The Stochastic Guaranteed Service Model with Recourse for Multi-Echelon Warehouse Management

Jörg Rambau, Konrad Schade[1]

Lehrstuhl für Wirtschaftsmathematik
Universität Bayreuth
Bayreuth, Germany

Abstract

The Guaranteed Service Model (GSM) computes optimal order-points in multi-echelon inventory control under the assumptions that delivery times can be guaranteed and the demand is bounded. Our new Stochastic Guaranteed Service Model (SGSM) with Recourse covers also scenarios that violate these assumptions. Simulation experiments on real-world data of a large German car manufacturer show that policies based on the SGSM dominate GSM-policies.

Keywords: Multi-echelon inventory control, guaranteed service model, stochastic programming, integer linear programming, real-world application

1. Introduction

Inventory control for a spare part distribution system follows two goals: deliver as promptly as possible to the end customer and minimize inventory costs. One option to deal with two goals at the same time is to impose a bound for one and optimize the other. For example: minimize inventory cost subject to a given service level, i.e., the fraction of demands that can be served immediately. This is the strategy that is used, e.g., by the so-called *guaranteed-service model*. See [1] which includes the idea of guaranteed service times for the first time, [2] for an extension to a tree structure network, and [3] where the model is extended to acyclic networks. In [4] the model was applied to the spare part distribution system of a large German car manufacturer. See also the work of Inderfurth [5, 6] and Minner [7].

The guaranteed service model characterizes, for a given service level, optimal order-points s for the widely accepted (s, S)-policies in multi-echelon inventory control (see [8] for the classical problem statement and the theoretical motivation for (s, S) policies). It can be considered as an advantage of the GSM that it

Email addresses: joerg.rambau@uni-bayreuth.de (Jörg Rambau), konrad.schade@uni-bayreuth.de (Konrad Schade)

[1]Supported by a grant of "Elitenetzwerk Bayern"

only makes decisions on the *safety stock level* s for the prescribed (s, S)-policy: even though (s, S)-policies may be suboptimal, they are transparent to human operators – it is much easier to make plausibility checks for safety stock levels than for models that computationally produce higher-dimensional decisions in a black-box. An additional advantage is that the GSM can be implemented and (approximately) solved as an integer linear program (see [3]).

The GSM, however, can only handle bounded demands and deterministic delivery times in the network. Extreme demands and missed internal delivery times produce situations that are not captured by the model, and thus the corresponding cost can not be accounted for by the GSM. There are, of course, other policies for multi-echelon inventory control – including sophisticated stochastic service models – with other strengths and weaknesses (see, e.g., [9] for the METRIC system, [10] for a survey, and [11] for a special version of a stochastic service model). In particular, in stochastic service models adding further restrictions, e.g., imposed by the business processes of a company, can render the method impractical, where as adding restrictions to the ILP model of the GSM to a certain extent does not affect the solution procedure too much.

Our contribution: We introduce the new *stochastic guaranteed service model with recourse (SGSM)* and apply two versions of it to the inventory control problem in a multi-echelon warehouse system of a spare part distributor. The model is a stochastic enhancement of the guaranteed service model by a recourse component and demand scenario sampling, so that all demand scenarios that are captured by the sampling process are handled inside the model. The benefit is that service levels are now an outcome of the model. The advantage of the GSM ILP model that can take further restrictions is maintained. The drawback is that recourse cost data for the cases of lost demands have to be given. (See [12] for background on stochastic programming.) The contribution of this article goes beyond the conference presentation [13] in the following aspects (among others):

- We introduce the new SGSM with a non-trivial complete recourse consisting of a transportation option besides the penalty cost for non-sales, i.e., requested parts that cannot be delivered in time.
- We solve the SGSM by a combination of sample average approximation with state-of-the art scenario reduction techniques. This way, a better coverage of unlikely but expensive scenarios is achieved without increasing the computation times in the MILP solver. Our new asymmetric distance function for the asymmetric scenario reduction takes into account the influence of the scenario reduction on the result of the optimization. To the best of our knowledge, this is new.
- We present a more comprehensive documentation of extended computational results, including a new comparison to one representative [11] of the class of stochastic service models that could be implemented to cope with our test data.

Simulation results on real-world data of a large German automobile manufacturer and Poisson-distributed demand with real-world intensity forecasts show that our inventory policies based on the SGSM dominate GSM-policies and yield better results than the considered stochastic service time model. One reason for this is, among others, that the service level guarantees of the GSM do not take into account that non-sales can have quite different impact on the total cost, which depends on the particular part and on the number of parts missing. It would be interesting from a theoretical point of view to also check performances on artificial randomized data. For this work, we focussed on the practical impact in real-world applications, for which randomized data is rarely representative. We emphasize that, for this reason, our simulation test is completely independent of the assumptions of the tested models – it rather represents our partner's process as closely as possible.

In the following section we introduce the modeling of the GSM and the SGSM before we show the methods used of scenario generation and scenario reduction in section 3. After the description of the simulation method and some computational results in section 4 we end with some conclusions.

2. Modeling

In this section we first give an introduction to the GSM. We use the ILP modeling approach as in [3]. Then we present the SGSM in two different ways. First, in 2.2 we introduce the SGSM as a two stage stochastic mixed-integer linear program with simple recourse. Second, in 2.3 we show an extension where the recourse action of the locations supplying the end customers are modeled as a transportation problem.

2.1. The Guaranteed-Service-Model

The GSM ILP follows the original work in [3], except for the integrality of the order-points, which is mandatory in spare-part systems with occasionally large, expensive parts at very small stock-levels.

Parameters of the model GSM are:

G	directed graph describing the warehouse network
N	number of warehouses
$N(G)$	set of nodes in G
$A(G)$	set of arcs in G
$D(G)$	set of leaves in G (warehouses delivering to end-customers)
h_i	inventory holding cost in location i
L_i	delivery time to location i
$\bar{s}_i^{\text{out}}$	given service time for a leaf $i \in D(G)$
$\Phi_i(x_i)$	upper bound for the demand in $i \in N(G)$ during the time period x_i

The model GSM uses the following variables for warehouses $i \in N(G)$:

s_i^{in}	service times guaranteed by the predecessors of i
s_i^{out}	service times guaranteed by i for its successors
x_i	time period that i needs to bridge with its inventory (i.e., the time between order and delivery of replenishments from the predecessors of i)
y_i	order-point in i

The model GSM now reads as follows:

$$
\begin{array}{lcll}
\min & \sum_{i=1}^{N} h_i\, y_i & & \\
s.t. & x_i \geq s_i^{\text{in}} - s_i^{\text{out}} + L_i & & \forall i \in N(G) \\
& s_i^{\text{in}} \geq s_j^{\text{out}} & & \forall (j,i) \in A(G) \\
& s_i^{\text{out}} \leq \bar{s}_i^{\text{out}} & & \forall i \in D(G) \\
& y_i \geq \Phi_i(x_i) & & \forall i \in N(G) \\
& x_i,\ s_i^{\text{in}},\ s_i^{\text{out}},\ y_i \geq 0 & & \forall i \in N(G) \\
& y_i \in \mathbb{Z} & & \forall i \in N(G)
\end{array}
$$

This is not quite an ILP yet because of the upper bound on the demand in the location i which is denoted by $\Phi_i(x_i)$. With standard piecewise-linear modelling techniques with additional binary variables, this model can approximately be transformed into an ILP (see [3]).

2.2. The Stochastic Guaranteed-Service-Model with Simple Recourse

We now address two major drawbacks of the GSM: the bounded demand (given by the prescribed service level) and the guaranteed delivery times inside the network. Whenever one of them happens to be violated, an action has to be taken that is not captured by the model which incurs a cost that is not taken into account by the model.

In order to incorporate the two aspects into the model in the simplest way, we introduce simple complete recourse for both delays and unmet demand. That is:

- Whenever the guaranteed delivery time of a warehouse is missed, there is some agent that for some cost per time unit delivers the part in time; this can also be interpreted as a penalty to pay for missed deadlines.
- Whenever a warehouse can not deliver a piece, there is some (other) agent that delivers the piece to the warehouse immediately; this can also be interpreted as a penalty to pay for unmet demand.

Of course, in practice, the recourse may be complete but most probably not simple. A real-world model of the recourse process in use depends on the particular application and requires data about the cost of courier services, the cost of a

damage in reputation, and the like. However, our first goal was to investigate how the recourse model as such would influence the resulting policy. And to this end, simple recourse is already telling, as we will see.

Formally, the SGSM has the following additional scenario and recourse parameters:

S	set of scenarios
p_s	probability of scenario $s \in S$
t_i	cost to compensate for one time unit of late delivery
c_i	cost to compensate for one piece of unmet demand
L_i^s	actual delivery time to i in scenario s
$\Psi_i^s(x_i)$	actual demand in i, during time period x_i in scenario s

Following the idea of simple recourse, the SGSM has the following additional recourse variables:

r_i^s	recourse variable for missed deadlines; "how many time units should be compensated at a cost of t_i per unit?"
q_i^s	recourse variable for missed pieces; "how many pieces should be compensated at a cost of t_i per unit?"

Since there is no obvious implementation of actions in the real world according to these recourse variables, they serve as penalties for each non-sale or missed lead time. The hope is that the SGSM can balance inventory costs and non-sales in a more detailed way than the GSM. At the same time, we maintain the modelling power of the MILP formulation: additional restrictions can be easier incorporated than in stochastic service models we know of.

The two-stage stochastic model SGSM now reads as follows:

$$
\begin{array}{lcl}
\min & \sum_{i=1}^{N} \left(h_i\, y_i + \sum_{s \in S} p_s (t_i\, r_i^s + c_i\, q_i^s) \right) & \\
s.t. & x_i + r_i^s \geq s_i^{\text{in}} - s_i^{\text{out}} + L_i^s & \forall i \in N(G),\ \forall s \in S \\
& s_i^{\text{in}} \geq s_j^{\text{out}} & \forall (j,i) \in A(G) \\
& s_i^{\text{out}} \leq \bar{s}_i^{\text{out}} & \forall i \in D(G) \\
& y_i + q_i^s \geq \Psi_i^s(x_i) & \forall i \in N(G),\ \forall s \in S \\
& x_i,\ s_i^{\text{in}},\ s_i^{\text{out}},\ r_i^s,\ q_i^s \geq 0 & \forall i \in N(G),\ \forall s \in S \\
& y_i,\ q_i^s \in \mathbb{Z} & \forall i \in N(G),\ \forall s \in S
\end{array}
$$

Again, a linearization of $\Psi(x_i)$ can be carried out by standard piecewise-linear modelling with additional binary variables.

2.3. Extension with External Suppliers and Lost Sales

The model with simple recourse from the previous section can be extended by modelling an explicit recourse process. We assume that unmet customer demands are lost. However, internal orders are backlogged. The locations that deliver parts to the end customers can order parts from external suppliers to prevent lost sales.

The external suppliers deliver the parts directly to the end customers so that there is no delay in the delivery. The costs of an order from an external supplier depends on the distance between the ordering location and the supplier. Of course the supplier do not have unlimited stock so that capacity constraints have to be taken into account. To concentrate on these recourse actions we assume that the delivery times in the system are fix. An extension with delivery time uncertainties would be straight forward.

We need some more notation to model the new situation

J	set of external suppliers
C_j	capacity of the external supplier j
q^s_{ji}	recourse variable for parts ordered by location i at supplier j
c_{ji}	costs for location i to order one part from supplier j

This leads us to the following model:

$$
\begin{array}{lcl}
\min & \sum_{i=1}^{N} \left(h_i\, y_i + \sum_{s\in S} p_s \sum_{j\in J} c_{ji}\, q^s_{ji} \right) & \\
s.t. & x_i \geq s_i^{\text{in}} - s_i^{\text{out}} + L_i & \forall i \in N(G) \\
& s_i^{\text{in}} \geq s_j^{\text{out}} & \forall (j,i) \in A(G) \\
& s_i^{\text{out}} \leq \bar{s}_i^{\text{out}} & \forall i \in D(G) \\
& y_i + \sum_{j\in J} q^s_{ji} \geq \Psi^s_i & \forall i \in N(G),\ \forall s \in S \\
& \sum_{i\in D(G)} q^s_{ji} \leq C_j & \forall j \in J,\ \forall s \in S \\
& x_i,\ s_i^{\text{in}},\ s_i^{\text{out}},\ q^s_{ji} \geq 0 & \forall i \in N(G),\ \forall j \in J,\ \forall s \in S \\
& y_i,\ q^s_{ji} \in \mathbb{Z} & \forall i \in N(G),\ \forall s \in S
\end{array}
$$

So far, this model does not have complete recourse. Therefore, we introduce an other recourse variable. As before, we enable for every location the possibility to pay a penalty for a non-sale if it can not deliver the ordered parts. For instance one can provide the customer with a replacement vehicle until the spare part can be delivered and the customer‘s car is fixed. The corresponding penalty recourse variable is denoted by q^s_i, as in the first model, and the penalty costs are denoted by c_i again.

Note, that by using the penalty recourse variables we force complete recourse but account for failure by some cost. The computational results in Section 4.3 suggest that the SGSM policies with the tested penalty values dominate GSM-policies in terms of both inventory and recourse cost, not only total cost. This means, the resulting SGSM policy, internally using those successful penalty

values, will perform better than the corresponding GSM policies also for *any other* penalty values.

We obtain a two stage stochastic model with complete recourse:

$$
\begin{array}{lll}
\min & \sum_{i=1}^{N} \left(h_i\, y_i + \sum_{s\in S} p_s \left(c_i\, q_i^s + \sum_{j\in J} c_{ji}\, q_{ji}^s \right) \right) & \\
s.t. & x_i \geq s_i^{\text{in}} - s_i^{\text{out}} + L_i & \forall i \in N(G) \\
& s_i^{\text{in}} \geq s_j^{\text{out}} & \forall (j,i) \in A(G) \\
& s_i^{\text{out}} \leq \bar{s}_i^{\text{out}} & \forall i \in D(G) \\
& y_i + q_i^s + \sum_{j\in J} q_{ji}^s \geq \Psi_i^s & \forall i \in D(G),\ \forall s \in S \\
& \sum_{i\in D(G)} q_{ji}^s \leq C_j & \forall j \in J,\ \forall s \in S \\
& y_i + q_i^s \geq \Psi_i^s & \forall i \in N(G)\backslash D(G),\ \forall s \in S \\
& x_i,\ s_i^{\text{in}},\ s_i^{\text{out}},\ q_{ji}^s \geq 0 & \forall i \in N(G),\ \forall j \in J,\ \forall s \in S \\
& y_i,\ q_{ji}^s \in \mathbb{Z} & \forall i \in N(G),\ \forall s \in S
\end{array}
$$

3. Scenario Generation and Reduction

An appropriate discrete approximation of the assumed distribution of the stochastic parameters in the model often needs many scenarios. The extensive form of the deterministic equivalent problem grows quite fast with the number of scenarios. This is the reason why we employ scenario reduction as described in Subsection 3.2. But first we wrap-up the basics about Sample-Average-Approximation (SAA) Methods for general discrete approximations of probability distributions in Subsection 3.1.

3.1. SAA-Method for Scenario Generation

To approximate the distributions of the stochastic parameters we generate random numbers according to the assumed distribution. These random numbers build the scenarios in the discrete distribution approximating the real distribution of the stochastic parameters. All samples are assigned probabilities proportional to the number of times they were generated. Sampling techniques like this are quite common in stochastic programming. See for example [12].

The idea of sampling techniques is to approximate a stochastic program

$$f(x) = \min_{x\in X} \left\{ c^T x + \mathcal{Q}(x, \boldsymbol{\xi}) \right\}. \tag{1}$$

Here $\mathcal{Q}(x, \boldsymbol{\xi})$ denotes the expected value of the optimal solution of the second stage problem $Q(x, \xi)$ depending on the actual realization ξ of $\boldsymbol{\xi}$.

Assume there is a possibility to get independent, identically distributed samples $\{\xi^1, \ldots, \xi^S\}$ of $\boldsymbol{\xi}$. The problem

$$\hat{f}(x) = \min_{x\in X} \left\{ c^T x + \sum_{s=1}^{S} Q(x, \xi^s) \right\} \tag{2}$$

can be solved conceptually easily and gives us an unbiased estimator for $f(x)$ the solution of the original problem. Further information to SAA can for example be found in [14].

3.2. Scenario Reduction: The Fast Forward Selection

The goal of scenario reduction is to approximate a discrete distribution with many scenarios by another discrete distribution with significantly fewer scenarios. There are several methods to achieve this goal, usually based on a metric on the space of all possible scenarios (see [15, 16, 17]).

An exact approach to find the best approximation with a fixed number of scenarios is to model the approximation problem as a p-median problem. In order to save computation time, we chose to apply the so-called *fast forward selection*, one of the heuristics introduced in [15, 16, 17].

The approximation of the delivery times and demand distributions is split into two parts. First, a number of samples $S = \{\xi^1, \ldots, \xi^S\}$ is generated according to the assumed distribution. These samples built a first discrete approximation where every scenario instance occurs with equal probability $p_s = 1/S$. Second, the resulting discrete distribution is fed into the fast-forward scenario reduction, i.e., it is approximated by a discrete distribution over a subset of scenarios of prescribed cardinality, which have, in general, non-uniform probabilities.

Let us now sketch the principle of scenario reduction, since we have to make some choices.

The approach to reduce the number of scenarios is based on a distance between two scenarios denoted by $d(\xi^1, \xi^2)$, a quantity that we have to define. When the set of scenarios S' is defined we add the probability p_s for all $\xi^s \in S \backslash S'$ to the scenario $\xi^{s'} \in S'$ which has minimal distance to ξ^s.

The fast forward heuristic works as follows. It uses the fact that it is quite easy to find the scenario $\xi^{s'} \in S$ for which the total distance to all $\xi^s \in S \backslash \xi^{s'}$, which is

$$\sum_{\xi^s \in S \backslash \xi^{s'}} p_s d(\xi^s, \xi^{s'}), \tag{3}$$

is minimal. As $p_s = 1/S$ for all scenarios it can be replaced by a combination of the other scenarios. Iterating this until the set S' includes the predefined number of scenarios is the idea of the fast forward heuristic.

Given the generated scenarios $s \in S = \{\xi^1, \ldots, \xi^S\}$, the distances d between the scenarios, and the cardinality of S', $|S'| = k$ the fast forward selection works as follows:

begin
  $S^0 = \{1, \ldots, S\}$
  $\bar{d} = d$
  for $i = 1, \ldots, k$ **do**
    $s'_i \in \operatorname{argmin}_{s \in S^{i-1}} \left\{ \sum_{j \in S^{i-1} \backslash s} \min_{i \notin S^{i-1} \backslash s} \left\{ \bar{d}(\xi^i, \xi^j) \right\} \right\}$
    $S^i = S^{i-1} \backslash s'_i$

$update(\bar{d}, s'_i)$
$S' = S^0 \backslash S^k$
for $s' \in S'$ **do**
$p'_{s'} = \frac{1}{S} + \sum_{s \in S^k | s' = \text{argmin}_{\{\bar{s} \in S'\}} d(\xi^{\bar{s}}, \xi^s)} \frac{1}{S}$
return S' and p'
end

where $update(\bar{d}, s'_i)$ is the following function:

begin
for $i = 1, \ldots, |S|$ **do**
for $j = 1, \ldots, |S|$ **do**
$\bar{d}(\xi^i, \xi^j) = \min \left\{ \bar{d}(\xi^i, \xi^j), \bar{d}(\xi^i, \xi^{s'_i}) \right\}$
end

In our computational tests we use two different kinds of distances between two scenarios. The first distance we will refer to as *symmetric distance*. For the lead time we just take the euclidean distance

$$d(L_i^1, L_i^2) = |L_i^1 - L_i^2|. \tag{4}$$

Since a demand scenario consists of different demand rates for every time interval, we have to compare piecewise linear functions. We define the distance between two demand scenarios Ψ_i^1 and Ψ_i^2 as

$$d(\Psi_i^1, \Psi_i^2) = \left| \frac{\alpha_i^{1,r} - \alpha_i^{2,r}}{2^r} \right|, \tag{5}$$

where $\alpha_i^{s,r}$ denotes the demand rate during the time interval r at s.

There is another option that leads to asymmetric distances. The idea is to anticipate that the approximation is constructed for the use in a stochastic optimization problem. Thus, we would like to find the approximation that yields the least change in the result of the optimization. To decide which scenario is more important for optimization, we need some information about the costs that occur in case of stockholding and in case of stockout. We have this information given as parameter h_i, costs for holding one piece in stock, and c_i costs for having a stockout of one piece.

This way, we can define the *asymmetric distance between two lead time scenarios* as

$$d(L_i^1, L_i^2) = |L_i^1 - L_i^2| \frac{c_i}{h_i} \tag{6}$$

if $L_i^1 > L_i^2$

$$d(L_i^1, L_i^2) = |L_i^1 - L_i^2| \frac{h_i}{c_i}, \tag{7}$$

otherwise.

The distance $d(\xi^1, \xi^2)$ describes the costs of deleting scenario ξ^1 and adding its probability p_1 to p_2, the probability of ξ^2.

The definition of asymmetric distance between two demand scenarios is based on the same idea but the order $\Psi_i^1(x_i) > \Psi_i^2(x_i)$ depends on the x_i as $\Psi_i^s(x_i)$ is piecewise linear. That is why we look at the values of $\Psi_i^s(x_i)$ where x_i equals the expected value of the delivery time to location i, L_i. So we define the following *asymmetric distance between demand scenarios*:

$$d(\Psi_i^1,\, \Psi_i^2) = \left| \frac{\alpha_i^{1,r} - \alpha_i^{2,r}}{2^r} \right| \frac{c_i}{h_i}, \tag{8}$$

if $\Psi_i^1(L_i) > \Psi_i^2(L_i)$

$$d(\Psi_i^2,\, \Psi_i^1) = \left| \frac{\alpha_i^{1,r} - \alpha_i^{2,r}}{2^r} \right| \frac{h_i}{c_i}, \tag{9}$$

otherwise.

We use these distances in the fast forward selection to determine the scenarios $s' \in S'$ and their new probabilities $p'_{s'}$.

The asymmetric reduction does not approximate the distribution itself as faithfully as the reduction technique based on symmetric distances. We get a bias in our approximation that depends on the fraction of h_i and c_i. It will be shown in the next section that this biased reduction indeed approximates better the solutions to the optimization problems because it takes into account the cost of ending up in a certain scenario. To the best of our knowledge, this is not yet standard in the Stochastic Programming literature.

4. Simulation

We performed comprehensive computational tests on real-world data from our partner.

4.1. General Issues

Before we report on our tests, we want to make some general remarks concerning some side-effects of modeling-decisions of the SGSM.

First, the SGSM can only take finite discrete distributions of demands and lead times. Second, all scenarios of the demand distributions must be represented by piecewise linear approximations in order to obtain an MILP formulation for the SGSM.

Our partner forecasts the demand for one month. The data include the expected total demand in the actual month, the expected total demand in the coming month and so on. Thus, a straight-forward approach would be to approximate the demand linearly during one month. However: If we simply assume linearity of the demand during one month, then the rough discretization of time into months leads to demand scenarios with too little variation over time.

We can, of course, choose a finer discretization of time in weeks or days. The finer the discretization is the more realistic becomes the demand function.

In order to get a feeling for this influence, we generated stochastic numbers denoting the demand over one month or one week. Figure 1 shows an example of differences in the scenarios for discretization in months and in weeks.

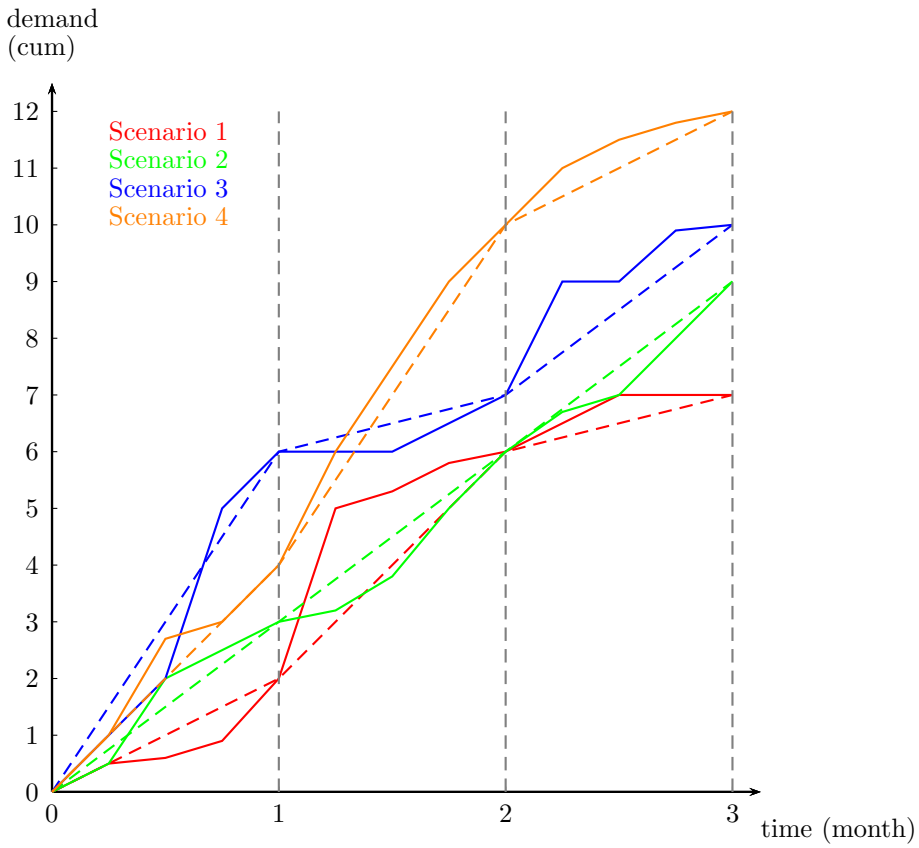

Figure 1: Different demand scenarios with discretization of time in month (dashed lines) and in weeks (solid lines)

A problem arises if the discretization of time becomes too small. The shorter the linear pieces in the demand functions, the more variables and constraints in the resulting MILP. This is the reason why the results in section 4.3 are all based on discretization in months or weeks. The discretization in days also does not lead to high savings compared to the one in weeks.

Besides the time discretization, the number of scenarios included in the model is the other quantity that is critical for the mere size and therefore to the computing time of the SGSM. Therefore, we check the effectivity of SAA with scenario reduction in our tests.

4.2. Test Data

We checked the SGSM on real data (inventory costs and demand intensities for 1127 spare parts) from our partner, a large German car manufacturer with a star-shaped two-echelon spare part distribution with one master warehouse (no. 0) and seven warehouses (nos. 1–7) for end customer service providers in the US. The model SGSM is not restricted to this special structure; it can be applied to any acyclic network structure.

The time horizon was chosen to be 25 months. The demand in the leaves of the network was generated randomly according to Poisson distributions with the given intensities from our historical data. Deviations of the delivery times up to 20 % were randomly generated. Stochastic data was identically reproduced for all policies under consideration. Replenishment orders in the simulation are triggered by (s, S)-policies, where the values for s are chosen by the models under consideration.

Note: The expected service times in the simulation are always equal to the service times computed in the respective models, i.e., on average there are no early deliveries (this is debatable; other options are work in progress). Moreover, the inventory costs in both the GSM and the SGSM are only approximations of the actual inventory costs. The simulation reports the actual (linear) inventory costs.

The SGSM produced scenarios by sampling from the Poisson distribution and was solved by Sample Average Approximation (SAA) (see Section 3.1. We tried a varying number of samples and scenario reduction as described in Section 3.2. The network topology was easy enough for all instances to solve in less than an hour for the testassortment of 1127 parts in the MILP solver `gurobi 3.0` up to an optimality gap of five percent. The calculation was carried out on a standard PC (CPU: *Intel(R) Core(TM) 2 Quad CPU Q9559 @ 2.83 GHz*, Mem: *8GB RAM*) using ubuntu 4.4.3.

In order to find out whether stochastic modelling as such has a positive impact on the result, we tried different parameter settings in the simulation experiments. The GSM is parametrized by the prescribed service level: we investigated the GSM with 90 % and with 96 % service level – called GSM(90%) and GSM(96%).

Moreover, in order to substantiate the benefit of a network model as opposed to a decentralized optimization of each separate warehouse, we give results for the decentralized policies DEZ(90%) and DEZ(96%) for a service level of 90% and 96%, respectively. In these models each location tries to reach the given service level target.

The cost coefficients are taken from cost estimates of our partner for inventory cost and the piece-based recourse cost (so-called "non-sales"). These coefficients are part and warehouse dependent and cannot be listed here.

4.3. Computational Results

Table 1 shows the benefits of sampling fifty scenarios followed by a reduction to three compared to sampling three scenarios. The results presented in this

table are the average costs of ten calculations with the given number of scenarios generated. The demands and delivery times are identical in all the simulations.

Table 1: results applying the SGSM with different scenario reduction techniques

Reduction	Inventory Cost	Recourse Cost	Total Costs
no $3 \to 3$	1 276 701.99	20 637 929.11	21 914 631.10
symmetric $50 \to 3$	1 368 387.73	5 799 112.62	7 167 500.35
asymmetric $50 \to 3$	1 487 010.23	1 876 708.99	3 363 719.22
no $50 \to 50$	1 659 602.59	1 528 288.18	3 187 890.77

We can see an enormous reduction in the total costs by applying the reduction techniques introduced in section 3. In the case of generating only three scenarios we observe a very high variability in the costs over the ten simulation runs. During ten simulations, the minimal total costs were 16 378 814.53, and the maximal total costs were 33 545 977.27. Applying the symmetric/asymmetric reduction technique the minimal total costs were 6 621 614.29/3 246 031.18 and the maximal total costs were 7 606 441.40/3 546 310.91, respectively. The costs occurring in the single simulation runs are listed in Appendix A.

These results show that applying scenario reduction leads to a much lower variability in the costs because also scenarios with small probability are taken into account.

We can see that the results for the asymmetric reduction are quite close to those where all the fifty generated scenarios are included in the model.

Table 2 includes the service levels in the different locations during the first of the ten simulation runs.

Table 2: Comparison of service levels (%)

Warehouse	$3 \to 3$	$50 \to 3$ sym	$50 \to 3$ asym	$50 \to 50$
0	75.4	85.4	73.1	88.9
1	92.4	94.4	95.6	96.8
2	92.5	94.1	95.3	96.3
3	92.1	94.2	95.2	97.0
4	92.0	94.1	95.1	96.2
5	93.0	94.8	96.6	97.5
6	92.0	94.0	96.1	96.3
7	92.9	95.0	95.9	97.0

The service levels in table 2 show the difference between the symmetric and the (new) asymmetric reduction technique. The asymmetric technique takes into account that for many parts the quotient h_i/c_i is greater for the leaf warehouses than for the master warehouse. Therefore, for the symmetric technique we get a higher service level in the master warehouse (no. 0), but lower service levels in the warehouses (nos. 1–7).

Simulating the situation modeled in the SGSM with simple recourse leads to the results listed in table 3.

Table 3: Results of simulation with poisson distributed demand and equal distributed delivery time

Model	Inventory Cost	Recourse Cost	Total Cost
(1) **DEZ 90%**	2 512 304.91	2 012 278.56	4 524 583.47
(2) **DEZ 96%**	2 987 207.93	1 018 603.98	4 005 811.91
(3) **GSM 90%**	2 496 925.63	1 831 689.48	4 328 615.11
(4) **GSM 96%**	2 983 078.69	963 058.22	3 946 136.91
(5) **SGSM** 50, months	1 555 212, 00	1 473 793.87	3 029 005.87
(6) **SGSM** 200 $\rightarrow$ 50, months, sym	1 560 619.50	1 497 695, 96	3 058 315, 46
(7) **SGSM** 200 $\rightarrow$ 50, months, asym	1 689 607.49	1 282 358.06	2 971 965 55
(8) **SGSM** 200 $\rightarrow$ 1, months	1 465 783.28	1 409 768.67	2 875 551.95
(9) **SGSM** 200 $\rightarrow$ 50, weeks, sym	1 867 149.91	893 382.95	2 760 532.86
(10) **SGSM** 200 $\rightarrow$ 50, weeks, asym	1 883 937.34	808 081.77	2 692 019.11

This table includes the average costs of the different approaches. Here we calculated the orderpoints s using all the different methods and run the simulation ten times with different demand and delivery time. For all different approaches the demand and delivery time in the simulation is identical.

The results for the decentralized method are a bit worse than the results, when the orderpoints are calculated by the GSM. Using one of the listed SGSM approaches leads to a cost reduction of 30% and more. Again, the asymmetric scenario reduction dominates the symmetric one. Another important aspect to notice is that the results using a discretization of time in weeks are remarkably better than results using a discretization in month. Results for each of the ten simulation runs for Model (4) and (10) can be found in Appendix A.

In Method (8) a special heuristic is applied (different from the fast forward reduction) that tries to find a critical scenario of the delivery time and the demand for every location. This shows that much of the problem's structure can be encoded into a single scenario. This heuristic works properly for the discretization in months and may be extended to finer discretization. This is work in progress.

The resulting service levels for the different methods in the first simulations are shown in table 4.

Table 4: Comparison of service levels (%)

Warehouse	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
0	94.1	96.0	94.2	96.3	78.9	89.7	90.0	71.3	92.4	89.8
1	97.6	98.5	97.7	98.4	96.7	96.8	96.9	96.8	97.2	97.3
2	97.1	98.0	97.1	97.9	96.4	96.6	96.7	96.6	96.9	96.9
3	97.4	98.3	97.5	98.3	96.7	97.0	97.0	97.0	97.6	97.8
4	97.4	98.1	97.4	98.1	96.5	96.5	96.6	96.4	96.8	96.8
5	98.4	99.1	98.4	99.1	97.4	97.5	97.5	97.2	97.8	97.7
6	97.1	98.0	97.1	98.1	96.5	96.7	96.7	96.1	96.7	96.7
7	97.4	98.3	97.4	98.2	97.2	97.3	97.5	97.5	97.6	97.7

The differences in the service levels of the symmetric and the asymmetric reduction are no longer substantial. The reason is that now the number of scenarios in the set S' is much higher; thus, both approaches lead to a good approximation of the distribution and its impact on resulting service levels.

As table 3 shows, the differences in the resulting costs are still remarkable. This is due to more scenarios in the more relevant parts of the distribution in the asymmetric reduction (high demand and delivery time if h_i/c_i is low and vice versa).

The results of simulations of the SGSM with external suppliers from which missing parts can be ordered and lost sales (introduced in subsection 2.3) are listed in Table 5:

Table 5: Results of simulation with external suppliers

Method	Inventory costs	Recourse Costs	Total Costs
(1) **DEZ 90%**	2 294 924.33	1 314 070.63	3 608 994.96
(2) **DEZ 96%**	2 471 509.73	1 130 821.16	3 602 330.89
(3) **GSM 90%**	2 268 247.71	1 311 477.48	3 579 725.19
(4) **GSM 96%**	2 451 235.28	1 121 440.97	3 572 676.25
(5) **SGSM 100**, weeks	2 294 965.77	792 949.54	3 087 915.31
(6) **SGSM** 200 $\rightarrow$ 50, weeks, sym	2 271 811.47	867 706.94	3 139 518.41
(7) **SGSM** 200 $\rightarrow$ 50, weeks, asym	2 230 176.18	689 392 01	2 919 568.19
(8) **SGSM** 300 $\rightarrow$ 75, weeks, sym	2 384 222.98	859 123.94	3 243 346.92
(9) **SGSM** 300 $\rightarrow$ 75, weeks, asym	2 230 359.11	607 771.65	2 838 130.76

The simulation works a little bit different to the one applied in Tables 1–4. Here the demand that can not be delivered immediately from the warehouses (nos. 1–7) to the end customers is lost. If the warehouses have not enough stock to deliver the ordered parts, there is the possibility to buy these parts from an external supplier. This recourse action causes costs depending on the distance between the warehouse and the external supplier. The supplier itself has limited stock so that the warehouses are not able to order any amount from them. If a demand at a warehouse can be neither delivered from stock nor ordered from an external supplier, the demand is lost.

Internal orders (from a warehouse to the master warehouse) are still backlogged, and the master warehouse delivers the demand as soon as possible to the ordering warehouse.

The ordering costs and the capacities of the external suppliers are not included in the data of our partner, so we had to set them artificially.

As we can see in the results of Table 5, the decentralized model and the GSM perform much better in the case with only one kind of uncertainty (demand uncertainty) than in the case of both, demand and delivery time uncertainty. The SGSM still outperforms the deterministic models achieving 10–20% of cost savings.

Table 6 show the resulting service levels of the different methods.

Here the service levels of the SGSM approaches are very similar to these of the decentralized model and the GSM, both with a prescribed service level of

Table 6: Comparison of service levels (%)

Warehouse	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0	84.9	87.2	85.0	87.6	88.2	88.2	88.0	88.7	88.8
1	94.2	94.3	94.2	94.3	94.5	94.5	94.6	94.6	94.6
2	94.1	94.1	94.0	94.1	94.3	94.3	94.3	94.3	94.4
3	94.4	94.5	94.4	94.5	94.7	94.7	94.7	94.7	94.7
4	93.9	94.0	93.9	94.0	94.1	94.1	94.1	94.2	94.2
5	94.3	94.5	94.3	94.5	94.7	94.8	94.8	94.8	94.8
6	93.6	94.0	93.6	93.9	94.0	94.0	94.0	94.0	94.0
7	94.1	94.2	94.1	94.3	94.4	94.5	94.4	94.5	94.5

96%. The costs in table 5 tell us that the SGSM treats different parts differently, while the GSM and the decentralized model cover 96% of the demand for every part, no matter what the costs h_i and c_{ji} are. This is the reason why the orderpoints calculated by the SGSM can lead to cheaper inventory costs and recourse costs at the same time.

Last we want to compare our model to a model that was introduced by Doğru, de Kok and van Houtum, see [11]. In the following we will refer to this model as DoKoHo. In the simulation we need to apply fix lead times as this is one assumption of the DoKoHo model. We simulate a situation that fits to the DoKoHo assumptions where demand is backlocked and there are penalty costs if a location is not able to deliver as demanded.

Table 7 shows the results of the simulation for the GSM, the SGSM and DoKoHo. There are some different parameter settings for DoKoHo where the penalty costs used in the model are multiplied by a factor (γ).

Table 7: Results of simulation with poisson distributed demand and fix lead time

Model	**Inventory Cost**	**Recourse Cost**	**Total Cost**
(1) **DoKoHo** ($\gamma = 1$)	1 450 564.94	2 510 522.12	3 961 087.06
(2) **DoKoHo** ($\gamma = 5$)	1 816 753.33	1 534 991.82	3 351 745.15
(3) **DoKoHo** ($\gamma = 10$)	1 954 727.95	1 387 486.39	3 342 214.34
(4) **GSM 96%**	1 834 581.46	1 980 314.45	3 814 895.91
(5) **SGSM** $300 \rightarrow 75$, weeks, asym	1 058 309.50	1 629 832.65	2 688 142.15

The DoKoHo model outperforms the GSM but causes higher costs than the SGSM. The simulation of the different models lead to the service levels that are shown in table 8.

The reason for the very low service levels at the master warehouse using the DoKoHo model compared to the ones using GSM or SGSM can be explained easily. In the DoKoHo model there are no explicit service times guaranteed to the successors. But the lower performance of the master warehouse is considered when the successor‘s safety stock is calculated. In the simulation we use a service time of zero for this case so the delivery of master warehouse is often late.

Table 8: Comparison of service levels (%)

Warehouse	(1)	(2)	(3)	(4)	(5)
0	48.0	53.7	55.4	91.9	83.3
1	96.8	98.7	99.0	98.1	97.4
2	96.4	98.4	98.6	97.8	97.1
3	96.6	98.7	98.9	97.9	97.4
4	96.3	98.2	98.4	97.6	96.5
5	97.0	98.9	99.1	98.7	97.6
6	96.7	98.1	98.2	97.8	96.6
7	96.6	98.5	98.7	97.9	97.5

As we do not consider penalty costs for master warehouse in the simulation, this does not effect the costs that we get applying the DoKoHo models.

5. Conclusion

We have introduced the Stochastic Guaranteed Service Model (SGSM), a stochastic programming version of the Guaranteed Service Model (GSM) for the computation of safety stock levels in a multi-echelon spare part distribution system of a large German car manufacturer.

Whereas the GSM makes assumptions that require extreme demand scenarios and missed delivery dates to be handled outside the model, the SGSM is capable of incorporating these volatilities inside the model, thereby accounting for the corresponding cost. The stochasticity needs to be captured by sufficiently large sample sizes: in our example we generated 200 scenarios most of the time and reduced them to 50 applying modified scenario reduction techniques. The resulting MILP models could be solved straight-forwardly in our example.

The SGSM makes some assumptions that are only approximations of reality (complete recourse, piecewise linear demand). However, our simulation was not restricted by these assumptions; it only checked the resulting policies, no matter what they assumed, and accounted for all the occurring costs. And in this quite realistic simulation experiment, the policies calculated with the SGSM performed extremely well. One reason for this is that the SGSM can have *structurally* different optimal solutions than the GSM: not all optimal SGSM solutions are extreme in the space of variables of the GSM. Thus the SGSM sometimes finds solutions that the GSM can never provide, no matter which parameter setting. And such solutions dominated GSM solutions in our simulations.

We therefore think that the SGSM can be applied routinely in spare part distribution systems like the one of our partner. Next, we will model the real-world recourse actions in more detail in order to find more realistic recourse cost values.

Appendix A. Results of the simulation runs

In this section we show some of the numerical results in detail. The average costs of the ten simulation runs listed here are given in tables 1 and 3.

The tables A.9–A.12 include the results that lead to the average costs of table 2.

Table A.9: No reduction ($3 \to 3$)

Run	Inventory Costs	Recourse Costs	Total Costs
1	1 262 180.67	20 904 545.37	22 166 726.04
2	1 227 922.08	17 572 609.36	18 800 531.44
3	1 238 190.82	22 341 954.25	23 580 145.07
4	1 294 183.52	19 942 497.73	21 236 681.25
5	1 310 570.53	21 047 322.98	22 357 893.51
6	1 256 992.73	20 860 600.50	22 117 593.23
7	1 325 988.56	15 528 825.97	16 378 814.53
8	1 262 180.67	20 904 545.37	22 166 726.04
9	1 241 824.84	32 304 152.43	33 545 977.27
10	1 311 050.20	18 169 764.58	19 480 814.58
Average	**1 273 108.46**	**20 957 681.85**	**22 183 190.31**

Table A.10: Symmetric reduction ($50 \to 3$)

Run	Inventory Costs	Recourse Costs	Total Costs
1	1 344 456.16	5 885 899.06	7 230 355.22
2	1 374 833.68	6 231 607.72	7 606 441.40
3	1 357 439.20	5 264 175.09	6 621 614.29
4	1 358 487.22	5 693 121.60	7 051 608.82
5	1 377 759.11	6 037 579.32	7 415 338.43
6	1 356 518.24	5 469 194.65	6 825 712.89
7	1 376 288.85	5 947 061.26	7 323 350.11
8	1 401 105.86	5 516 129.63	6 917 235.49
9	1 358 232.81	5 828 321.03	7 186 553.84
10	1 378 756.13	6 118 036.79	7 496 792.92
Average	**1 368 387.73**	**5 799 112.62**	**7 167 500.35**

Table A.13 and A.14 include the results for the single runs of **GSM 96%** (4) and **SGSM** $200 \to 50$, weeks, asym (10) of table 3.

References

[1] Simpson, In-process inventory, Operations Research 6 (1958) 863–873.

[2] S. Graves, S. Willems, Optimizing strategic safety stock placement in supply chains, Manufacturing & Service Operations Management 2 (1) (2000) 68–83.

Table A.11: Asymmetric reduction (50 → 3)

Run	Inventory Costs	Recourse Costs	Total Costs
1	1 485 860.93	1 980 706.25	3 466 567.18
2	1 478 170.53	1 767 860.65	3 246 031.18
3	1 476 507.35	1 833 396.79	3 309 904.14
4	1 508 549.33	1 743 703.09	3 252 252.42
5	1 509 959.07	1 860 989.82	3 370 580.89
6	1 475 040.72	1 865 894.95	3 340 935.67
7	1 502 305.31	1 888 979.81	3 391 285.12
8	1 495 706.29	1 910 136.29	3 405 842.58
9	1 480 737.28	1 826 744.86	3 307 482.14
10	1 457 633.50	2 088 677.41	3 546 310.91
Average	**1 487 047.03**	**1 876 708.99**	**3 363 719.22**

Table A.12: No Reduction (50 → 50)

Run	Inventory Costs	Recourse Costs	Total Costs
1	1 650 596.38	1 531 849.61	3 182 445.99
2	1 657 496.08	1 391 982.95	3 049 479.03
3	1 655 262.68	1 520 144.48	3 175 407.16
4	1 652 312.18	1 512 184.03	3 164 496.21
5	1 660 352.51	1 511 081.84	3 171 434.35
6	1 657 754.36	1 577 041.20	3 234 795.56
7	1 663 863.19	1 561 981.43	3 225 844.62
8	1 673 298.06	1 509 069.42	3 182 367.48
9	1 669 855.12	1 525 359.34	3 195 214.46
10	1 655 235.31	1 642 187.46	3 297 422.77
Average	**1 659 602.59**	**1 528 288.18**	**3 187 890.77**

[3] T. Magnanti, Z.-J. Shen, J. Shu, D. Simchi-Levi, C.-P. Teo, Inventory placement in acyclic supply chain networks, Operations Research Letters 34 (2006) 228–238.

[4] K. Schade, Lagerhaltungsstrategie für mehrstufige Lagerhaltung in der Automobilindustrie, Master's thesis, Universität Bayreuth (2008).

[5] K. Inderfurth, Safety stock optimization in multi-stage inventory systems, International Journal of Production Economics 24 (1-2) (1991) 103 – 113. doi:10.1016/0925-5273(91)90157-O.
URL http://www.sciencedirect.com/science/article/pii/0925527391901570

[6] K. Inderfurth, Safety stocks in multistage divergent inventory systems: A survey, international journal of production economics 35 (1994) 321–329.

[7] S. Minner, Dynamic programming algorithms for multi-stage safety stock optimization, OR Spectrum 19 (1997) 261–271, 10.1007/BF01539783.
URL http://dx.doi.org/10.1007/BF01539783

Table A.13: GSM with a prescribed service level of 96%

Run	Inventory Costs	Recourse Costs	Total Costs
1	2 977 194.04	953 576.93	3 930 770.97
2	2 985 082.45	957 196.39	3 942 278.84
3	2 981 576.40	1 017 176.75	3 998 753.15
4	2 988 402.12	945 210.41	3 933 612.53
5	2 983 057.11	1 085 943.06	4 069 000.17
6	2 992 958.19	914 166.88	3 907 125.07
7	2 997 632.67	881 167.01	3 878 799.68
8	2 974 985.18	962 779.43	3 937 764.61
9	2 971 511.93	927 915.91	3 899 427.84
10	2 978 386.83	985 449.47	3 963 836.30
Average	**2 983 078.69**	**963 058.22**	**3 946 136.91**

Table A.14: SGSM $200 \rightarrow 3$ asymmetric reduction with time discretization in weeks

Run	Inventory Costs	Recourse Costs	Total Costs
1	1 869 685.34	794 347.91	2 664 033.25
2	1 894 667.10	742 510.22	2 637 177.32
3	1 892 332.95	834 364.91	2 726 697.86
4	1 874 564.45	764 024.10	2 638 588.55
5	1 875 607.36	997 203.15	2 872 810.51
6	1 883 575.61	772 162.01	2 655 737.62
7	1 879 543.93	812 130.95	2 691 674.88
8	1 894 412.96	808 671.05	2 703 084.01
9	1 885 131.65	767 510.04	2 652 641.69
10	1 889 852.08	787 893.38	2 677 745.46
Average	**1 883 937.34**	**808 081.77**	**2 692 019.11**

[8] A. Clark, H. Scarf, Optimal policies for a multi-echelon inventory problem, Management Science 6 (1960) 475–490.

[9] C. Sherbrooke, Metric: A multi-echelon technique for recoverable item control, Operations Research 16 (1968) 122–141.

[10] A. Diaz, M. C. Fu, Multi-echelon models for repairable items: A review, Document in Decision, Operations & Information Technologies Research Works http://hdl.handle.net/1903/2300, University of Maryland (2005). URL `http://hdl.handle.net/1903/2300`

[11] M. Doğru, A. de Kok, G. van Houtum, Optimal control of one-warehouse multi-retailer systems with discrete demand, Working paper (2005).

[12] J. R. Birge, F. Louveaux, Introduction to Stochastic Programming, Springer, 1997.

[13] J. Rambau, K. Schade, The stochastic guaranteed service model with recourse for multi-echelon warehouse management, in: Proceedings of the

International Symposium on Combinatorial Optimization (ISCO 2010), Vol. 36 of Electronic Notes in Discrete Mathematics, Elsevier, 2010, pp. 783–790, to appear.

[14] A. Shapiro, Monte carlo sampling methods, in: A. Ruszczynski, A. Shapiro (Eds.), Stochastic Programming, Vol. 10 of Handbooks in Operations Research and Management Science, Elsevier, 2003, pp. 353 – 425. doi:DOI: 10.1016/S0927-0507(03)10006-0.

[15] H. Heitsch, Stabilität und approximation stochastischer optimierungsprobleme, PhD dissertation, Humbold-Universität zu Berlin (Nov. 2007).

[16] H. Heitsch, W. Römisch, Scenario reduction algorithms in stochastic programming, Computational Optimization and Applications 24 (2003) 187–206.

[17] R. Henrion, C. Küchler, W. Römisch, Discrepancy distances and scenario reduction in two-stage stochastic mixed-integer programming, JOURNAL OF INDUSTRIAL AND MANAGEMENT OPTIMIZATION 4 (2) (2008) 363–384.