Chapter 3
Dealing with Demand Uncertainty using Sample Average Approximation

The sample average approximation (SAA) method is an approach for solving stochastic optimization problems by using Monte Carlo simulation. In this technique proposed by Shapiro and Homem-de Mello (1998), the expected objective function of the stochastic problem is approximated by a sample average estimate derived from a random sample. The resulting sample average approximating problem is then solved by deterministic optimization techniques. The process is repeated with different samples to obtain candidate solutions along with statistical estimates of their optimality gaps.

This approach has been used in the literature as a method of avoiding the difficulty of dealing with a large number of scenarios. Linderoth and Wright (2003) applied the SAA technique to several linear programming models using parallelization in a computational grid to accelerate the solution obtaining process. Kleywegt et al. (2002) presented important theoretical considerations regarding the method for combinatorial problems and illustrated them with numerical examples of some applications. Verweij et al. (2003) demonstrated the application of SAA to routing problems with large numbers of scenarios (up to 21,694) and obtained solutions with optimality gaps of approximately 1.0%. Santoso et al. (2005) proposed an application that was specifically aimed at supply chain design and applied it to a real study of the beverage industry. More recently, Schütz et al. (2009) applied the SAA methodology to a supply chain design problem for food products.

In this chapter we present the development of SAA techniques to deal with the demand uncertainty considered in the stochastic programming model presented in chapter 2. We show how we can approximate the solution by means of statistical bounds to be obtained by repeatedly solving the problem considering samples from the original scenario set. Moreover, we show how one can use the SAA technique to estimate the minimum number of scenarios that guarantee certain statistical properties for the estimated optimal solution under a Monte Carlos sampling framework. At last, we present a case study

where the mathematical model proposed in chapter 2 is used to study the supply chain investment planning process for the distribution of petroleum products in northern Brazil.

3.1 Sample Average Approximation

Consider the problem:

$$v^\star = \min_{x \in X} \left\{ f(x) = \mathbb{E}_\Omega \left[F(x, \xi) \right] = \int_\Omega G(x, \xi) g(x) dx \right\} \tag{3.1}$$

where g is the density function of ξ. Note that the two-stage stochastic programming problem with recourse is a particular instance of problem 3.1. This can be straightforwardly seen if one defines

1. $X = \{x \mid Ax = b\}$
2. $f(x) = c^T x + \mathcal{Q}(x)$
3. $\mathcal{Q}(x) = \mathbb{E}_\Omega[Q(x, \xi)]$
4. $Q(x, \xi) = \min_y \left\{ q^T y \mid Wy = h(\xi) - Tx \right\}$

where x is a n-dimensional vector of first-stage variables, A is a $m \times n$ matrix, b is a m-dimensional vector, c is a n-dimensional vector representing the first-stage decision costs, $\xi \in \Omega$ represents the possible realizations of uncertainty, y is a p-dimensional vector representing the second-stage decisions, T and W are matrices of size $q \times n$ and $q \times p$, respectively, q is a p-dimensional vector representing the second-stage costs, and h is m-dimensional vector.

The main difficulty in solving problem 3.1 is related with the calculation of the expected value $\mathbb{E}_\Omega \left[F(x, \xi) \right]$ due to its multi-dimensional characteristics. The approach proposed in the SAA method seeks to obtain an approximation of this value, by considering a sample of N realizations of the random variable ξ. In this sense, following Shapiro and Homem-de Mello (1998) we can define our Sample Average Approximation (SAA) problem as

$$\hat{v}_N = \min_{x \in X} \left\{ \tilde{f}_N(x) = \frac{1}{N} \sum_{n=1,\ldots,N} F(x, \xi^n) \right\} \tag{3.2}$$

Let $\hat{y}_N$ denote the optimal solution of problem 3.2. Note that $\hat{v}_N$ and $\hat{y}_N$ are random in the sense that they are functions of the corresponding random sample. However, for a particular realization $\xi^1, \ldots, \xi^N$ of the random sample, problem 3.2 is deterministic and, thus, can be solved by appropriate optimization techniques.

Since we are trying to approximate $f(x)$, it is important to keep in mind two important properties of the SAA problem 3.2:

Property 1. *$\tilde{f}_N(x)$ consists of an unbiased estimator for $f(x)$.*

Proof: It is not difficult to see that:

$$\mathbb{E}_\Omega\left[\tilde{f}_N(x)\right] = \frac{1}{N}\mathbb{E}_\Omega\left[\sum_{n=1,\ldots,N} F(x,\xi^n)\right] = \frac{1}{N}Nf(x) = f(x) \quad \square \tag{3.3}$$

Property 2. *$\hat{v}_N$ is a lower bound for $v^\star$.*

Proof: Note that:

$$v^\star = \min_{x\in X}\left\{\mathbb{E}_\Omega\left[F(x,\xi)\right]\right\} = \min_{x\in X}\left\{\mathbb{E}_\Omega\left[\frac{1}{N}\sum_{n=1,\ldots,N} F(x,\xi^n)\right]\right\} \tag{3.4}$$

With this in mind, we can then state the following:

$$\min_{x\in X}\left\{\frac{1}{N}\sum_{n=1,\ldots,N} F(x,\xi^n)\right\} \leq \frac{1}{N}\sum_{n=1,\ldots,N} F(x,\xi^n) \tag{3.5}$$

Taking the expectation on both sides, we have that:

$$\mathbb{E}_\Omega\left[\min_{x\in X}\left\{\frac{1}{N}\sum_{n=1,\ldots,N} F(x,\xi^n)\right\}\right] \leq \mathbb{E}_\Omega\left[\frac{1}{N}\sum_{n=1,\ldots,N} F(x,\xi^n)\right] \tag{3.6}$$

According with 3.2, we can rewrite 3.20 as

$$\mathbb{E}_\Omega\left[\hat{v}_N\right] \leq \mathbb{E}_\Omega\left[\frac{1}{N}\sum_{n=1,\ldots,N} F(x,\xi^n)\right] \tag{3.7}$$

which implies that

$$\mathbb{E}_\Omega\left[\hat{v}_N\right] \leq \min\left\{\mathbb{E}_\Omega\left[\frac{1}{N}\sum_{n=1,\ldots,N} F(x,\xi^n)\right]\right\} = v^\star \quad \square \tag{3.8}$$

(a) Lower bound approximation

Provided the above properties, we are still left with the task of calculating the lower bound $\mathbb{E}_\Omega[\hat{v}_N]$, which again is not a trivial task. To circumvent this drawback, we rely on a sampling approach to come up with an approximation for it. For this purpose, we generate M independent samples $\xi^{nm}, n = 1,\ldots,N, m = 1,\ldots,M$. For each batch m of N samples, we solve the following SAA problem

$$\hat{v}_N^m = \min_{x\in X}\left\{\frac{1}{N}\sum_{n=1,\ldots,N} F(x,\xi^{nm})\right\} \tag{3.9}$$

Each of the M problems 3.9 provides a realization of the random variable $\hat{v}_N$. Therefore, we can define an approximation for $\mathbb{E}_\Omega[\hat{v}_N]$ as

$$L_{NM} = \frac{1}{M} \sum_{m=1,\ldots,M} \hat{v}_N^m \tag{3.10}$$

Following the ideas that we used to demonstrate Property 1, it is straightforward to see that L_{NM} represents an unbiased estimate for $\mathbb{E}_\Omega[\hat{v}_N]$ and therefore, a good candidate to approximate the lower bound of the original problem 3.1. To construct a confidence interval for L_{NM}, we can build upon the Central Limit Theorem, which states that

$$\sqrt{M}\,[L_{NM} - \mathbb{E}_\Omega[\hat{v}_N]] \Rightarrow \mathcal{N}(0, \sigma_L^2) \tag{3.11}$$

where σ_L^2 is the variance of $\hat{v}_N^m, m = 1, \ldots, M$, and "$\Rightarrow$" denotes distributional convergence to a normal distribution with mean 0 and variance σ_L^2. To approximate σ_L^2, we can use the sample variance estimator s_L^2, which is defined as

$$s_L^2 = \frac{1}{M-1} \sum_{m=1,\ldots,M} (\hat{v}_N^m - L_{NM})^2 \tag{3.12}$$

And finally, provided a tolerance α, we can define a $(1-\alpha)\%$ confidence interval for L_{NM} as

$$\left[L_{NM} - \frac{z_\alpha s_L}{\sqrt{M}}, L_{NM} + \frac{z_\alpha s_L}{\sqrt{M}} \right] \tag{3.13}$$

where z_α is the standard normal deviate such that $P(z \leq z_\alpha) = 1 - \alpha$.

(b) Upper bound approximation

An upper bound can be obtained by noting that, for any feasible solution $\hat{x}$, we have immediately from 3.1 that $f(\hat{x}) \geq v^\star$. Therefore, by selecting $\hat{x}$ to be a near-optimal solution, for example using the SAA problem 3.5, and by using some unbiased estimator of $f(\hat{x})$, we can obtain an estimate of an upper bound for $v^\star$. To obtain such an estimate, we generate T independent samples $\xi^{nt}, n = 1, \ldots, \overline{N}; t = 1, \ldots, T$ and define

$$\hat{f}_{\overline{N}}^t(\hat{x}) = \frac{1}{\overline{N}} \sum_{n=1,\ldots,\overline{N}} F(\hat{x}, \xi^{nt}) \tag{3.14}$$

which is again unbiased estimator for $f(x)$ and $\overline{N}$ is such that $\overline{N} > T > N$. We highlight the idea of using a larger sample size $\overline{N}$ and sample batch size T, in this case, in order to improve precision. In general, when it comes to two-stage stochastic programming problems, the evaluation of f provided a fixed solution $\hat{x}$ is not computationally demanding and can also benefits from decomposition and parallelization techniques.

We can then use the average value defined by

$$U_{\overline{N}T}(\hat{x}) = \frac{1}{T} \sum_{t=1,\ldots,T} \hat{f}^t_{\overline{N}}(\hat{x}) \tag{3.15}$$

as an estimate of $f(\hat{x})$. Note that here we consider the upper bound estimator $U_{\overline{N}T}(\hat{x})$ as dependent on the solution $\hat{x}$ selected. In the same spirit of what we did for the lower bound, by applying the Central Limit Theorem, we have that

$$\sqrt{T}\,[U_{\overline{N}M} - f(\hat{x})] \Rightarrow \mathcal{N}(0, \sigma^2_U) \tag{3.16}$$

where σ^2_U is the variance of $\hat{f}^t_{\overline{N}}(\hat{x}), t = 1, \ldots, T$, and "$\Rightarrow$" denotes distributional convergence to a normal distribution with mean 0 and variance σ^2_U. We can replace σ^2_U by the sample variance estimator s^2_U, which is given by

$$s^2_U = \frac{1}{T-1} \sum_{t=1,\ldots,T} (\hat{f}^t_{\overline{N}}(\hat{x}) - U_{\overline{N}T}(\hat{x}))^2 \tag{3.17}$$

And finally, provided a tolerance α, we can define a $(1-\alpha)\%$ confidence interval for $U_{\overline{N}T}(\hat{x})$ as

$$\left[U_{\overline{N}T}(\hat{x}) - \frac{z_\alpha s_U}{\sqrt{T}}, U_{\overline{N}T}(\hat{x}) + \frac{z_\alpha s_U}{\sqrt{T}} \right] \tag{3.18}$$

where z_α is the standard normal deviate such that $P(z \leq z_\alpha) = 1 - \alpha$.

(c) Estimating the gap

Provided that we have available estimates 3.10 and 3.15, we may wish to estimate the optimality gap $f(\hat{x}) - v^\star$. Consider the difference

$$GAP_{NM\overline{N}T}(\hat{x}) = U_{\overline{N}T} - L_{NM} \tag{3.19}$$

It follows by the Law of Large Numbers that $GAP_{NM\overline{N}T}(\hat{x})$ converges to $f(\hat{x}) - v^\star$ with probability one as N, M, $\overline{N}$, and T tends to ∞. Moreover, since that $\hat{x}$ is not the optimal solution, then $f(\hat{x}) - v^\star$ is strictly positive. The

variance s^2_{GAP} of $GAP_{NM\overline{N}T}(\hat{x})$ is then estimated by

$$s^2_{GAP} = s^2_U + s^2_L$$

We have to keep in mind that three different sources of uncertainty contributes to the error in the statistical estimator $GAP_{NM\overline{N}T}(\hat{x})$ of the gap $f(\hat{x}) - v^\star$, namely

1. variance of $U_{\overline{N}T}$
2. variance of L_{NM}
3. bias $v^\star - \mathbb{E}_\Omega[\hat{v}_N]$

Remind that $U_{\overline{N},T(\hat{x})}$ and L_{NM} are unbiased estimators of $f(\hat{x})$ and $\mathbb{E}_\Omega[\hat{v}_N]$, respectively. Moreover, their variances can be estimated from the samples and may be reduced by either increasing sample sizes $\overline{N}$, M, and T. In addition to that, we have that $GAP_{NM\overline{N}T}(\hat{x})$ is an unbiased estimator of $f(\hat{x}) - \mathbb{E}_\Omega[\hat{v}_N]$, and that $f(\hat{x}) - \mathbb{E}_\Omega[\hat{v}_N] > f(\hat{x}) - v^\star$. That is, $GAP_{NM\overline{N}T}(\hat{x})$ overestimates the true gap $f(\hat{x}) - v^\star$, and has bias $v^\star - \mathbb{E}_\Omega[\hat{v}_N]$. Shapiro and Homem-de Mello (1998) show that, for ill conditioned problems, this bias may be relatively large and tends to zero at a rate of $O\left(N^{-1/2}\right)$. Therefore, the bias can be reduced by increasing the sample size N of the SAA problem 3.2 or by using a more sophisticated sampling technique (by using Latin Hypercube Sampling, for example). Nevertheless, an increase in N leads to a larger problem instance to be solved, while increases in $\overline{N}$, M, and T to reduce components 1 and 2 of the error lead only to more instances of the same size to be solved.

3.2 Scenario generation using SAA

In stochastic programming approaches, a random process can be either represented by continuous or discrete random variables. However, stochastic programming problems with continuous random variables can only be solved in small or illustrative examples in the best case. In fact, it is frequently impossible to evaluate a possible solution in this kind of problems. For this reason, the discrete representation of random variables using a finite set of possible outcomes becomes essential in actual decision-making problems under uncertainty.

In order to create a discrete representation of the random phenomenon considered in the model presented in chapter 2, we rely on an sampling strategy. That is, after identifying a particular model that best represents the continuous stochastic process, a repeated random generation of this model is performed to

produce a discrete approximation in the form of a scenario set. Consequently, in order for this approximation to be accurate, a high number of scenarios is usually necessary. Provided that the computational burden of a stochastic programming model rapidly increases with the number of scenarios, we must carefully manage the size of the scenario set in order to reconcile scenario generation and computational tractability.

We can use the framework present on section 3.1 as means of managing the scenario set size. Following this idea, it is possible to rely on the sampling framework to achieve prespecified confidence levels. Kleywegt et al. (2002) showed that, for combinatorial problems such as 2.1 - 2.18, assuming that the SAA problem 3.2 is solved up to a optimality gap δ, the sample size required to ensure the attainment of ϵ-optimality with probability $1-\alpha$ can be bounded by:

$$N \geq \frac{3\sigma_{max}^2}{(\epsilon-\delta)^2} \log\left(\frac{2^n}{\alpha}\right) = \frac{3\sigma_{max}^2}{(\epsilon-\delta)^2} \left[n \log 2 - \log \alpha\right] \tag{3.20}$$

where $\epsilon \geq \delta$, $\alpha \in [0,1]$, $n = |A_K| \times |L_K| \times |T|$, and 2^n represents the total number of possible first-stage solutions, considering that all first-stage variables are binary. In 3.20, the term σ_{max}^2 is defined as the maximal variance of certain function differences in the optimal solution (Kleywegt et al., 2002). The main drawback related with bound 3.20 is that it can be highly conservative for practical applications, thus yielding large sample sizes. Nevertheless, 3.20 suggests that the sample size required to reach complete convergence grows at most linearly with the size of the first-stage variable solution space.

A practically convenient alternative for estimating the minimum number of scenarios can be reached by the use of confidence intervals for the objective value of the SAA problem 3.2. Recall that the expected cost value is a random variable itself in this context, we can use sampling theory to obtain an estimation of the sample size N, based on the degree of confidence expected for the solution(Kleywegt et al., 2002). Following this idea, let

$$\hat{g}_N(w,y) = \min_{w,y} \left\{ \sum_{l,t} CKL_{lt} w_{jt} + \sum_{a,t} CKA_{at} y_{at} + \sum_{n=1}^{N} \frac{1}{N} Q(w,y,\xi^n) \right\} \tag{3.21}$$

be the optimal objective function for the SAA problem, provided the given sample $\xi^1, \ldots, \xi^N$ of size N, and let

$$g_n(w,y,\xi^n) = \sum_{l,t} CKL_{lt} w_{jt} + \sum_{a,t} CKA_{at} y_{at} + Q(w,y,\xi^n) \tag{3.22}$$

be the objective function evaluated for scenario ξ^n. The Monte Carlo sampling variance estimator of the result for this stochastic programming problem is given by

$$s_N = \sqrt{\frac{\sum_{n=1}^{N} \left(\hat{g}_N(w,y) - g_n(w,y,\xi^n)\right)^2}{N-1}} \tag{3.23}$$

We can then state the $1-\alpha$ confidence interval for $\hat{g}_N(w,y)$ as

$$\left[\hat{g}_N(w,y) - \frac{z_{\alpha/2}s_N}{\sqrt{N}}, \hat{g}_N(w,y) + \frac{z_{\alpha/2}s_N}{\sqrt{N}}\right]$$

where $z_{\alpha/2}$ is the standard normal deviate such that $P(z \leq z_{\alpha/2}) = 1 - \alpha/2$. Finally, once we define a maximum percent deviation β, we have that

$$N = \left(\frac{z_{\alpha/2}s_N}{(\beta/2)\hat{g}_N(w,y)}\right)^2 \tag{3.24}$$

Note that the term $(\beta/2)\hat{g}_N(w,y)$ represents the fraction of the total cost one wishes to consider as the confidence interval absolute size for each side. For example, if one wishes to attain a confidence interval of 5% around the expected total cost, then $\beta = 0.1$. In practical terms, the choice of the number of scenarios should take into account the trade-off between the computational effort to obtain a solution and the quality level required for the solution.

3.3 Case Study

In this section we present an application of the SAA technique combined with the model presented in chapter 2 to a real case study on the distribution of petroleum products in northern Brazil.

(a) Case description

The transport in the region considered is primarily performed using waterway modals, which are strongly affected by seasonality issues regarding the navigability of rivers. For this study, four different products were considered - diesel, gasoline, aviation fuel and fuel oil - to be distributed over 13 bases, 3 of which have sea terminals. Three supply sources were considered including one refinery and two external supply locations. The external supply, coming from Paulínia (SP) and São Luiz (MA), represents the connection of the regional logistics network under study with the rest of the country. The case study does not include international commercialization. Four distinct modes of transportation are considered including waterways (using large ferries and

small boats), roadways and pipelines. Waterway transportation is generally performed by large ferries, which are typically used during periods of river flooding, and by smaller boats, which are able to navigate the rivers during droughts (i.e., low water level seasons). However, the use of small boats as means of transportation is expensive and only carried out when the use of large ferries is not possible.

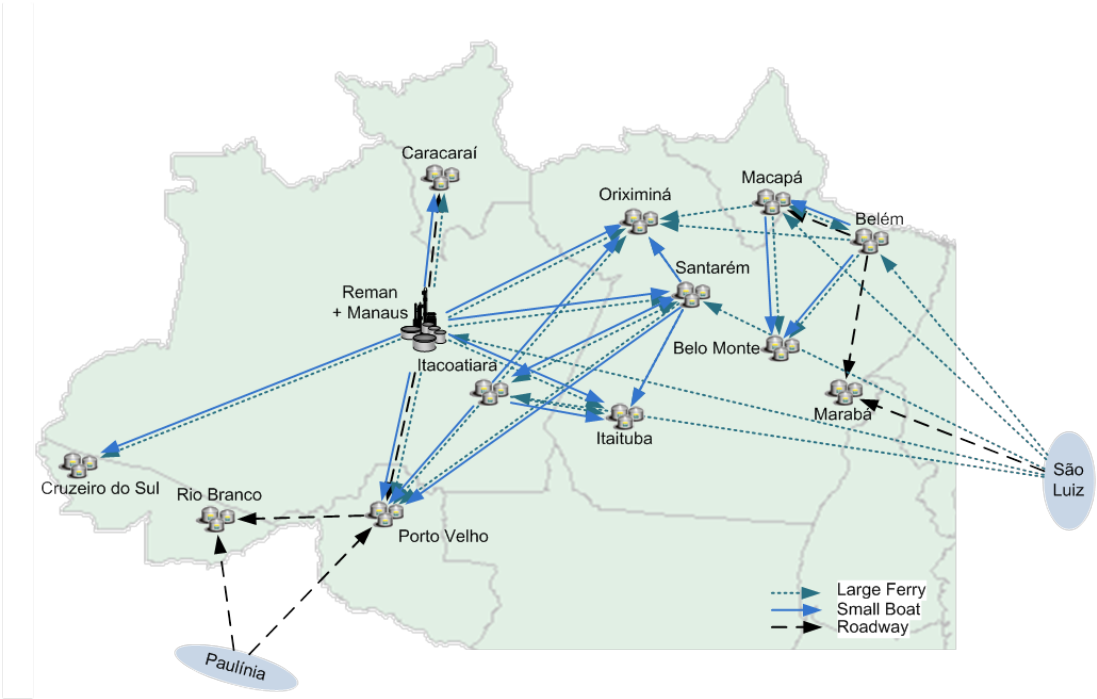

Figure 3.1: Case study distribution network

Figure 3.1 schematically represents the network under study. The region considered comprises approximately 3.7 million km^2, which represents nearly 43% of Brazil's national territory. As shown in this figure, the bases of Manaus (AM), Itacoatiara (AM), Santarém (PA), Macapá (AP), and Belém (PA) are particularly relevant because they act also as distribution points of the supply coming from São Luiz (MA).

Depending on the season, these arcs may or may not be available for navigation. Table 3.1 shows how the seaworthiness was modeled in various parts of the region under study. The checkmarks represent periods in which the given stretch is available for navigation by that mode. Observe that during certain times of the year, the Cruzeiro do Sul base remains completely isolated from communication with the system (3^{rd}quarter), while the base of Caracaraí can only rely on supply via the roadway mode during the first two quarters of the year.

Origin	Destination	Mode	1st Q.	2nd Q.	3rd Q.	4th Q.
Itacoatiara	Itaituba	Ferries			✓	✓
		Small Boats	✓	✓		✓
Itacoatiara	Porto Velho	Ferries	✓	✓	✓	✓
		Small Boats	✓	✓	✓	✓
Manaus	Caracaraí	Ferries				✓
		Small Boats			✓	
Manaus	Cruzeiro do Sul	Ferries		✓		
		Small Boats	✓			✓
Manaus	Itaituba	Ferries			✓	✓
		Small Boats	✓	✓		
Manaus	Porto Velho	Ferries			✓	
		Small Boats	✓	✓		✓
Santarém	Itaituba	Ferries			✓	✓
		Small Boats	✓	✓		
Santarém	Porto Velho	Ferries			✓	✓
		Small Boats	✓	✓		✓

Table 3.1: Seaworthiness between locations

Figure 3.2 shows the level of demand for each of the bases. Manaus (AM) is the main hub of the region's demand, followed by Porto Velho (RO), Belém (PA) and Macapá (AP).

The portfolio of projects considered in the study consists of 28 local projects and one arc project. Such projects are considered mutually independent and can therefore be combined. Table 3.2 represents the portfolio of investments considered, showing the site where each investment will be conducted and the type of project.

	Locations					
Projects	Manaus	Macapá	Santarém	Belém	Cruzeiro do Sul	Itacoatiara
Diesel tank	✓	✓	✓	✓	✓	✓
Gasoline tank	✓	✓	✓	✓	✓	✓
Av. Fuel Tank	✓	✓	✓	✓	✓	✓
Fuel Oil tank	✓		✓	✓		
Pumps/sub.	✓	✓	✓			✓
Pier	✓		✓			

Table 3.2: Investment portfolio for locations

Three distinct types of investments are considered at each location including investments in storage capacity that increase the location's capacity for processing and storing a given product, investments in pumps and substations

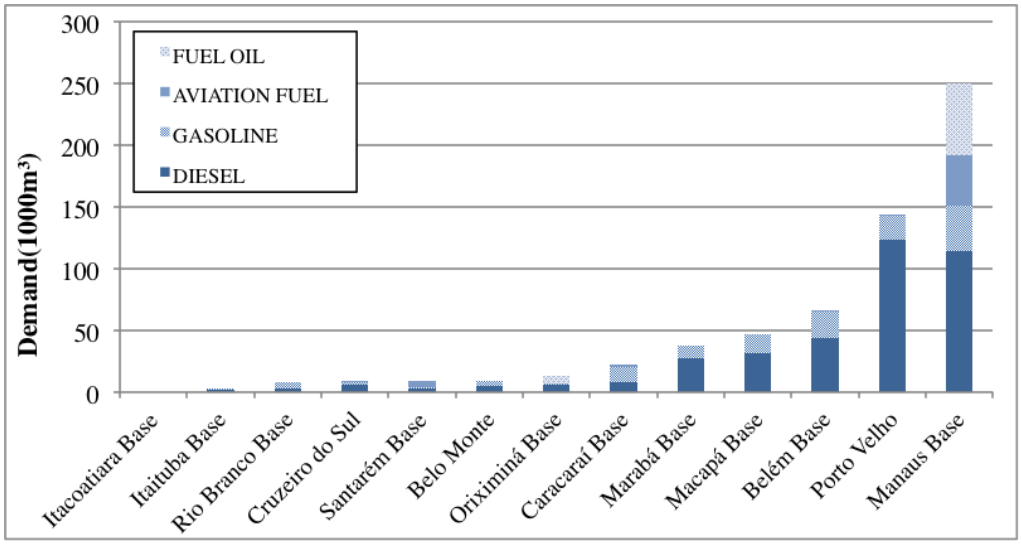

Figure 3.2: Case study demand levels

that reduce the operating costs and increase the ability to rotate the tanks, and investments in the construction of a new pier, which make the demurrage cost curves per handled volume smoother. The investment portfolio also has an investment available for the implementation of a multi-product pipeline that connects the bases of Porto Velho and Rio Branco.

The planning horizon considered was 8 years, which are divided into a total of 32 quarterly periods. All of the costs considered in the model are discounted to a present value under a yearly interest rate of 6.8%.

To take into account the uncertainty in demand levels for petroleum products, we generated scenarios by the following first-order autoregressive model:

$$D_{lpt} = D_{lpt-1}\left[1 + \omega_p + \sigma_p \epsilon\right] \tag{3.25}$$

where ω_p represents the forecasted average growth rate for the consumption of product p over the planning horizon, σ_p represents the estimated maximum deviation of the product p consumption and ϵ is a random error that follows a standard normal distribution. The estimate of the maximum deviation used was simplified as being identical for each product due to the lack of data regarding the historical consumption of the products in the studied region. This estimation was made based on an analysis of the annual Brazilian petroleum products consumption series over the last 40 years. Each scenario represents a possible demand curve for the entire time horizon considered and for each product and distribution base in the considered problem. Figure 3.3 gives an

example of 50 demand scenarios for diesel consumption in Manaus.

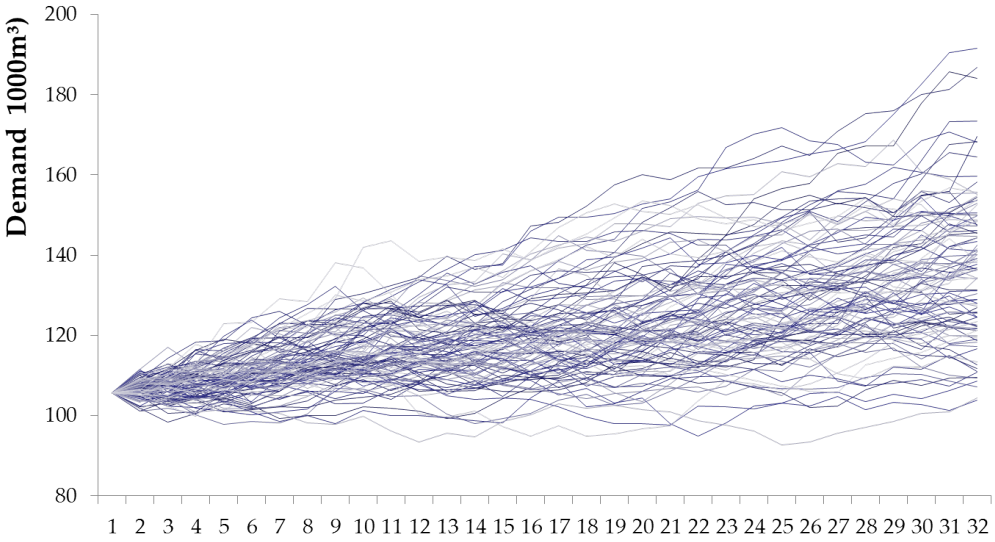

Figure 3.3: Example of demand scenarios

(b) Results

The mathematical model and the scenario generation routines were implemented in AIMMS 3.12. The mixed-integer linear programming (MILP) model was solved using CPLEX 11.2. Table 3.3 describes the size of the instances for the case study in question together with the mean and standard deviation of the solution time for solving each SAA problem. All of the experiments were performed using a Pentium Quad-Core 2.6 GHz with 8 GB RAM. To obtain estimates of the upper and lower limits, experiments were performed with N equal to 20, 30 and 40. These values for N were defined approximating the true values obtained using the estimate of Monte Carlo sampling standard deviation (Equation 3.23) for $N = 50(s_{50})$ considering $\beta = 0.1$, and three different values for α, namely 0.05, 0.025, and 0.01, yielding samples with approximated sizes of 20, 30, and 40, respectively. The average solution time ranged from 532.83s for instances with 20 scenarios to 1,472.05s for those with 40 scenarios. To obtain the lower limits, we performed 50 replications (i.e., $M = 50$), with a time limit of 3,600s and a relative GAP of 1% defined as stop criteria.

N	# Variables	# Constraints	Average(s)	Standard Deviation(s)
20	250400	304144	532.83	967.26
30	455504	374880	971.42	1500.20
40	606864	499360	1472.05	864.41

Table 3.3: Summary of model sizes

Thirty-six distinct candidate solutions were generated for $N = 20$, 22 solutions for $N = 30$, and 19 solutions for $N = 40$. We developed the following experimental procedure in order to avoid the complete (and thus time consuming) evaluation of all candidate solutions. First, all of the candidate solutions were previously assessed with 50 replications. From this first evaluation, we selected the three solutions that showed the best results in terms of solution gap and subsequently further evaluated them with 1000 replications in order to increase the precision of the estimates.

Table 3.4 shows the best results for each experiment in terms of the lower and upper limits estimated for the solution of the real problem. The results suggest that the configuration of the experiment with 50 replications (M = 50) for the lower bound was considered satisfactory given that the deviation obtained for the lower limit is approximately 1%. For the upper limit, it should be noted that its variability is related to the number of scenarios considered in obtaining the lower limit, and it is reduced from 13.4% ($N = 20$) to 4.9% ($N = 40$). This effect is related with the fact that, in general, a larger number of scenarios implies a more comprehensive investment profile in terms of its ability to cope with higher demands, which makes the system more robust with respect to variations in the total costs of meeting the demand (i.e., smaller fluctuations in second-stage costs).

N		Lower Limit	Upper Limit
20	Amount(MM$)	800.12	818.76
	St. Dev. (MM$)	9.81	109.66
	% Deviation	1.2%	13.40%
30	Amount(MM$)	801.25	821.67
	St. Dev. (MM$)	10.22	50.63
	% Deviation	1.2%	6.20%
40	Amount(MM$)	805.28	817.12
	St. Dev. (MM$)	8.22	40.03
	% Deviation	1.0%	4.90%

Table 3.4: Experiment results: statistical limits (lower and upper)

Table 3.5 shows the statistics obtained on the estimate of the optimality gap for the three best solutions obtained in each experiment. The experiments suggest a reduction of the estimated variability of the gap for the experiments with larger number of scenarios, which supports the hypothesis that these solutions are close to the real optimal solution of the problem. In practical terms, this optimality gap is considered acceptable, given the uncertainty inherent in the input data that is considered deterministic. This result is noteworthy, especially because this estimator is biased (as discussed in to 3.1(c)), thus, such an estimate always corresponds to an upper limit of the real gap.

N		gap		
	Solution	Value(MM$)	%	St. Dev.(MM$)
20	A	19.61	2.4%	107.41
	B	26.40	3.2%	72.82
	C	18.64	2.3%	110.10
30	A	23.18	2.8%	63.73
	B	26.95	3.3%	82.40
	C	20.42	2.5%	51.57
40	A	17.38	2.1%	44.03
	B	11.83	1.4%	41.22
	C	14.85	1.8%	49.43

Table 3.5: Experiment results: estimative of the optimality gap

Table 3.6 provides the solutions with the lowest GAP obtained from the

three experiments including Solution Number 3 for $N = 20$ and $N = 30$, and Solution Number 2 for $N = 40$. As can be observed from Table 3.6, the profile of investments has little variability between experiments. An investment was made in fuel oil storage for Santarém in the first period in all runs, which indicates the attractiveness of this investment. This may be explained by the central position of Santarém in the petroleum products distribution network of the region and by the low level of tankage for fuel oil currently available at that location. Other investments also tend to have low variability in their positioning along the time horizon. The greatest variability was seen in the investment in pumps and substations in Macapá, which is directly related to the existence of an anticipated increase in demand for fuel oil in Belo Monte[1], which is transported from São Luiz. The solutions also suggest that Santarém is a strategic location for the logistics of products other than fuel oil because the model suggests investing in three tanking projects in the region. Another relevant observation is related to the projects that did not constitute

Project	$N = 20$	$N = 30$	$N = 40$
	Period Invested	Period Invested	Period Invested
Manaus av. fuel	7	7	6
Santarém diesel	21	16	17
Santarém gasoline	24	22	16
Santarém Fuel Oil	1	1	1
Belém diesel	29	24	27
Macapá pumps/sub.	23	27	26

Table 3.6: Investment profiles of solution 3 for $N = 20$, solution 3 for $N = 30$ and solution 2 for $N = 40$

the optimal portfolio. None of the projects for the physical expansion of the marine terminals (piers) is selected for investment, which suggests that the terminal system as modeled is appropriate to the scenarios of demand for the products considered. The pipeline connecting the Porto Velho and Rio Branco bases turned out to be not economically attractive and was not included in the optimal portfolio of investments in any of the simulated scenarios. This is probably because of the high cost of that project and the existence of an alternative road coming from Paulínia that, despite its high costs, is more economically efficient. The demand was completely satisfied in all experiments.

[1]The data used considers the construction of Belo Monte hydroelectric dam, which will be the second-largest hydroelectric dam complex in Brazil and the world's third-largest in installed capacity. As a consequence, it is forecasted an increase in the demand for petroleum products in the region.

3.4 Conclusions

In this chapter we presented the SAA methodology to solve the problem of investment planning in the supply chain of petroleum product distribution in northern Brazil considering the uncertain demand for such products in the region. Moreover we showed how we can use the SAA approach as a scenario reduction technique and how we can organize the experiments in order to obtain statistically certified good solutions. The results show that it was possible to obtain solutions with acceptable estimates of optimality gaps in practical terms (i.e., in terms of the solution quality and the computational time required to obtain the solutions) even with a relatively small number of scenarios. In the proposed approach, it is possible to delineate reasonably acceptable confidence intervals and thereby define the total number of scenarios required to statistically guarantee that the solutions obtained. It is important to highlight that the amount of scenarios required are strongly related with the variability of the recourse cost of the particular instance considered.

The case study showed that from the proposed portfolio, only six projects comprise the optimal portfolio of investments. The results suggest that the Santarém region has a particular strategic importance for the planning as half of these investments were assigned to that region. Another important observation is the finding that many projects in the set of possible investments were not relevant to the optimization of the logistics in the region for the data set considered.

The results seem to be in line with what has been observed in the literature(Linderoth and Wright, 2003; Kleywegt et al., 2002; Verweij et al., 2003; Santoso et al., 2005; Schütz et al., 2009) when it comes to the successful application of the SAA methodology to solve practical large-scale problems. It can be observed that, even for a modest number of scenarios (ranging from 20 to 40, in this case), the method can provide high quality solutions with relatively small optimality gaps. Therefore, the proposed methodology can support the decision making process, while identifying solutions and statistically ensuring its quality without the need for time-consuming discussions of the adequacy of possible methods for generating scenarios.