade-off between quality of solution and computational time. This value of $n_{\max}$ will be used from now on.

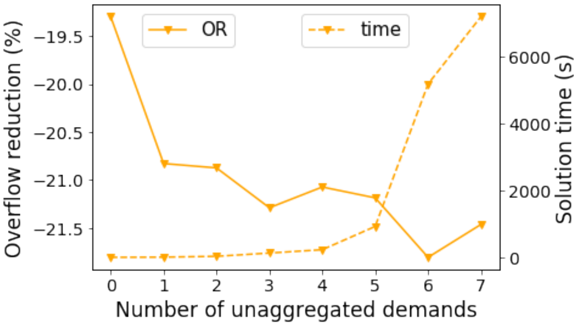

**Figure 5: Overflow reduction and solution time of the analytic model for different values of $n_{\max}$ (with instance A).**

### 4.3.3 Number of pieces in the linear approximation

In order to make the analytic model solvable by CPLEX, we first linearized it using the curve approximation described in Section 3.1.2. Increasing the value of $n_{\text{pces}}$ adds pieces in the linear approximation of the curve representing the the expectation of overflow, which, in turn, makes the model more complex by adding new constraints as well as terms in the objective function. It is then expected that by increasing the value of $n_{\text{pces}}$, the solution time should increase significantly. In order to determine a suitable value for this parameter, we proceeded similarly as in Section 4.3.2. In other words, we compared overflow reduction and solution time for an experiment in which we gradually increased the value of $n_{\text{pces}}$ from 2 to 20. Results of this experiment appear in Figure 6. It can be seen that a fairly good trade-off between quality of solution and computational time is offered for a maximal number of linear pieces in the region $7 \leq n_{\text{pces}} \leq 13$. The value of $n_{\text{pces}}$ was fixed to 10 in the remaining experiments.

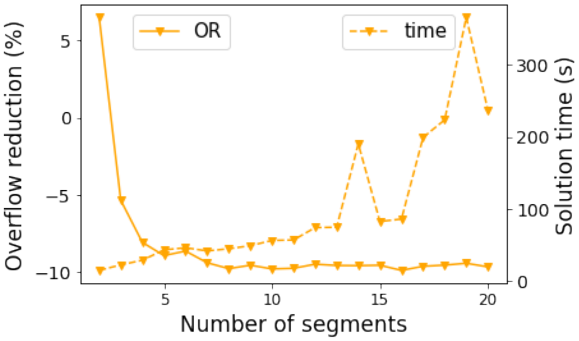

**Figure 6: Overflow reduction and solution time of the analytic model for different values of $n_{\text{pces}}$ (with instance B).**

## 4.4 Comparison of the models

In this subsection, we compare overflow reduction and solution time for the deterministic, the SAA and the analytic models. Given that the deterministic model does not take into account stochasticity of the demand, it should not allow for an overflow reduction that is better than the stochastic models. Moreover, since the SAA and the analytic models both correspond to approximations of the same stochastic model (defined in Section 2.2), their percentage of overflow reduction is expected to be about the same. On the other hand, solution time could differ greatly between all the models. Note that since the deterministic model can be solved in only a few seconds, its solution time was not included in the figures below.

### 4.4.1 Size of the problem

In order to analyze the influence of the size of the problem on overflow reduction and computational time, we ran an experiment in which various proportions of the shipping plan were fixed, i.e., some $x_{d\omega}$ were kept fixed during the resolution. For the sake of the experiment, the fixed demands were chosen randomly. However, in real operations, the airline may want to fix priority shipments or commodities of a specific type (e.g., perishables).

Results for overflow reduction and solution time for instance $B$ are shown in Figure 7. Note that the smaller the percentage on the horizontal axis, the larger the problem. When less than 50 % of the demands are fixed, the deterministic model offers an overflow reduction that is significantly lower than with the stochastic models. As expected, the SAA model and the analytic model deliver similar performances as far as overflow reduction is concerned. However, for smaller percentages of fixed demands, the time needed to solve the SAA model increases considerably. When all demands are allowed to be reassigned, the analytic model is close to ten times faster than the SAA model. Therefore, the analytic model seems to be less sensitive to the size of the problem. For this reason, this model appears to be more suitable to deal with large instances or with medium-size problems in which a demand is not the finest unit (e.g., if a demand is divisible into a fixed number of pieces).

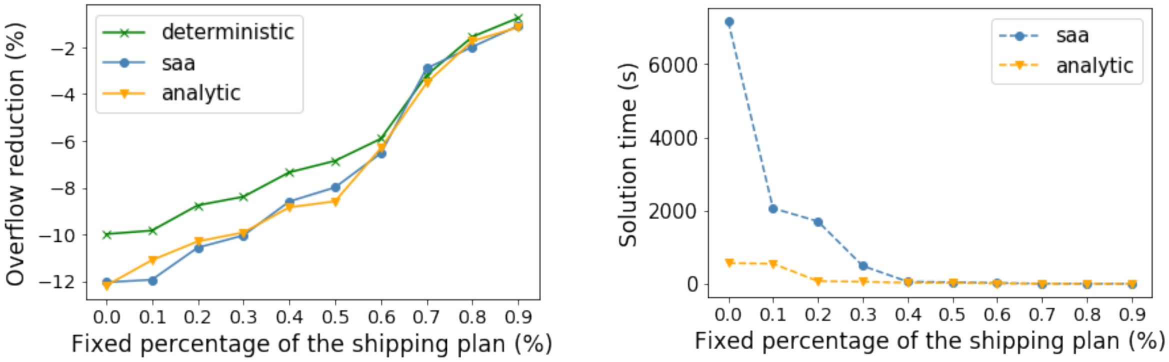

**Figure 7: Overflow reduction (%) and solution time (s) for different fixed percentage of the shipping plan (with instance B).**

### 4.4.2 Maximal number of demand reassignments

As explained in Section 2, reoptimizing the shipping plan for a large number of demands is not always advisable (due to logistical and administrative reasons). To limit the number of demand reassignments, constraint (3) can be added to our three models. We analyze now the influence of this constraint on the quality of the solution. We carried out an experiment in which the maximal number of demand reassignments was made to vary from 100 to 2 500 (out of a maximum of 11 357 possible reassignments). For this experiment, instance C with 30 % of the shipping plan fixed was used. This instance, which has a maximal overbooking rate of 1.05, was chosen on purpose. Indeed, when planes are heavily overbooked, the value of the expected cargo overflow tends to the value of the actual cargo overflow. This can be seen from the curves on the right of Figure 1. When the average of the load factor is high, all the curves, regardless of their variance, tend to the curve with zero variance, i.e., the deterministic case. It follows that great improvements of the shipping plan can be made by reassigning the actual overflow of cargo, which is exactly what the deterministic model does. Further improvements, that are possible only when taking into account the stochasticity of the demands, will then be made only when a relatively large number of demands are allowed to be reassigned. Therefore, an instance with a relatively small overbooking rate was chosen so that the improvements due to stochasticity can stand out. The results of this experiment are presented in Figures 8. For the reasons explained above, all three models deliver similar performances when the number of reassignments is small. However, starting from 500 reassignments allowed, both stochastic models offer a significantly greater overflow reduction than the deterministic one. As for computational time, the solution time of the analytic model does not increase much when more changes are allowed, which is in stark contrast with the SAA model. Indeed, its solution time is more than 10 times larger than the one of the analytic model for a maximal number of changes of 2 500.

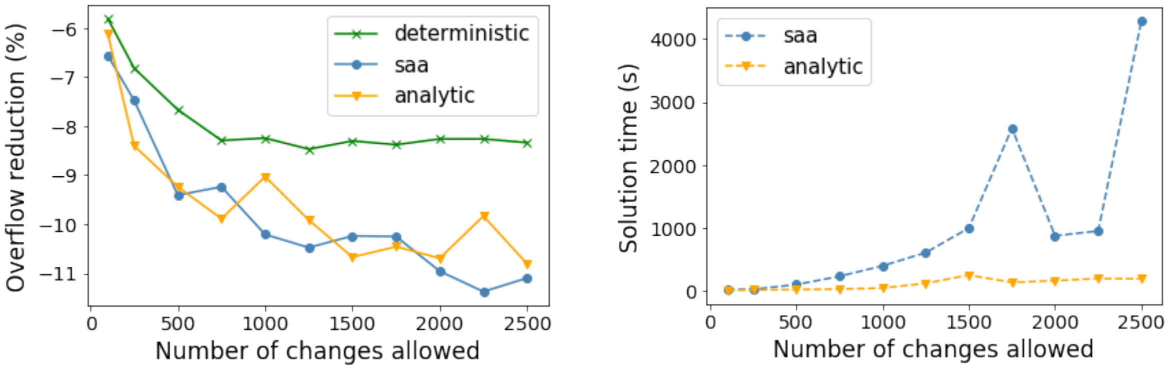

**Figure 8: Overflow reduction (%) and solution time (s) for different maximal numbers of demand reassignments (with instance C).**

### 4.4.3 Robustness

When developing the probability density (4), we assumed that a precise knowledge of the show-up rate distribution is not required to integrate the important stochastic information in the optimization model. We made the hypothesis that only a good estimation of the no-show rate and of the first two moments are necessary. To test this hypothesis, we replaced the normal distribution in (4) by an asymmetric distribution that has for probability density the piecewise constant function represented on the left of Figure 9. This distribution was constructed so that its mean (i.e., 1.00) and standard deviation (i.e., 0.31) are almost the same as the the ones of the normal distribution (i.e., $\mu_d = 1.00$ and $\sigma_d = 0.30$). On the other hand, since its skewness and its excess kurtosis are 0.66 and 0.64 respectively, this distribution is significantly different from a normal distribution (for which skewness and excess kurtosis are both 0). The resulting mixture distribution was then used to generate the weight vector samples (as described in Section 4.2). Moreover, the scenarios of the SAA model were generated from this new mixture distribution.

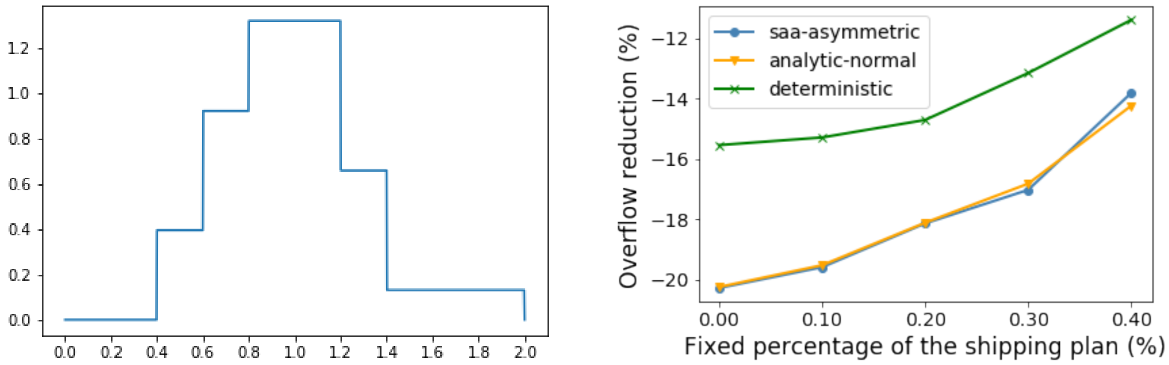

**Figure 9: Overflow reduction with the asymmetric distribution (with instance A).**

The graph on the right of Figure 9 shows overflow reduction for different fixed percentages of the shipping plan when tested with weight vector samples generated from the asymmetric distribution. The curves of the analytic model (assuming a normal distribution) and the SAA model (with scenarios from the asymmetric distributions) almost overlap even though only the SAA model assumes the same distribution as the one with which it is tested. This suggests that as long as the no-show rate, the mean and the variance can be well estimated, the analytic model is, at least to some extent, robust to the precise shape of the show-up rate distribution.

The fact that the SAA model does not perform better than the analytic model might be a consequence of the higher load factors of instance A. Indeed, as shown in Figure 1, the higher the load factor, the less meaningful an error on the variance of the load factor is on the expectation of overflow. As a result, the analytic model might not perform as well with lower load factors. However, this is not a major limitation since minimizing the expectation of overflow primarily makes sense in the context of overbooking.

## 5 Conclusion

In this paper, we introduced new stochastic models to address the cargo overflow reduction problem in an air transportation network. One of the main novelties of our approach is to represent the show-up rate using the mixture of a normal and a constant distributions. It has the advantage of capturing the important no-show phenomenon while still taking into account variability in the weight of the goods that are actually brought to the airport. We provided an analytic expression for the expectation of overflow, as well as different ways of approximating it as a piecewise linear or quadratic function. This approximated expression can then be used as the objective function of a stochastic integer program that can be solved without resorting to a standard scenario-based approach. Since no scenarios have to be constructed, input preparation time is considerably reduced. Only an estimation of the no-show rate and of the first two moments of the show-up rate distribution are required.

The analytic model has then been implemented and tested against an implementation of the SAA model. For all the instances, our implementation of the analytic model has shown to provide overflow reductions similar to the ones obtained by the SAA model, but with calculation times much smaller. When the problem at hand is large, the gain in calculation time can be up to a factor of 10. Moreover, our approach has shown robustness with regards to the shape of the actual show-up rate distribution.

It is important to mention that this research does not represent yet a valid economic study of the cost savings that could be realized from a reoptimization of the shipping plan because the required actual demand data were not available to the authors at the time of the research. However, it constitutes an important first step by determining a way to reduce calculation time without significantly altering the quality of solutions. Furthermore, this paper identifies a reduced set of demand information required in order to obtain a good model. An important second step would be to evaluate the efficiency of our approach by testing our model with an industrial dataset. Additionally, further research should be conducted to study how the analytic model could be extended to deal with both weight and volume at the same time.

As a final remark, note that even though our approach was developed with a view to applications in the air cargo industry, it could also be suitable for other contexts where the phenomenon of no-show represents an important aspect of planning.

## 6 References

K. Amaruchkul and V. Lorchirachoonkul. Air-cargo capacity allocation for multiple freight forwarders. Transportation Research Part E: Logistics and Transportation Review, 47(1):30–40, 2011. ISSN 1366-5545. doi: https://doi.org/10.1016/j.tre.2010.07.008.

F. Azadian, A. E. Murat, and R. B. Chinnam. Dynamic routing of time-sensitive air cargo using real-time information. Transportation Research Part E: Logistics and Transportation Review, 48(1):355–372, 2012. ISSN 1366-5545. doi: https://doi.org/10.1016/j.tre.2011.07.004.

C. Barz and D. Gartner. Air cargo network revenue management. Transportation Science, 50(4):1206–1222, 2016.

Boeing. World air cargo forecast 2016–2017. 2016.

B. Feng, Y. Li, and Z.-J. M. Shen. Air cargo operations: Literature review and comparison with practices. Transportation Research Part C: Emerging Technologies, 56:263–280, 2015.

D. L. Han, L. C. Tang, and H. C. Huang. A markov model for single-leg air cargo revenue management under a bid-price policy. European Journal of Operational Research, 200(3):800–811, 2010.

R. Kasilingam. Air cargo revenue management: Characteristics and complexities. European Journal of Operational Research, 96:36–44, 02 1997a. doi: 10.1016/0377-2217(95)00329-0.

R. G. Kasilingam. An economic model for air cargo overbooking under stochastic capacity. Computers & industrial engineering, 32(1):221–226, 1997b.

Y. Levin, M. Nediak, and H. Topaloglu. Cargo capacity management with allotments and spot market demand. Operations Research, 60, 04 2012. doi: 10.2307/41476362.

T. Levina, Y. Levin, J. McGill, and M. Nediak. Network cargo capacity management. Operations Research, 59(4):1008–1023, 2011. doi: 10.1287/opre.1110.0929.

S. Luo, M. Çakanyildirim, and R. G. Kasilingam. Two-dimensional cargo overbooking models. European Journal of Operational Research, 197(3):862–883, 2009. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2007.09.047.

O. E. Nahum, Y. Hadas, and A. Kalish. A combined freight and passenger planes cargo allocation model. Transportation Research Procedia, 37:354–361, 2019. ISSN 2352-1465. doi: https://doi.org/10.1016/j.trpro.2018.12.203.

J. Patomtummakan and N. Nananukul. Air cargo decision support system. International Information Institute (Tokyo). Information, 20(4A):2405–2415, 2017.

Y. Peng, P. Wang, X. Zhao, M. Chen, J. Zhang, and F. Zhang. A data-driven air cargo redistribution model based on multiple programming. International Journal of Modern Physics B, 33(17):1950176, 2019.

A. Popescu, P. Keskinocak, E. Johnson, M. LaDue, and R. Kasilingam. Estimating air-cargo overbooking based on a discrete show-up-rate distribution. INFORMS Journal on Applied Analytics, 36(3):248–258, 2006. doi: 10.1287/inte.1060.0211.

B. Shepherd, A. Shingal, and A. Raj. Value of air cargo: Air transport and global value chains. Montreal: The International Air Transport Association (IATA), 2016.

M. Steadie Seifi. Multimodal transportation for perishable products. PhD thesis, Department of Industrial Engineering & Innovation Sciences, 3 2017.

Y.-J. Wang and C.-S. Kao. An application of a fuzzy knowledge system for air cargo overbooking under uncertain capacity. Computers & Mathematics with Applications, 56(10):2666–2675, 2008. ISSN 0898-1221. doi: https://doi.org/10.1016/j.camwa.2008.02.049.

A. Wannakrairot and N. Phumchusri. Two-dimensional air cargo overbooking models under stochastic booking request level, show-up rate and booking request density. Computers & Industrial Engineering, 100:1–12, 2016. ISSN 0360-8352. doi: https://doi.org/10.1016/j.cie.2016.08.001.

C. Zhang, R. Luo, and Z. Chen. An optimization model of cargo space allocation for air cargo agent. In 2010 7th International Conference on Service Systems and Service Management, pages 1–5, June 2010. doi: 10.1109/ICSSSM.2010.5530228.