International Journal on Artificial Intelligence Tools

Extrapolating from Limited Uncertain Information in Large-Scale Combinatorial Optimization Problems to obtain Robust Solutions *

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

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

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 and evaluated with real-world applications of harvesting and supplying timber from forests to mills and the well known knapsack problem with uncertainty.

Keywords: uncertainty; robustness; optimization.

1. Introduction

In this paper, we deal with real-life problems in which uncertainty is associated with elements that have an ordered relationship. In these problems, measurement/estimation errors are common, and they result in a partially incorrect representation of the modeled problem. 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 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 continuous/discrete probability distributions.

In this paper we introduce an approach for handling these situations, by extrapolating data about changes over the original formulation of the uncertain parameters,

*This paper is an extended version of L. Climent et. al. in proceedings of ICTAI-2014.

based on the order relationship associated with the domains of such parameters. 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 (Ref. 2) for the forestry problem of timber supply. We also apply the extrapolation approach to the knapsack problem with uncertain weights. For designing the extrapolation approach, we have been inspired by some ideas in Refs. 3 and 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 tie 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. When this occurs, a new solution can be computed. However, it requires computation time. This means that in on-line problems, the solution might not be computed in time. Furthermore, the loss of a solution typically causes several negative effects in the modeled situation (even if a new solution is found). For example, in a task assignment of a production system with several machines, it could cause the shutdown of the production system, the breakage of machines, the loss of the material/object in production, etc. All these negative effects will probably entail an economic loss as well.

To reduce the chances of losing a solution, it is important to obtain a robust solution, which has a high likelihood of remaining valid given the uncertainties of the problem. And this is indeed the motivation of the extrapolation technique presented in this paper. Our technique is able to extrapolate the likelihood associated with parameters whose values are uncertain according to a cumulative distribution function. There is a requirement for performing such actions: the possible elements associated with the uncertain parameter must have a significant order relationship over them. Thus, minimum levels of robustness are ensured for the uncertain parameters of the problems based on the probability information that we extrapolate. In addition, this model can be combined with another other optimization criterion.

This paper is structured as follows: First, in Section 2 some definitions used in the paper are explained. Section 3 gives a brief review of related work, which helps to motivate our approach. In Section 4, we introduce an approach for extrapolating from limited uncertain information associated with some parameters of the problems. In Section 5 and Section 6 we apply our approach to well-known problems with uncertainties. Finally, we give some conclusions in Section 7.

2. Technical Background

In this section we provide some standard definitions for modeling combinatorial optimization problems in the CP and OR field. See Ref. 5 for integrated methods for optimization in such fields.

Definition 2.1. A Constraint Satisfaction Problem (CSP) is represented as a triple $P = \langle \mathcal{X}, \mathcal{D}, \mathcal{C} \rangle$ where $\mathcal{X} = \{x_1, \ldots, x_n\}$ is a finite set of variables , $\mathcal{D} = \{D_1, \ldots, D_n\}$

is a set of domains such that for each variable $x_i \in \mathcal{X}$ there is a set of values D_i that the variable can take, and $\mathcal{C} = \{C_1, \ldots, C_m\}$ is a finite set of constraints which restrict the values that the variables can simultaneously take.

Definition 2.2. A Constraint Satisfaction and Optimization Problem (CSOP) is an augmented model of CSP that introduces some objective functions. The objective is to maximize/minimize the set of functions $f(s)$ for $s \in \mathcal{S}(\text{CSP})$ (see Ref. 6).

Definition 2.3. 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.

Below, we show a definition extracted from Ref. 7. In such reference, there are also detailed explanations about robustness and other characteristics of solutions of problems that come from uncertain and/or dynamic environments.

Definition 2.4. The most robust solution within a set of solutions is the one with the highest likelihood of remaining a solution after any type of change.

3. 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 Ref. 8 and a summary of CP approaches can be found in Ref. 9. There are also other fields that deal with uncertain environments, such as metaheuristic approaches (for instance evolutionary computation, see Ref. 15). In this section, we consider some CP and OR approaches, focusing on models that assume that some parameters of the problem can take any value within some range. There are also other ways of representing uncertainty; for instance, as previously mentioned, in Refs. 3 and 4, the constraints and domains may become more restricted over time. However, we focus on modeling uncertainty in the measurement/estimation of certain parameters. In Section 3.1 we discuss models that do not consider probabilistic data. In addition, in Section 3.2 we discuss stochastic models.

3.1. *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 uncertainty 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 uncertainty sets can be discrete or continuous. In this framework, a conservative definition of a robust solution was given in Ref. 10.

Definition 3.1. A solution is robust iff it satisfies all the constraints, whatever the realization of the data from the uncertainty sets.

According to this definition, a solution is robust iff it satisfies all the constraints for all the possible combinations of values from the uncertainty sets associated with the uncertain parameter. A CSP model that considers Definition 3.1 with discrete uncertain domains is the Mixed CSP model (MCSP), whose main object 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 (Ref. 11). Another type of model, the Uncertain CSP model (UCSP), is an extension of the MCSP because it considers continuous domains (Ref. 1). However, Definition 3.1 is very conservative and demanding since there are often no solutions that satisfy this requirement for every uncontrollable 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 the less conservative approach that is described next.

In Ref. 13, 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. $\Gamma_i \in [0, |J_i|]$, where $J_i \subseteq S$ is the set of random variables of the ith uncertain constraint. Note 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-life 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 may be important that some variables are able to satisfy all the values in the uncertainty set, but it is more useful to ensure that all variables are able to satisfy some of the values in their

uncertainty sets. Thus, in the latter case robustness is spread more evenly among all of the random variables. In addition, the above described technique does not take into consideration situations with limited existent probabilistic data associated with the values estimated.

3.2. *Stochastic programming*

Another area that deals with optimization subject to uncertainty is stochastic programming (Ref. 14). 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 (Ref. 16). In either approach, an 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 U_i. 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 random variables produce different scenarios. The likelihood of occurrence of a scenario, when the random variables represent independents events is $\prod_{i=1}^{i=|S|} p(s_i = u_i)$, where $u_i \in U_i$ and p is a probability mass function associated with s_i. There are also scenario-based approaches (e.g. Ref. 12) that generate a large number of possible scenarios in order to compute probabilistic information about the uncertainties. However, computing such a large number of possible combinations carries a high cost. This disadvantage is exacerbated for LSCO problems, since solving these problems already requires a large amount of computation. Therefore, adding scenario-generation may well involve a greater computation time than the time available for solving a real-life LSCO problem.

Another approach is to use chance constraints (Ref. 2). These constraints are composed of at least one random variable that makes it possible to state a minimum probability of satisfying the constraint. Thus, if we consider a LO model (see Definition 2.3) that includes some random variables, a chance constraint would be defined as: $p(Ax \leq b) \geq \beta$, where "p" denotes the probability of satisfaction and β is the set of prescribed confidence levels.

The advantages of the stochastic approaches mentioned above are that they are less conservative than Definition 3.1 and they are able to spread the robustness across all of the random variables. The major limitation of such approaches in the situations we envisage is the precise and extensive knowledge of probabilities that such models require, with the exception of scenario-based approaches, which as previously mentioned require a high extra cost. 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

non-existent, and in these situations stochastic models cannot be applied.

4. Extrapolating from Limited Uncertain Data

The present approach was motivated by the limitations of the models and strategies described in the previous section. The new method involves extrapolating from an estimated value in a way that can be combined subsequently with a particular stochastic approach, that of chance constraints. By means of the β parameter of the chance constraints, we are able to fix robustness bounds for each random variable. 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 Sections 5 and 6 we show how this extrapolation method can be combined with chance constraints to solve real-life optimization problems.

The main requirement of this approach is that for each uncertainty set, whether continuous or discrete, there is a monotonic relation over the elements of each domain of the problem for which the model has been generated. This order relationship is required for the representation of uncertainty. In lieu of detailed information regarding probabilities, we can use any cumulative probability distribution over an ordered domain that has a corresponding monotonic relation. This section covers the explanation of the uncertain ordered intervals, the probabilities associated with them and the cumulative distributions.

4.1. *Uncertain ordered intervals*

The present approach to uncertain ordered domains is based on the concept of nominal values.

Definition 4.1. The nominal value of an uncertain parameter represents the most likely value that such parameter will take given the study conditions.

Typically, after measuring the variables associated with a problem, a nominal value (denoted as $\widehat{u}_i$) will 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 settings, an estimate can be made of the maximum amount by which the random variable might deviate from $\widehat{u}_i$ (denoted as $\widehat{e}_i$). This interval of deviation U_i of s_i can be denoted as: $[(\widehat{u}_i - \widehat{e}_i), (\widehat{u}_i + \widehat{e}_i)]$. Here we assume for simplicity that $\widehat{u}_i$ is at the midpoint of the interval, although our methods do not depend on this assumption. In the problems to be described in Sections 5 and 6, the error estimate is a fixed percentage of the nominal value. This follows standard practice in our main domain of application. In this case, the interval of the random variable 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 Refs. 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 magnitude of "restrictive" changes over the solution space that they could handle. To support this assumption, there had to be an ordered relationship over the domains. For example, a time-buffer following a scheduled task (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 restrictive (e.g. delays in 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 considered in this paper, it is often the case that values on one side of an estimate represent cases that allow larger or smaller 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 there are true values with varying likelihoods given an uncertain estimated value.

To illustrate, let $s_i \in S$ be a random variable with an ordered domain $[u_1, u_2, u_3]$ that is strictly increasing/decreasing. If the value u_2 allows larger restrictive deviations than u_1 and u_3 allows larger restrictive deviations than u_2 (according to the given order relationship), then a solution that is feasible for $s_i \leq u_3$ is more likely to be valid 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. Here we call solutions more "robust" if they 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 in the opposite direction (u_1 is the value that allows the greatest deviation), then the robustness ordering is opposite as well. That is, solutions feasible for $s_i \geq u_1$ are the most robust, i.e. they are the ones most likely to be valid.

4.2. *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 model (see Section 3.2), the probability distribution defined over the uncertain ordered interval expresses the likelihood 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 $u_i \in U_i$, the uncertainty set associated with s_i. In this case, the calculated 'probability' is directly related to solution robustness because, given a solution, the higher this probability value, the higher the likelihood that the real value of the random variable is feasible for this solution. Therefore, in the example just described, 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 in Table 1. 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 the restrictiveness order. On the one hand, when they represent demands, it is more likely that the real demand is lower or equal to a large value than to a small one. On the other hand, if we consider capacities, it is more likely than the real capacity is greater or equal to a small value than to a large one. In this case, if a capacity value meets a certain demand, the solution is more likely to be feasible for lower demands than the expected one.

Table 1. Examples of Ordered Domains and the Probabilities (p) associated with their Random Variables.

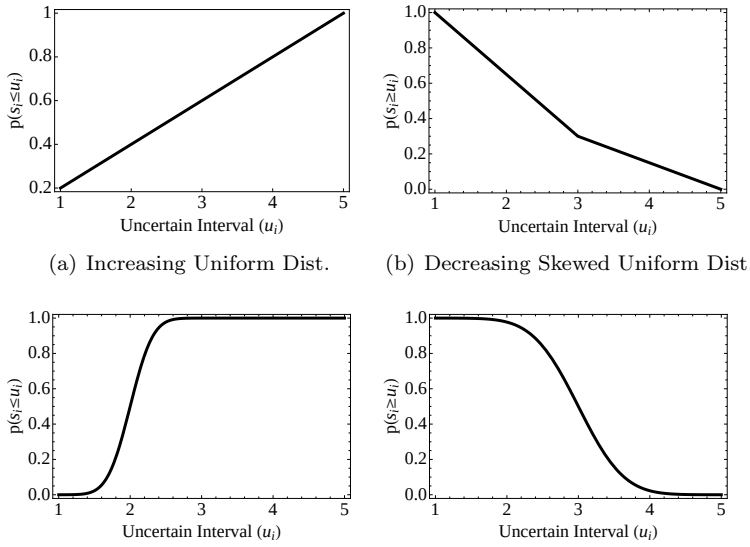
DOMAIN	**PROBABILITIES** (p)
Demand: $\overrightarrow{[1,2,3]}$	$p(s_i \leq 1) < p(s_i \leq 2) < p(s_i \leq 3)$
Capacity: $\overleftarrow{[1,2,3]}$	$p(s_i \geq 1) > p(s_i \geq 2) > p(s_i \geq 3)$

4.3. *Cumulative probability distributions*

In the present approach we are assuming that there is no specific probabilistic information associated with the interval of uncertainty; therefore, we need a way of extrapolating it. For this purpose we use the ordering of values over the uncertain interval. As mentioned earlier, the maximum deviation 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 $(0, 1]$.

In this paper, we describe the extrapolation using cumulative probability distributions that come from either the uniform or the normal distributions (see Figure 4). However, any cumulative distribution (with strictly monotonic increasing/decreasing probability) could be used instead. This is because we are not using the distribution to model probabilities but only to order the likelihoods associated with domain values on either side of the estimate. (A specific distribution may, however, reflect our intuitions about the rate at which likelihoods change.) With the uniform distribution there is a constant increase/decrease in the cumulative probability over the interval on each side of the estimate ($\widehat{u}_i$). For the normal distribution, the increase/decrease is more abrupt for values closer to $\widehat{u}_i$.

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 sufficient, especially



(a) Increasing Uniform Dist. (b) Decreasing Skewed Uniform Dist.

(c) Shifted Increasing Normal Dist. (d) Decreasing Normal Dist.

Fig. 1. Cumulative Distributions.

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. Without loss of generality, we use continuous probabilities distributions for the description of our approach. However, discrete probabilities distributions can also be used (such as binomial or truncated distributions) for discrete domains.

We can also accommodate situations in which the cumulative probability associated with the estimated value for a random variable (nominal value) is known. We use this variation in situations in which we know from past experiences that our estimates tend to be higher/lower precise. 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 1(c)) or the distribution is skewed and there are two specific gradients for values greater and lower than the nominal value (see Figure 1(b)).

The extrapolation of the uniform skewed cumulative probability distribution is defined in Equation 1 for the increasing and decreasing case:

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

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

The equation for the increasing cumulative normal distribution is:

$$p(s_i \leq u_i) = p(s_i \leq \widehat{u}_i) + \frac{1}{\sigma\sqrt{2\pi}} \int_{\hat{u}_i}^{u_i} e^{-\frac{(u_i - \hat{u_i})}{2\sigma^2}} \, \mathrm{d}u \tag{2}$$

A similar equation holds for the decreasing case (see Figure 1(d)) except that the integral limits are reversed. Since we are using this distribution as a stand-in for the actual unknown distribution, it seems reasonable to use the normal distribution in its standardized form. This gives:

$$p(s_i \leq u_i) = p(s_i \leq \widehat{u}_i) + \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{z^2}{2}} \, \mathrm{d}z \tag{3}$$

Unfortunately, the resulting integral is still not easy to resolve. Perhaps the most straightforward strategy is to use a table of values for the cumulative normal distribution, choosing a value for the expression associated with the value of u_i and interpolating when necessary.

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. 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 can be used. In line with the motivation given in Section 3, we introduce two models incorporating chance constraints in Sections 5 and 6.

5. A Case Study from Forestry

To illustrate these ideas in practice, we present an application of our approach to a well known real-life LSCO problem of forestry. This type of problem is of great importance for many countries that export timber. The process of harvesting stands from forests starts with the cutting of trees that belong to the selected area for harvest. Subsequently, the trees are bucked into different types of log-products. To estimate the likely harvest, 3D measurements are made of a sample of tree stems by means of laser scans. There are various simulation programs that can use this information to determine the patterns in which the stem should be cut, which allows the product assortment to be pre-estimated. However, there are many factors that

contribute to the uncertainty in the estimation of such timber offer, for instance: (i) environmental factors, such as bad weather conditions, pests, etc. that change the expected growth of the trees, and (ii) error measurements due to the quality of the laser scans (some trees may be partially occluded by others in the images, or factors such as fog can affect the quality of the images).

The logistics process includes transportation of different types of timber to the mills (customers that set the demand). There are different costs associated with the transportation process, which should be minimized in order to obtain the maximum profit from the sales of logs. To this end, we develop a linear model for such problems in Section 5.1. In Section 5.2 for illustrative purposes we apply this model to a toy instance and then to a large instance.

5.1. *Model of forestry problem that handles uncertainty*

In this section we describe a Mixed Integer Programming (MIP) model, which is a type of LO model (see Definition 2.3) in which some variables are constrained to be integers. The model presented includes our extrapolation technique (see Section 4) and it also minimizes the transportation costs. We control the level of robustness of the random variables by incorporating a chance constraint in the model (see Section 3.2) 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 $p(s_i \leq u_i)$. Since the random variables that we are analyzing represent capacities, the type of error that could invalidate a solution is that the real value is lower than the nominal value, therefore, the appropriate cumulative distribution for such random variables is the decreasing one: $p(s_i \geq \widehat{u}_i)$ (see Equation 1). The uncertain intervals U_i are computed given the input parameters $\widehat{u}_i$ and $\widehat{e}_i$ (see Section 4). The components of this MIP model are as follows:

(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 supplying one unit of the log-product t from the forest i to the mill j.
$\widehat{u}_{ikt}$: Capacity estimated for the forest i for each of its stands k for each type of log-product t.
$\widehat{e}_{ikt}$: Percentage of variability for the forest i for each stand k for each type of log-product t.
d_{jt} : Demand of the mill j of the type of log-product t.
β_{ikt} : Vector of minimum probabilities.

N : Large constant.

(3) Variables:

$b_{ik} \in \{0, 1\}$ (1 if stand k from forest i is harvested, otherwise 0).
a_{ijt} = Amount supplied of log-product t from forest i to mill j.
$u_{ikt} \in [(\widehat{u}_{ikt}(1 - \widehat{e}_{ikt})), (\widehat{u}_{ikt}(1 + \widehat{e}_{ikt}))]$ = Minimum capacity selected for the forest i for the stand k for the type of log-product 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:

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

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 six constraints. The first ensures that the demands of all mills are satisfied. The constraints $2-5$ 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 this purpose, constraints $3 - 5$ fix the value of an auxiliary vector of variables called x. Finally, the sixth constraint ensures a minimum level of robustness for each random variable only if its associated stand is harvested.

5.2. *Evaluation of forestry problem using the uncertainty model*

In this section we first show results for a toy problem (due to its easy representation). Subsequently, we evaluate a large scale instance in order to check the scalability of our approach with LSCOs problems. We show the results of applying the extrapolation method introduced in Section 4 and the MIP model introduced above. Note that we cannot directly apply (without extrapolating probabilities first) a stochastic approach (see Section 3.2) because a probability is not associated with every value in the uncertainty set. For extrapolating the probabilities, we used the decreasing uniform cumulative distribution (see Equation 1). This distribution is skewed (see Figure 1(b)) for stands in which the probability associated with the estimated nominal value is not 0.5. For such cases, non-stochastic approaches (such as Ref. 13, see Section 3.1) do not take into consideration the limited probabilistic data. Therefore, non-robust values may 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. However, with our approach we can assign the greater values to random variables with the highest associated likelihoods. For the implementation of the model we used the Numberjack modeling package and the CPLEX solver. The experiments were run on a 2.3 GHz Intel Core $i7$ processor. For bounding the error of measurement, we adopt a typical estimate used in this industry: 10% ($\widehat{e}_{ik} = 0.1$, $\forall i \forall k$).

5.2.1. *Toy instance*

A pictorial representation of an instance of the forestry problem is shown in Figure 2. There are two forests, three stands and two customers, each of whom demands 100 units of a log-product. For simplicity there is only one type of log-product and the cost of harvesting any stand is zero. The figure shows the customer demands in units of log product d_j ($j \in \mathcal{M}$) and the number of units of log product estimated for each stand $\widehat{u}_{ik}$ ($i \in \mathcal{F}, k \in \mathcal{S}$). It also shows the transportation costs for supplying each unit of log product from each forest to each costumer c_{ij} ($i \in \mathcal{F}, j \in \mathcal{M}$). The first forest is composed of two stands. The capacity estimated for the first stand is $\widehat{u}_{11} = 55$ units; for the second it is $\widehat{u}_{12} = 50$ units. The second forest has only one stand with an estimate of $\widehat{u}_{21} = 100$ units.

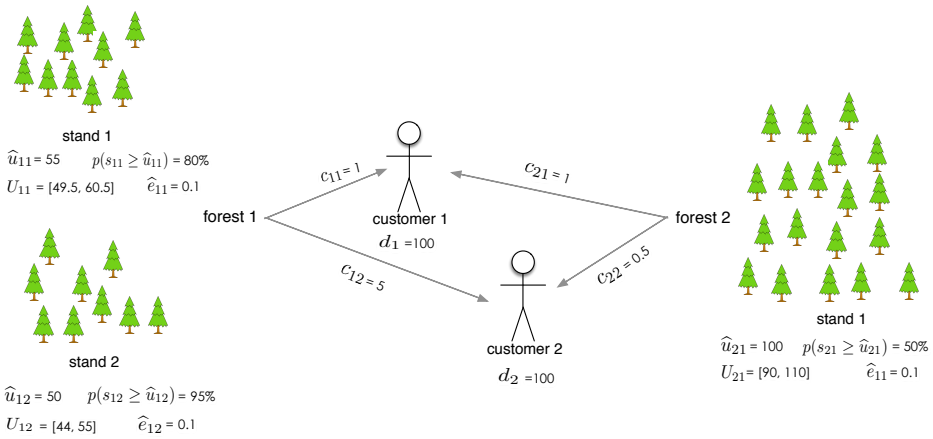


Fig. 2. Forestry Industry Example with $|\mathcal{F}| = 2$ (no. forests), $|\mathcal{S}| = 3$ (no. stands), $|\mathcal{M}| = 2$ (no. mills) and $|\mathcal{L}| = 1$ (no. of types of log-products).

Table 2 shows several solutions obtained for the problem described above for different robustness bounds β (cf. constraint No. 6). We also computed a solution for the deterministic case, in which the expected values are completely certain (there are neither chance constraints nor uncertainty sets). The computation time was approximately $0.093s$ in each case (the increase of computation time for robust

solutions was insignificant). The lowest probability of this instance is $p(s_{21} \geq \widehat{u}_{21}) = 0.5$ (the other probabilities are: $p(s_{11} \geq \widehat{u}_{11}) = 0.8$ and $p(s_{12} \geq \widehat{u}_{12}) = 0.95$), therefore, only for higher values of β can we see the effect of the chance constraints. Thus, for the deterministic approach and the non-deterministic approach with $\beta \in (0, 0.5]$ the solution obtained is optimal because it has the lowest possible cost: 150. In this case 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 this case, the obtained solution will become invalid.

Table 2. 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			

More robust solutions can be obtained by sacrificing optimality, and consequently increasing the cost associated with the deterministic case. 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 2, $s_{21} \geq 97$ 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 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. For $\beta > 0.76$ there is no solution; therefore, this β bound provides the most robust solution for this instance.

5.2.2. *Large scale instances*

In order to check the scalability of our approach for LSCOs problems, we implemented a random instance generator (using the random uniform distribution) based on the information of real-life applications of our industrial partner. We set the number of forests to 30, each of them composed by 8 stands, which means that the total number of stands are 240. Since we fixed the number of types of log-products to 10, there are 2400 random variables. We would like to mention that these input parameters represent bigger instances than the typical real-life instances that our industrial partner deals with. The number of customers that buy log-products in a certain interval of time (e.g. a week) is not usually very high. For this reason, we considered 10 customers, each demanding a random amount in $[0, 150]$ of each log-product. The transportation costs from the forests to the costumers for all the types of log-products were also randomly selected in $[0.1, 10]$. The cost of harvesting a stand (any of them) was set to 50 units. We computed random nominal values in $[0, 50]m^3$ for each type of log-product t for each stand k of each forest i and their associated probabilities ($p(s_{ikt} \geq \widehat{u}_{ikt})$) were randomly computed in $[0.5, 0.9]$.

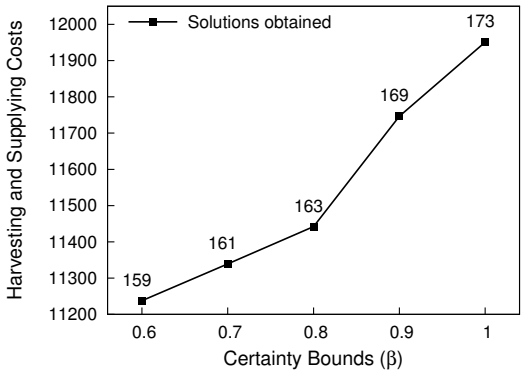


Fig. 3. Costs and No. of harvested blocks (above the points) of the solutions obtained.

Figure 3 shows the associated costs of the solutions obtained for different levels of robustness (β) and also the number of stands that were necessary to harvest for satisfying the demands (numbers located next to the points). The scalability of our approach was very good, since a robust solution was obtained in less than six seconds. In contrast to the previous instance analyzed, it is possible to obtain a solution that satisfies all the constraints of the problem for all the values of all the uncertain sets (for $\beta = 1$ we obtain the most robust solution). Of course, this robustness involves an increment in the harvesting and supplying cost (see in Figure 3 how 14 stands more are harvested from $\beta = 0.6$ to $\beta = 1$). The trade-off between robustness and optimality can be interpreted as follows for problems with uncertain stocks (like the one analyzed here). 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" (Ref. 13). 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 solutions ensures their robustness since they have a high likelihood of remaining valid faced with a lower amount of the original stock available than assumed. However, these buffers in the stocks have the disadvantage that they also entail scrap costs (see the column of costs in Table 2 for the two greatest β values and vertical axis in Figure 3). Even so, and as mentioned in Section 1, 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.

6. Knapsack Problem with Uncertainty

To show the generality of the approach presented in this paper, we also evaluate it with a well known problem from the literature on combinatorial optimization: the knapsack problem (Ref. 17). In this problem there is a set of items, and each type of item has an associated weight and value. Items are to be collected, given a bound on the total allowable weight. The objective is to select the combination of items that is the most valuable without exceeding this bound.

An issue that we may face in solving a knapsack problem in an uncertain and/or dynamic environment is that the measured weights of the items are uncertain. Then, depending on the level of certainty of such measurements and also their range of possible variability, some solutions will be more robust than others. Typically, the approaches from the literature assume that the uncertainty about the weight measurements is evenly distributed (see for instance Ref. 13). However, in real applications some measurements are usually less accurate than others (for instance, in items that are very big/small for the scale used, items composed by a liquid/gas with certain tiny evaporation rate, etc.). In other cases it may be desirable to specify a different robustness for different items (neither situation is considered in Ref. 13). For instance, it may be useful to set the certainty bound of a valuable object at a higher level than the certainty bound of a cheaper item, since the more valuable an item it is, the greater the loss in value for a solution if the item weighs more than expected (and therefore cannot be collected). In this section we explain the model for such knapsack problems with uncertainty and then we present some experimental results for a real-life problem.

6.1. *Model for the knapsack problem with uncertainty*

In this section we present a CSOP model (see Definition 2.2) for the problem with uncertainty described above that incorporates the extrapolation technique proposed

in this paper. For modeling uncertainty in the weights of the items, these must be represented as random variables, which generates non-linear constraints. This motivates modeling the problem as a CSOP and solving it with CP techniques. To convert rational values to integer values (necessary for applying CP techniques), we increase the order of magnitude of the values and then round them. The appropriate cumulative distribution for such random variables is an increasing one: $p(s_i \leq \widehat{u}_i)$ (see Equation 1). This is because the random variables represent weights and therefore the type of error that could invalidate a solution is one in which the actual value is greater than the nominal value. The parameters of the problem are as follows:

- Sets of types of items $\mathcal{Z}$ ($i \in \mathcal{Z}$).
- m_i: Maximum allowed quantity of elements of type of item i.
- v_i: Value of the item i.
- $\widehat{w}_i$: Estimated weight of the item i.
- $\widehat{e}_i$: Percentage of variability in the weight of item i.
- β_i : Vector of minimum certainty bounds.
- W : Maximum total weight bound.

The CSOP model is as follows:

$$\mathcal{X} = \{x_1, x_2, ..., x_{|\mathcal{Z}|}, w_1, w_2, ..., w_{|\mathcal{Z}|}, s_1, s_2, ..., s_{|\mathcal{Z}|}\}$$

$$\mathcal{D} = \begin{cases} [0, m_i] & x_i \\ [(\widehat{w}_i(1-\widehat{e}_i)), (\widehat{w}_i(1+\widehat{e}_i))] & w_i/s_i \end{cases}$$

$$\begin{array}{ll} \max & \sum_i v_i x_i \\ \text{s.t.} & \sum_i w_i x_i \leq W \\ & p(s_i \leq w_i) \geq \beta_i \quad \forall i \end{array}$$

As in the usual knapsack problem, the objective is to maximize the total value of all the elements collected. As before, the uncertain intervals U_i of the random variables $s_i \in S$ are computed given the input parameters $\widehat{w}_i$ and $\widehat{e}_i$ (see Section 4). The first constraint ensures that the total maximum weight bound is not exceeded by the collected items. The level of robustness of the random variables (and therefore the maximum weights of the items accepted by the solution) is controlled by the remaining constraints in the model, which are chance constraints (see Section 3.2).

6.2. *Evaluation with the knapsack problem with uncertainty*

To evaluate our approach we used a real-life problem presented in Ref. 18. In this problem, there is a set of advertisements of certain durations and values, which can be selected for broadcasting up to a specific number of times. Table 3 (from Ref. 18) shows the characteristics of the commercials. The authors of the original article do

not consider uncertainties in the durations of the advertisements. For this reason, their obtained solution will not remain a solution if any advertisement undergoes a delay of a few seconds.

Table 3. List of commercials and their characteristics.

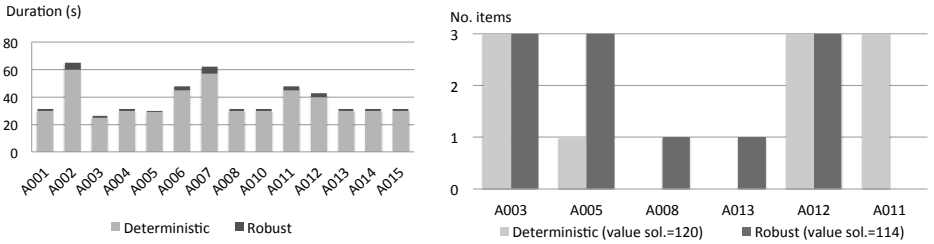
Advert A0XX (x_i)	01	02	03	04	05	06	07	08	10	11	12	13	14	15
Max. no. (m_i)	2	1	3	5	3	2	1	4	4	3	3	3	2	2
Duration (s) ($\widehat{w}_i$)	30	60	25	30	29	45	59	30	30	46	41	30	30	30
Value (v_i)	9	9	9	9	9	14	17	9	9	14	14	9	9	9

Again, we will bound the uncertainty by assuming that there might be a delay of 10% ($\widehat{e}_i = 0.1$) over the estimated durations of the advertisements. In contrast to the problem evaluated in Section 5.2, we consider that all estimates have the same certainty ($p(s_i \leq \widehat{w}_i) = 0.5, \forall i$). However, the demanded certainties vary according to the value associated with the commercial, since, as previously mentioned, it makes sense to make the certainty bounds greater for items with greater value. For this instance we use: $\beta_i = 0.6$ for $v_i = 9$, $\beta_i = 0.8$ for $v_i = 14$ and $\beta_i = 0.9$ for $v_i = 17$. We implemented the model with the Numberjack modeling package and solved it with Mistral2 solver. The experiments were run on a 2.3 GHz Intel Core *i*7 processor. For extrapolating probabilities, we used the increasing uniform cumulative distribution (see Figure 1(a) and Equation 1).

Figure 4(a) shows the maximum durations that the advertisements can undergo. For the deterministic case, the commercials cannot undergo any delay, since only the estimated duration is took into account. However, the approach presented in this paper allows a certain delay for each commercial, which is typically called "buffer" time (with scheduling problems). And as noted, it is desirable that the more valuable items have a greater buffer time. Note that, for this instance, the more valuable items have longer expected durations, and consequently, the solutions found by our approach allow greater delays for these commercials (see items 'A002', 'A006', 'A007', 'A011' and 'A012' in the figure).

The solutions obtained by the deterministic approach and the approach presented in this paper, for a maximum broadcasting time of six minutes ($W = 360$), are shown in Figure 4(b) (adverts in increasing duration order). The deterministic solution was found in 49.10s and the robust solution in 2m 55s. We would like to highlight that the deterministic approach tends to select items with greater duration, while our approach discards some of them because it considers their buffers, which are greater (see item 'A011' in the figure). As expected, there is trade-off between robustness and optimality. For this instance, obtaining this robust solution entails a decrease of 6 units in the value of the deterministic solution. However, if any commercial of the deterministic solution were to undergo a delay, there would be a depreciation of 9, 14 or even 17 units because at least one of the commercials selected could not be broadcast. Considering that in real-life problems of this sort, the value of the items could be much greater (for instance hundreds or thousands

of $), this could entail a dramatic loss of income.



(a) Max. allowed durations of the adverts.

(b) Solutions for $W = 360$s.

Fig. 4. Evaluation of the knapsack problem with uncertainty.

7. Conclusions and Future Work

In this paper we present a strategy for dealing with real-life combinatorial optimization problems (specially useful for LSCO problems) in which some of the data is uncertain. Since information about uncertainty is limited, our approach extrapolates data about changes over the original formulation of the uncertain parameters, based on the order of the elements of the uncertain domains. Although we could have used other stochastic methods, in this work our approach was combined with chance constraints to obtain solutions that have a high likelihood of remaining valid in spite 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 an advantage over non-stochastic models in that in combination with chance constraints, it is able to spread robustness more uniformly across all uncertain parameters, because for each one, a minimum specific robustness bound can be fixed. In addition, it does not requires the extra high cost of computing a large number of scenarios. 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 a type of real-life LSCO problem 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 also applied our approach to a knapsack real-life problem proposed in the literature. By means of the CSOP model designed for knapsack problems with uncertainty, we obtained a solution that can handle greater weights/durations in the items collected and also maximizes their total value. In addition, our approach allows us to fix different minimum certainty bounds to the items. We can therefore handle the well-known trade-off between robustness and optimality by means of minimum robustness bounds.

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)*, 10(1):3, 2009.
2. A. Charnes and W. W. Cooper. Chance-constrained programming. *Management science*, 6(1):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)*, 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*, 49:49–78, 2014.
5. J. Hooker. *Integrated methods for optimization*. Springer, 2007.
6. K. Ghedira. *Constraint Satisfaction Problems: CSP Formalisms and Techniques*. John Wiley & Sons, 2013.
7. L. Climent. *Robustness and Stability in Dynamic Constraint Satisfaction Problems*. PhD thesis, Universitat Politécnica de Valéncia, 2013. *https://riunet.upv.es/handle/10251/34785*.
8. A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.
9. G. Verfaillie and N. Jussien. Constraint solving in uncertain and dynamic environments: A survey. *Constraints*, 10(3):253–281, 2005.
10. A. L. Soyster. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research*, 21(5):1154–1157, 1973.
11. 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)*, pages 175–180, 1996.
12. S. A. Tarim, S. Manandhar and T. Walsh. Stochastic constraint programming: A scenario-based approach. *Constraints*, 11(1):53–80, 2006.
13. D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.
14. J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer, 2011.
15. S. Yang, Y. S. Ong and Y. Jin. *Evolutionary computation in dynamic and uncertain environments* Springer Science & Business Media, 2007.
16. T. Walsh. Stochastic constraint programming. In *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI-02)*, pages 111–115, 2002.
17. G. Mathews. On the partition of numbers. *Proceedings of the London Mathematical Society*, 1(1):486–490, 1896.
18. O. Peasah, S. Amponsah, and D. Asamoah. Knapsack problem: A case study of garden city radio (GCR), kumasi, ghana. *African Journal of Mathematics and Computer Science Research*, 4(4):170–176, 2011.