Extrapolating from Limited Uncertain Information to obtain Robust Solutions for Large-Scale Optimization Problems

Laura Climent*, Richard Wallace*, Barry O'Sullivan* and Eugene Freuder*
*Insight Centre for Data Analytics. University College Cork, Cork, Ireland
Email: {laura.climent,richard.wallace,barry.osullivan,eugene.freuder}@insight-centre.org

***Abstract*—Data uncertainty in real-life problems is a current challenge in many areas, including Operations Research (OR) and Constraint Programming (CP). This is especially true given the continual and accelerating increase in the amount of data associated with real-life problems, to which Large Scale Combinatorial Optimization (LSCO) techniques may be applied. Although data uncertainty has been studied extensively in the literature, many approaches do not take into account the partial or complete lack of information about uncertainty in real-life settings. To meet this challenge, in this paper we present a strategy for extrapolating data from limited uncertain information to ensure a certain level of *robustness* in the solutions obtained. Our approach is motivated by real-world applications of supply of timber from forests to saw-mills.**

***Keywords*-uncertainty; robustness; optimization;**

I. Introduction

Data uncertainty can be due to (i) the dynamic and unpredictable nature of the real world (e.g., future events and changes, stochastic demand/request, etc.), but also to (ii) the level of available information for modeling the problem and its correctness (e.g., partial or inadequate information, measurement errors, etc.) [1]. In this paper, we deal with real-life problems in which uncertainty is associated with elements that have an ordered relationship. These problems can be affected by either type of uncertainty. In these problems, measurement/estimation errors are common, and they result in a partially incorrect representation of the modeled problem. The same occurs when the original measurements/estimations were taken properly, but because of external factors (for instance, the environment), the estimated values become partially/completely incorrect.

In real-life problems, repeated observation of an uncertain and/or dynamic environment allows the calculation of statistics related to the uncertain data in the problem (frequently, these statistics can be expressed in the form of probabilities). Nevertheless, in most cases real-life problems do not have much detailed probabilistic data associated with them. This often happens in LSCO problems because of the large sizes of the domains and the difficulty of gathering probabilities for individual values in the domain (discrete domains) or probability distributions (continuous domains).

In this paper we introduce an approach for handling these situations, by extrapolating data about changes over the original formulation of the uncertain parameters, based on their order. After such extrapolation, classic probabilistic models can be used to solve the problem. As a real-life application example, we design a linear optimization model combined with chance constraints [2] for the forestry problem of timber supply. For this extrapolation, we have been inspired by some ideas in [3], [4]: when incorrect measurements/estimations involve domains with a significant order, the dynamism is often equivalent to relaxations/restrictive modifications over the related domains and constraints.

Such restrictions reduce the solution space (contrary to the relaxations) and therefore, a solution obtained for the problem with the erroneous data might not be a solution to the problem with the correct data. In order to reduce the chances of losing a solution, it is important to obtain a robust solution, which has a high likelihood to remain valid, faced with the uncertainties of the problem. And this is indeed the motivation of the extrapolation technique presented in this paper. In this way, the linear optimization model presented for the forestry problem, ensures minimum levels of robustness in the uncertain parameters of the problems based on the probability information that we extrapolate. In addition, this model can be combined with the maximization/minimization of other optimization criteria.

This paper is structured as follows: First, in Section II some definitions used in the paper are explained. Section III gives a brief review of related work, which helps to motivate our approach. In Section IV, we introduce an approach for extrapolating from limited uncertain information associated with some parameters of the problems. In Section V we present a linear optimization model for ensuring minimum robustness levels in the uncertain parameters of the forestry problem. Finally, we give some conclusions in Section VI.

II. Technical Background

In this section we provide some standard definitions for modeling problems in the CP and OR field.

Definition 2.1: A Constraint Satisfaction Problem (CSP) is represented as a triple $P = \langle \mathcal{X}, \mathcal{D}, \mathcal{C} \rangle$ where $\mathcal{X}$ is a finite set of variables $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, $\mathcal{D}$ is a set of domains $\mathcal{D} = \{D_1, D_2, ..., D_n\}$ such that for each variable $x_i \in \mathcal{X}$ there is a set of values D_i that the variable can take, and $\mathcal{C}$ is a finite set of constraints $\mathcal{C} = \{C_1, C_2, ..., C_m\}$ which restrict the values that the variables can simultaneously take.

Definition 2.2: A Linear Optimization (LO) problem can be expressed in the canonical form:

$$\begin{array}{ll} \text{maximize/minimize} & c^T x + d \\ \text{subject to} & Ax \leq b \\ & x \geq 0 \end{array}$$

where $x \in \mathbb{R}^n$ is a vector of decision variables and n is the number of variables of the problem, the vector of coefficients $c \in \mathbb{R}^n$ and constant value $d \in \mathbb{R}$ form the objective function, A is an $m \times n$ constraint matrix and $b \in \mathbb{R}^m$ is the vector of constant terms of the m constraints.

III. Approaches for dealing with uncertainty

Uncertainty associated with the parameters of models representing real-life problems has been treated extensively in the literature. An interesting discussion of OR approaches can be found in [5]. In addition, a summary of CP approaches for dealing with uncertainty can be found in [6]. In this section, we focus on two kinds of models which assume that some parameters of the problem can take any value in an range of values. There also exist other approaches that consider other ways of representing the uncertainty (for instance, as previously mentioned, in [3], [4] the constraints and domains may become tighter over the time). However, in this paper, we focus on this representation of the uncertainty of the variables because it allows to model uncertainty in the measurement/estimation of certain parameters (which is another representation than the used in [3], [4]). In Section III-A we discuss models that do not consider probabilistic data associated with the intervals. In Section III-B we discuss stochastic models.

A. Non-stochastic Uncertainty Sets

When we are not entirely sure about the precise value of some parameters, we can often extract information about the problem that indicates, with a certain confidence, intervals of values that these parameters can take. These intervals are typically called *uncertainty sets* ($\mathcal{U}$). A parameter associated with an uncertain set $U_i \in \mathcal{U}$ is called an uncontrollable variable (typically in CP approaches) or random variable (typically in OR approaches). In order to avoid confusion over this term, in the rest of the paper we refer to such variables as random variables and each one is denoted as $s_i \in S$. Just as with deterministic domains, the uncertain sets can be discrete or continuous. In this framework, a conservative definition of a robust solution was given [7].

Definition 3.1: A solution is robust if it satisfies all realizations of the constraints from the uncertainty set.

A CSP model that considers Definition 3.1 with discrete uncertain domains is the Mixed CSP model (MCSP) [8], whose main objective is to find an assignment of decision variables (which are the usual variables of the CSP model) that satisfy all the possible values that the uncontrollable variables can take. Another type of model, the Uncertain CSP model (UCSP) [1] is an extension of the MCSP because it considers continuous domains. However, Definition 3.1 is very conservative and demanding since there are often no solutions that satisfy this requirement for every random variable. In addition, in optimization problems, if there is a solution that satisfies such demands, the cost to "pay" in the criterion to optimize is often very high. This is one difficulty that motivated the development of a less conservative approach, as described below.

In [9], a Γ_i parameter is defined, which fixes the number of random variables that are allowed to change for the ith uncertain constraint. If a random variable is allowed to change, this means that the solution will remain a solution if the variable takes any value in its uncertainty set. Note that $\Gamma_i \in [0, |J_i|]$, where $J_i \subseteq S$ is the set of random variables of the ith uncertain constraint. We would like to point out that when $\Gamma_i = 0$, the obtained solution is non-robust (only the value estimated/measured for the random variables is considered). However, when $\Gamma_i = |J_i|$ the obtained solution is one of the most robust for the ith uncertain constraint (because all possible uncertain values of the random variables are considered).

There are disadvantages when we try to apply the latter approach to real problems in which most or all of the random variables are likely to change from the original value estimated. This is especially true for LSCO problems because they have a very high number of random variables. In such cases, it is not only important that some variables are able to satisfy all the values in the uncertain set, but it is more important to ensure that all the variables are able to satisfy some of the values in their uncertainty sets. Note that in the latter case the robustness is spread more evenly among all of the random variables. In addition, this technique does not take into consideration situations where the estimated values have certain probabilistic dynamism data associated.

B. Stochastic Programming

Another area that deals with optimization subject to uncertainty is stochastic programming. Recently, a version of stochastic programming has been developed in which constraints are treated within the framework of stochastic programming; consequently, it is known as stochastic constraint programming. In either approach, each uncertain decision variable is considered to be a random variable with an associated probability mass function, which indicates the probability that the variable will take any particular value in its uncertain domain $\mathcal{U}$. In using such models, it is assumed that probabilistic data can be derived from empirical observations or historical evidence.

In both stochastic programming and stochastic constraint programming, decision variables are organized into scenarios in which a given decision leads to one of a set of possible events, each of which is the occasion for another decision, and so forth. The different combinations of the

random variables produce different scenarios. The likelihood of occurrence of an scenario, when the random variables represent independents events is $\prod_{k=1}^{k=|S|} p(s_i = a_i)$, where p is a probability mass function associated with s_i. Furthermore, there are also approaches that use the chance constraints [2]. These constraints are composed of at least a random variable, which makes possible to state a minimum probability of satisfaction for such constraint. Thus, considering a LO model (see Definition 2.2) with some random variables, a chance constraint would be defined as: $p(Ax \leq b) \geq \beta$, where "p" is the abbreviation of "probability" and β is the set of prescribed confidence levels.

The main advantages of the mentioned stochastic approaches are that they are less conservative than Definition 3.1 and in addition they are able to spread the robustness across all of the random variables. However, a major weakness of all these stochastic approaches in the situations we envisage is the precise and extensive knowledge of probabilities that such models require. In fact, each value in a domain interval must be associated with a specific probability. However, in many applications information regarding uncertainty is likely to be incomplete, erroneous or even nonexistant. Since stochastic models require precise probability distributions, they cannot be applied if such information is unknown. For this reason, in this paper we propose an extrapolation approach that can be combined subsequently with an stochastic approach. As a result, we have the benefits that the stochastic approaches provide even when there is a lack of data about the probabilities of the uncertainty sets.

In Section V we present a combination of the extrapolation method presented and chance constraints for the real-life forestry problem. We selected the chance constraints instead of other approaches that generate all possible scenarios due to the cost of computing a large number of possible combinations (especially in large combinatorial problems). This disadvantage is exacerbated when the uncertain intervals are continuous; in such cases a discretization must be done in order to generate all scenarios.

IV. Extrapolating from Limited Uncertain Data

The limitations of the two types of approach just described have motivated the present approach. Its main requirement is that for each uncertain set, whether continuous or discrete, there is an order over the elements that compose it. In lieu of detailed information regarding probabilities, we will use some cumulative probability distribution over each ordered domain that has an associated monotonic relation. This section covers the explanation of the uncertain ordered intervals, the probabilities associated with them and the cumulative distributions.

A. Uncertain Ordered Intervals

Typically, after observing and measuring the variables associated with a given problem, a nominal value (denoted as $\widehat{u}_i$) can be associated with each random variable s_i; this represents an estimate that is subject to some partially known/unknown amount of uncertainty. In addition, following collection and analysis of error measurements carried out in similar real-life instances, an estimate can be made of the range of possible values that the random variable might take (denoted as $\widehat{e}_i$). Thus, the interval U_i of s_i can be denoted as: $[(\widehat{u}_i - \widehat{e}_i), (\widehat{u}_i + \widehat{e}_i)]$. Note that in this case $\widehat{u}_i$ is assumed to be at the midpoint of the interval, even if this is not the only possible case, for simplicity, in the following we consider the symmetric case. In the example to be described in Section V, the error estimate is a fixed percentage of the nominal value. This follows standard practice in this domain of application. In this case, the interval of the random variables can be denoted as: $[(\widehat{u}_i(1 - \widehat{e}_i)), (\widehat{u}_i(1 + \widehat{e}_i))]$, where $\widehat{e}_i \in [0, 1]$.

The present work follows some ideas presented in [3] and [4] on Dynamic CSPs. In that work, values that were greater/lower than a certain value could be more or less robust than this value depending on the amount of "restrictive" changes over the solution space that they could handle. For such limited assumption, there must be an ordered relationship over the domains. For example, a time-buffer after a scheduled task (which is achieved by selecting greater values as starting times for the following closest tasks) makes the schedule more robust because it can handle changes that are potentially more restrictive (due to, for instance a delay in some previous tasks) with respect to possible start times. Moreover, a longer buffer subsumes a shorter one, so that the probability that the longer buffer absorbs a restrictive change affecting such task, is necessarily greater than the one associated a shorter buffer.

As will be demonstrated shortly, in the situations under consideration in this paper, it is often the case that values to one side of an estimate represent cases that allow greater or lesser restrictive deviations in the same sense. This means that we can use the same kind of reasoning in these situations that was used in the non-probabilistic work cited above on dynamic changes that affect variables with ordered domains. Here, however, instead of alterations of varying likelihood we have true values of varying likelihood given an uncertain estimated value.

To illustrate, let $s_i \in S$ be a random variable with an ordered domain that is strictly increasing or decreasing $[u_1, u_2, u_3]$. If the value u_2 allows greater restrictive deviations than u_1 and u_3 allows greater restrictive deviations than u_2 (according to the order relationship given by the problem), then a solution that is feasible for $s_i \leq u_3$ is more robust than one that is only feasible for $s_i \leq u_1$. As an example, if the real value of this uncertain variable is u_2, the first solution remains a solution but the second does not. The most robust solution/s are the ones that are feasible for $s_i \leq u_3$ because they will remain solutions for any value that the variable s_i takes (it allows all the possible restrictive deviations). However, when the 'restrictive relation' among

values is the opposite (u_1 is the value that allows the greatest deviation), then the robustness order is the opposite as well. That is, solutions that are feasible for $s_i \geq u_1$ are the most robust ones.

B. Probabilities Associated with the Intervals

Regarding the probabilities associated with the ordered domain values of the random variables, and this is a variation from the classical stochastic original model (see Section III-B), the probability distribution defined over the uncertain ordered interval expresses the probability that the random variable s_i takes a value greater or equal ($p(s_i \geq u_i)$), or lower or equal ($p(s_i \leq u_i)$) than $u_i \in U_i$ (recall that U_i is the uncertainty set associated with s_i). This probability is directly related to the robustness of the solutions because, given a solution, the higher this probability, the higher the likelihood that the real value of the random variable is feasible with this solution. Now, we can go a step further and use the same logic to order domain values with respect to the amount of change in the actual value with respect to a putative estimate, which is the estimate we use as part of a solution. Thus, In this case the relation $p(s_i \geq u_1) > p(s_i \geq u_2) > p(s_i \geq u_3)$ implies that a solution with value u_1 as an estimate for the value of s_i is more likely to remain a solution than if we chose a value u_2, etc.

These statements are illustrated by the example in Table I. In this example, we consider three integer values $[1, 2, 3]$ that represent either demands or capacities in the modeled problem. The arrows over the domains represent their restrictiveness order. For instance, when they represent demands, the greater values allow greater deviations (for a certain capacity of a product, higher demands restrict the problem more). For this reason, the probability is higher when the random variables are estimated to be to the highest values. However, when they represent capacities, the lower values are more conservative, in that if a capacity value amply meets a certain demand, the solution is more likely to be feasible for lower capacities than the expected ones.

Table I
EXAMPLES OF ORDERED DOMAINS AND THE PROBABILITIES (p) ASSOCIATED WITH THEIR RANDOM VARIABLES.

DOMAIN	**PROBABILITIES** (p)
Demand: $\overrightarrow{[1, 2, 3]}$	$p(s_i \leq 1) < p(s_i \leq 2) < p(s_i \leq 3)$
Capacity: $\overleftarrow{[1, 2, 3]}$	$p(s_i \geq 1) > p(s_i \geq 2) > p(s_i \geq 3)$

C. Cumulative Probability Distributions

In this paper, we are assuming a lack of probabilistic information associated with the interval of uncertainty; therefore, we need a way of extrapolating it. For this purpose, we take the ordering of values over the uncertain interval into account. As mentioned earlier, the maximum variation over the value estimated/measured is $\widehat{e_i}$. Therefore, the probability that a random variable takes a value in the uncertain interval $[(\widehat{u}_i * (1 - \widehat{e_i})), (\widehat{u}_i * (1 + \widehat{e_i})]$ is one. Hence, the probability associated with the value capable of handling the largest deviation is one and the value least able to accommodate deviations has an associated probability close to zero. (It is not exactly zero since it is possible that the random variable takes this value). For the remaining values in the continuous or discretized uncertain interval we extrapolate their likelihood according to a cumulative distribution function in the interval $(0, 1]$.

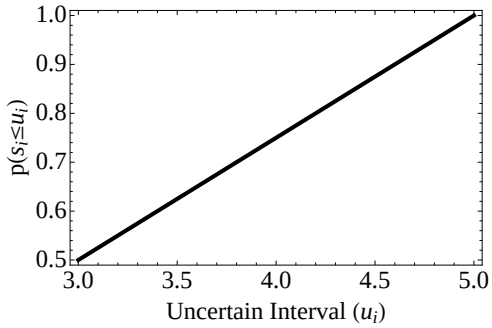

Figure 1. Increasing Uniform Cumulative Distribution.

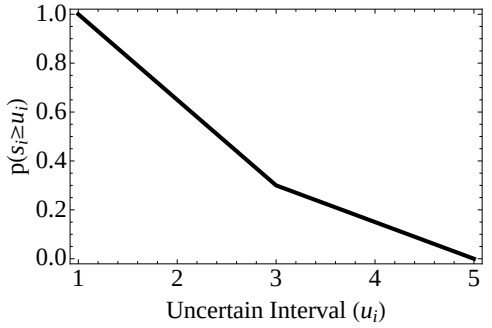

Figure 2. Decreasing Skewed Uniform Cumulative Distribution.

In this paper, we describe the extrapolation using cumulative probability distributions that come from either the uniform or the normal distributions. However, any cumulative distribution (with strictly monotonic increasing/decreasing probability) could be used instead. The uniform distribution represents a constant increase (see Figure 1) or decrease of the cumulative probability over the value estimated ($\widehat{u}_i$). However, for the normal distribution the increase or decrease over $\widehat{u}_i$ is more abrupt for the values that are closer to $\widehat{u}_i$ (see Figure 3 and Figure 4).

The selection of the cumulative distribution depends on what the random variable associated with the uncertain interval represents. In some cases, such as capacities and demands, the uniform cumulative distribution may be suffi-

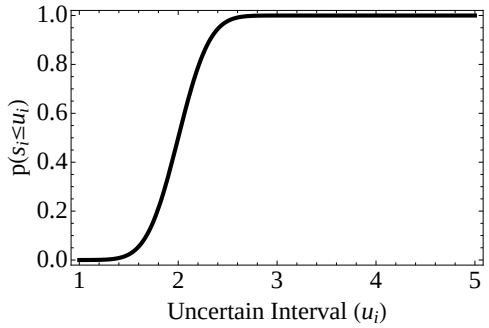

Figure 3. Shifted Increasing Normal Cumulative Distribution.

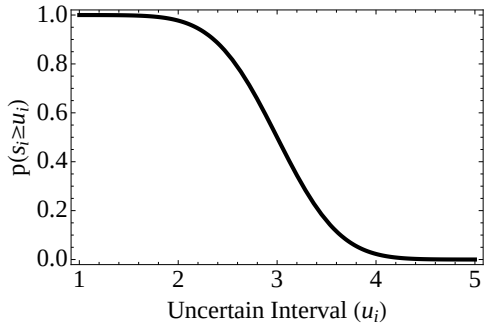

Figure 4. Decreasing Normal Cumulative Distribution.

cient, especially if the differences in value are more or less evenly distributed. However, for other domains such as life expectancy or breakdowns, the differences over the values are not evenly distributed because some extreme values are highly unusual given a particular estimated value. Here, the normal cumulative distribution would be more appropriate.

We can also accommodate situations in which the cumulative probability associated with the estimated value for a random variable (nominal value) is known. Such situations include cases where there are external environmental factors such as storms, floods, etc. that decrement the level of certainty of the estimated values. In these cases, the likelihood has to be fixed for the nominal value when we extrapolate the remaining information for the uncertain interval. For these situations, either the cumulative distributions are the same but the position of the nominal value is shifted (see Figure 3) or the distribution is skewed and there are two specific gradients for values greater and lower than the nominal value (see Figure 2). The extrapolation of the uniform skewed cumulative probability distribution is defined in Equation 1 for the increasing case (see Figure 1) and Equation 2 for the decreasing case (see Figure 2).

Increasing skewed uniform cumulative distribution:

$$p(s_i \leq u_i) = p(s_i \leq \widehat{u}_i) + p' * (u_i - \widehat{u}_i)$$

$$p' = \begin{cases} \frac{1-p(s_i \leq \widehat{u}_i)}{\widehat{e}_i * \widehat{u}_i} & \text{if } u_i \geq \widehat{u}_i \\ \\ \frac{p(s_i \leq \widehat{u}_i)-(\sim 0)}{\widehat{e}_i * \widehat{u}_i} & \text{if } u_i < \widehat{u}_i \end{cases} \quad (1)$$

Decreasing skewed uniform cumulative distribution:

$$p(s_i \geq u_i) = p(s_i \geq \widehat{u}_i) + p' * (\widehat{u}_i - u_i)$$

$$p' = \begin{cases} \frac{p(s_i \geq \widehat{u}_i)-(\sim 0)}{\widehat{e}_i * \widehat{u}_i} & \text{if } u_i > \widehat{u}_i \\ \\ \frac{1-p(s_i \geq \widehat{u}_i)}{\widehat{e}_i * \widehat{u}_i} & \text{if } u_i \leq \widehat{u}_i \end{cases} \quad (2)$$

For cases in which the likelihood of the nominal value is unknown, the value is assumed to be at the midpoint of the uncertain interval with cumulative probability $p(s_i \geq \widehat{u}_i)$ or $p(s_i \leq \widehat{u}_i) = 0.5$. Note that in this case, p' is the same for the two conditional cases of Equation 1 and Equation 2. Thus, the cumulative distributions have only one constant gradient over the entire uncertain interval. After extrapolating the probabilities associated with the random variables, stochastic constraint programming techniques from the literature can be used. In line with the motivation given in Section III, we introduce a linear approach with chance constraints in Section V.

V. A Case Study from Forestry

To illustrate these ideas in practice, we present an example in which we apply the concepts and approach developed in this paper to a well known real-life LSCO problem of forestry. This type of problem is of great importance for many countries that export timber obtained from forest harvesting. The impact of OR in the forestry industry is indisputable (see for instance, the Edelmann Prize [10]).

The process of harvesting stands from forests starts with the cutting of the trees that belong to the selected area for harvest. Subsequently, the trees are bucked into the different types of log-products. For such purposes, 3D forest measurements are made by means of laser scans that estimate the capacities of the forests. Nevertheless, as pointed out in Section I, there are many factors that contribute to the uncertainty of the estimation of such timber offer, for instance: (i) environmental factors, such as bad weather conditions, pests, etc. that change the expected growing of the trees and (ii) error measurements due to the bad quality of the laser scans.

The logistics process includes transportation of different types of timber to the mills; these represent the customers that fix the demand. There are different costs associated with the transportation process, and it is needless to mention that these transportation costs should be minimized in order to obtain the maximum profit from the sales of logs. For

this reason, we have developed a linear model for such problem in Section V-A. Then, in Section V-B for illustrative purposes we apply this model to a toy instance.

A. Linear Model for Uncertainty

In this section, we propose a Mixed Integer Programming (MIP) model, which is a type of LO model (see Definition 2.2) in which some variables are non-integers. The model presented includes the extrapolation technique proposed in this paper (see Section IV) and it also minimizes the transportation costs. We control the level of robustness of the random variables by incorporating a chance constraint to the model (see Section III-B) for each random variable s_i where $0 < i \leq |S|$. According to the order relationship of each uncertain interval U_i, the probability of the chance constraint is $p(s_i \geq u_i)$ or $p(s_i \leq u_i)$.The uncertain intervals U_i are computed given the input parameters $\widehat{u}_i$ and $\widehat{e}_i$ (see Section IV). Following, we present this MIP model.

1) Sets:

   Sets of forests $\mathcal{F}$ $(i \in \mathcal{F})$.
   Sets of mills $\mathcal{M}$ $(j \in \mathcal{M})$.
   Sets of stands $\mathcal{S}$ $(k \in \mathcal{S})$ for all the forests.
   Sets of types of logs $\mathcal{L}$ $(t \in \mathcal{L})$.

2) Parameters:

   g_{ik} : Cost of harvesting the stand k from forest i.
   c_{ijt} : Cost of supply associated to a_{ijt}.
   $\widehat{u}_{ikt}$: Capacity estimated for the forest i for each of its stands k for each type of log t.
   $\widehat{e}_{ikt}$: Percentage of variability from $\widehat{u}_{ikt}$.
   d_{jt} : Demand of the mill j of the type of log t.
   β_{ikt} : Vector of minimum probabilities.
   N : Large constant.

3) Variables:

   $b_{ik} \in \{0, 1\}$ (equal to 1 if we harvest the stand k from forest i, otherwise equal to 0).
   a_{ijt} = Amount of supply from forest i to mill j of the log type t.
   u_{ikt} = Minimum capacity selected for the forest i for the stand k for the type of log t.
   x_{ikt} = Vector of auxiliary variables (equal to u_{ikt} if b_{ik} is equal to 1, otherwise equal to 0).
   s_{ikt} = Vector of random variables associated with the uncertain capacities.

4) Mathematical Model:

$$
\begin{array}{lll}
\min & z = \sum_i \sum_k b_{ik} g_{ik} + \sum_i \sum_j \sum_t a_{ijt} c_{ijk} & \\
\text{s.t.} & \sum_i a_{ijt} - d_{jt} = 0 & \forall j, t \\
& \sum_k x_{ikt} - \sum_j a_{ijt} \geq 0 & \forall i, t \\
& x_{ikt} \leq N b_{ik} & \forall i, k, t \\
& x_{ikt} \leq u_{ikt} + N(1 - b_{ik}) & \forall i, k, t \\
& u_{ikt} \leq x_{ikt} + N(1 - b_{ik}) & \forall i, k, t \\
& u_{ikt} - (\widehat{u}_{ikt} + (\widehat{e}_{ikt} \widehat{u}_{ikt})) \leq 0 & \forall i, k, t \\
& u_{ikt} - (\widehat{u}_{ikt} - (\widehat{e}_{ikt} \widehat{u}_{ikt})) \geq 0 & \forall i, k, t \\
& p(s_{ikt} \geq u_{ikt}) + N(1 - b_{ik}) - \beta_{ikt} \geq 0 & \forall i, k, t
\end{array}
$$

Briefly, the objective function of the model is composed of the sum of the supply costs and the costs associated with the harvested stands. In addition, there are eight constraints. The first one ensures that the demands of all mills are satisfied. The second, third, fourth and fifth constraints are the linearization of the constraint $\sum_k u_{ikt} b_{ik} - \sum_j a_{ijt} \geq 0, \forall i, t$, which ensures that the supply does not exceed the capacity of the stands selected for harvesting. For such purpose, the third, fourth and fifth constraint fix the value of an auxiliary vector of variables called x. The sixth and seventh constraints determine the uncertain intervals of the random variables. Finally, the eighth constraint ensures a minimum level of robustness for each random variable only if its associated stand is selected for harvesting.

B. Evaluation of our Approach

A graphical representation of a forestry instance is shown in Figure 5. Note that for simplicity there is only one type of log and the cost of harvesting any stand is zero. There are two forests, three stands and two customers, each of whom demands 100 units of log product. Regarding the error of measurements, we consider a typical bounds estimate in this industry: 10% ($\widehat{e}_{ik} = 0.1$, $\forall i \forall k$). When solving such instances, non-stochastic approaches (such as [9], see Section III-A) do not take into consideration the limited probabilistic data. Therefore, non-robust values can be assigned to random variables with a low likelihood of being greater/lower (according to the order relationship given by the problem) or equal to the value estimated (such as s_{21} in Figure 5, where $p(s_{21} \geq \widehat{u}_{21}) = 50\%$). However, with our approach we can assign the most robust values to the random variables with the lowest associated likelihoods.

Following we show the results of applying the extrapolation method introduced in Section IV and the MIP model introduced above. Note that we cannot directly apply (without using our extrapolation approach first) a stochastic approach (see Section III-B) because a probability is not associated with every value in the uncertainty set. We also computed the solution of the deterministic case, in which the values expected are completely certain (there are neither chance constraints nor uncertainty sets in the model). For the implementation we used the Numberjack modeling package and the CPLEX solver. The experiments were run on a 2.3 GHz Intel Core $i7$ processor and the computation time was approximately $0.093s$ for all the solutions. We would like to highlight that the increase of computation time of our robust solutions in comparison with the deterministic solution was insignificant for this instance. For extrapolating the probabilities, we used the decreasing cumulative distribution $p(s_i \geq \widehat{u}_i)$ because the type of error that could invalidate a solution is that the real value is lower than the nominal value (here, we used the uniform distribution, see Figure 2).

Table II shows several solutions obtained for different robustness bounds (β). The lowest probability is $p(s_{21} \geq$

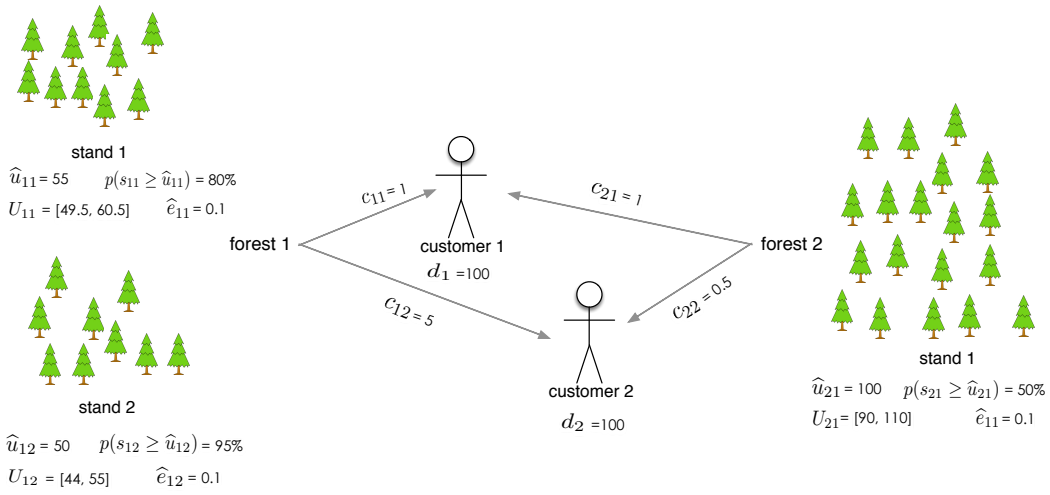

Figure 5. Forestry Industry Example ($\mathcal{F} = 2, \mathcal{S} = 3, \mathcal{M} = 2, |\mathcal{L}| = 1$).

$\widehat{u}_{21}) = 0.5$, therefore, only for higher values of β can we appreciate the effect of the chance constraints. Thus, for the deterministic approach and $\beta \in (0, 0.5]$ the solution obtained is optimal because it has the lowest possible cost: 150. Note that forest 1 provides 100 units of product to the customer 1 and forest 2 provides 100 units to the customer 2. However, this solution is not very robust because there is 50% of probability that forest 2 has less capacity than expected and in such case, the obtained solution will become invalid.

More robust solutions can be obtained by sacrificing optimality, and consequently increasing the cost. For ensuring a higher robustness in the random variable s_{21}, lower values of capacity than the nominal one have to be selected. Thus, in Table II $s_{21} \geq 97$ for $\beta = 0.65$ and $s_{21} \geq 94$ for $\beta = 0.76$. As a consequence, for fulfilling the demand of costumer 2, forest 1 has to provide him the rest of product. Note that in contrast to customer 1, the cost of supplying customer 2 varies significantly depending on the forest. In particular, the supply cost of forest 2 is only one tenth of the supply cost of the forest 1. Forest 2 has 5 extra units of product according to its nominal values; therefore, it can supply up to 5 units of product without sacrificing the robustness associated with the nominal values (this is the case for $\beta = 0.65$). Nevertheless, for $\beta = 0.76$, customer 2 requires 6 extra units from the forest 1, and for this reason the amount selected for one of the stands is one unit greater than its nominal value, with the decrease of robustness that this fact entails ($p(s_{12} \geq 51) = 0.76$). Note that there is a probability of at least 76% that each stand of each forest has a capacity greater or equal than selected for such a solution and consequently, remaining a solution. We would like to note that for $\beta > 0.76$ there is no solution, therefore this β bound provides the most robust solution for this instance.

The trade-off between robustness and optimality can be interpreted as follows for problems with uncertain stocks (such as the analyzed one). Robustness is achieved by assuming a lower estimate of capacity, which in some cases can produce an overstock of product, which is known as "the price of robustness". Because this solution is based on selecting a lower capacity than the nominal value estimated, if the latter is correct, the difference between these two values is equivalent to the amount of overstock. Considering overstock in the solutions provides them with robustness since they have a high likelihood to remain valid if faced with a lower availability of the original stock than assumed. However, these buffers in the stocks have the disadvantage that they also entail scrap costs (see the column of costs in Table II for the two greatest β values). Even so, and as mentioned in Section I, it is important to take into account that non-robust solutions for this type of optimization problems can lead to serious economic losses when resources are less than expected, and therefore customers cannot be adequately supplied. This not only entails the loss of a sale, but also customer dissatisfaction, with a resulting loss in demand for the product or use of the service.

Table II
SOLUTIONS OBTAINED FOR DIFFERENT CERTAINTY BOUNDS (β).

β	**SUPPLY**	**AMOUNT**	**PROBABILITY** (p)	**COST**
deterministic	$a_{11} = 100$	$s_{11} \geq 55$	$p(s_{11} \geq 55) = 0.8$	cost of $a_{11} = 100$
approach &	$a_{12} = 0$	$s_{12} \geq 50$	$p(s_{12} \geq 50) = 0.95$	cost of $a_{12} = 0$
$\forall \beta_i \in (0, 0.5]$	$a_{22} = 100$	$s_{21} \geq 100$	$p(s_{21} \geq 100) = 0.5$	cost of $a_{22} = 50$
				Total cost $= 150$
$\forall \beta_i = 0.65$	$a_{11} = 100$	$s_{11} \geq 55$	$p(s_{11} \geq 55) = 0.8$	cost of $a_{11} = 100$
	$a_{12} = 3$	$s_{12} \geq 50$	$p(s_{12} \geq 50) = 0.95$	cost of $a_{12} = 15$
	$a_{22} = 97$	$s_{21} \geq 97$	$p(s_{21} \geq 97) = 0.65$	cost of $a_{22} = 48.5$
				Total cost$= 163.5$
$\forall \beta_i = 0.76$	$a_{11} = 100$	$s_{11} \geq 55$	$p(s_{11} \geq 55) = 0.8$	cost of $a_{11} = 100$
	$a_{12} = 5.2$	$s_{12} \geq 51$	$p(s_{12} \geq 51) = 0.76$	cost of $a_{12} = 26$
	$a_{22} = 94.8$	$s_{21} \geq 94.8$	$p(s_{21} \geq 94.8) = 0.76$	cost of $a_{22} = 47.4$
				Total cost$= 173.4$
$\forall \beta_i > 0.76$	no solution			

VI. CONCLUSIONS AND FUTURE WORK

In this paper we present a strategy for dealing with real-life LSCO problems in which some of the data is uncertain. Since information about uncertainty is limited, our approach extrapolates the estimates that we have based on the order of elements in the uncertain domains. Although we could have used other stochastic methods, in this work our approach was combined with chance constraints in order to obtain robust solutions, which have a high likelihood of remaining valid in the face of differences in the actual values of the uncertain parameters. The main advantage of the presented approach over pure stochastic models is that it can deal with a lack of probabilistic information. In addition, it has the advantage over non-stochastic models in that by combining it with chance constraints, it is able to spread the robustness more uniformly across all uncertain parameters, because for each one, a minimum specific robustness bound can be fixed. Another advantage is that our approach is able to consider situations with limited existent probabilistic data associated with the values estimated.

We have applied our approach to an illustrative problem based on real-life problems from the forestry industry. For this purpose, we designed a MIP model that selects the best stands of forests to harvest for satisfying customers' demand while minimizing costs. In addition it ensures minimum robustness bounds according to the uncertain capacities of the forest stands. We can therefore handle the well-known trade-off between robustness and optimality, which can be controlled by means of minimum robustness bounds. In the future, we plan to apply this approach to real-life forestry data in collaboration with our industrial partner.

ACKNOWLEDGMENT

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

REFERENCES

[1] N. Yorke-Smith and C. Gervet, "Certainty closure: Reliable constraint reasoning with incomplete or erroneous data," *Journal of ACM Transactions on Computational Logic (TOCL)*, vol. 10, no. 1, p. 3, 2009.

[2] A. Charnes and W. W. Cooper, "Chance-constrained programming," *Management science*, vol. 6, no. 1, pp. 73–79, 1959.

[3] L. Climent, R. J. Wallace, M. A. Salido, and F. Barber, "Finding robust solutions for constraint satisfaction problems with discrete and ordered domains by coverings," *Artificial Intelligence Review (AIRE), DOI 10.1007/s10462-013-9420-0*, 2013.

[4] L. Climent, R. J. Wallace, M. A. Salido, and F. Barber, "Robustness and stability in constraint programming under dynamism and uncertainty," *Journal of Artificial Intelligence Research*, vol. 49, pp. 49–78, 2014.

[5] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust optimization*. Princeton University Press, 2009.

[6] G. Verfaillie and N. Jussien, "Constraint solving in uncertain and dynamic environments: A survey," *Constraints*, vol. 10, no. 3, pp. 253–281, 2005.

[7] A. L. Soyster, "Convex programming with set-inclusive constraints and applications to inexact linear programming," *Operations Research*, vol. 21, no. 5, pp. 1154–1157, 1973.

[8] H. Fargier, J. Lang, and T. Schiex, "Mixed constraint satisfaction: A framework for decision problems under incomplete knowledge," in *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96)*, 1996, pp. 175–180.

[9] D. Bertsimas and M. Sim, "The price of robustness," *Operations Research*, vol. 52, no. 1, pp. 35–53, 2004.

[10] R. Epstein, R. Morales, J. Seron, and A. Weintraub, "Use of OR systems in the chilean forest industries," *Interfaces*, vol. 29, no. 1, pp. 7–29, 1999.