Generating Scenario Trees for Hydro Inflows

Begoña Vitoriano, Santiago Cerisola, Andrés Ramos
Instituto de Investigación Tecnológica
Universidad Pontificia Comillas
c/ Alberto Aguilera, 23
28015 – Madrid SPAIN
Tel: +34 91 542 28 00 Fax: +34 91 541 11 32
e-mail: bvitoriano@doi.icai.upco.es

Abstract: The operation planning of a hydroelectric system, requires some random variables to be represented in a form suitable for quantitative models. Stochastic programming models require a scenario tree to be developed. In this paper a procedure to generate a scenario tree representing the time evolution of natural inflows in a hydroelectric system is presented. The procedure involves non-linear programming, linear regression techniques and deviation variables.

Keywords: Stochastic programming, scenario tree, linear regression, non-linear programming, deviation variables.

I. INTRODUCTION

In decision making under uncertainty it is essential to represent uncertainties in a form suitable for a model. In the operation planning of a hydroelectric system, stochastic programming models are usually used. A good multistage scenario tree, modelling the evolution of natural inflows, is essential to obtain valid results.

These inflows are considered random variables supposed to follow unknown continuous distributions, and one scenario of the tree represents all the random variables in every time period.

The aim of this paper is to present a procedure to obtain discrete distribution approximations of the continuous distributions, obtaining a multistage scenario tree representing the uncertainty about these variables. This procedure involves a non-linear programming model and we also present an alternative non-linear goal programming model with a linear objective function.

The scenario tree will be defined for a hydro system, where inflows are measured in different measuring points corresponding to the same or different basins, hence a multivariate tree must be defined.

In section II, the basic underlying ideas when a scenario tree is used are presented, and the next sections are devoted to develop the procedure to obtain the scenario tree for natural inflows.

This procedure is divided in two phases: univariate and multivariate. The optimisation models required for obtaining an univariate tree, and the moments estimation method and inter-period statistical analysis to develop it, are presented in section III and IV, respectively.

Section V focuses on obtaining the multivariate tree from a univariate by means of spatial statistical analysis.

Finally, an example to clarify the procedure and show the different trees obtained with the two proposed optimisation models, is compiled in section VI.

II. APPROXIMATING CONTINUOUS DISTRIBUTIONS BY DISCRETE DISTRIBUTIONS

When a continuous distribution is to be represented by a discrete approximation, it is possible to resort to sampling. In order to make sure that the distribution properties of the sample are close to those of the underlying distribution, the number of outcomes has to be large. However, stochastic programming models require scenario trees with limited outcomes. Therefore, there is a need of generating the outcomes in a different way, approximating by a discrete distribution with a prefixed number of possible values.

The standard approach for approximating a continuous distribution by a discrete distribution is the following: 1) Divide the outcome region into intervals, 2) select a representing point in each interval, and 3) assign a probability to each point. An example of such approach is the bracket median method described in [1], but in [2] it was pointed out that this method and similar approaches systematically underestimate the moments of the original distributions.

Different authors have developed scenario generation systems in different contexts ([3], [4], [5], [6] and [7]). In this paper, we develop a method to generate a limited number of outcomes for inflows with statistical properties close or equal to those of an underlying multivariate distribution over multiple stages. The method used was defined in [8], there the basic idea is to minimise the square distance between the statistical properties of the discrete approximation and the underlying distribution.

Following this method, we present a procedure to generate a scenario tree for natural hydro inflows. We use different measuring points, so a tree for every point could be defined. Nevertheless, randomness is not well represented by independent trees because there is high correlation between the inflows in different points. So multivariate trees will be defined and obtained.

III. GENERATING UNIVARIATE TREES: OPTIMISATION MODELS

Although we focus on multivariate trees, we present how to obtain an univariate tree because the multivariate will be obtained from this one.

Let *NT* be the number of stages and N_t the number of (conditional) outcomes in stage *t* desired. We assume initially a symmetrical tree, meaning that the number of branches is the same for all conditional distributions in the same stage. Let *S* be the set of all specified statistical properties and S_{VALi} be the specified value of statistical property *i* in *S*.

Define *x* to be the outcome vector and *p* to be the probability vector. Let $f_i(x,p)$ be the mathematical expression for statistical property *i* in *S*. Finally, let w_i be the assigned weight for statistical property *i* in *S*.

We want to construct *x* and *p* so that the statistical properties of the approximating distribution match as well as possible the specified statistical properties of the continuous distribution. So the general description of the model is as follows:

$$\begin{aligned} \min \quad & \sum_{i\in S} w_i(f_i(x,p)-S_{VALi})^2 \\ & \sum p_j = 1 \\ & p \geq 0 \end{aligned} \qquad (1)$$

In general, the resulting non-linear optimisation problem is not convex. For this reason the solution might be (and probably is) a local solution. Nevertheless, for our proposes it is satisfactory to have a solution with distribution properties equal to or close to the specifications, so an objective value equal or close to zero indicates that the distribution of scenarios has a perfect or good match with the specifications.

Another alternative model is proposed with a linear objective function and non-linear constraints, using goal programming and deviation variables:

$$\begin{aligned} \min \quad & \sum_{i\in S} w_i(n_i+d_i) \\ & f_i(x,p)-S_{VALi}+n_i-d_i=0 \qquad \forall i\in S \\ & \sum p_j = 1 \\ & p\geq 0, n_i, d_i \geq 0 \quad \forall i \in S \end{aligned}$$

(2)

The resulting problem is also non-convex, so a local optimum might be reached again. Nevertheless, non-linearity is weaker in this model.

The statistical properties considered for our specific problem are the four central moments of first four orders: expected value, variance, third moment (associated with skewness) and fourth moment (associated with kurtosis). Also the worst case event may be considered but we have not included it in this study.

The moments are different for the first stage and for the following, because the moments in the second and each successive stage must be conditional moments. So the optimisation model is solved by stages, with different problems for first and successive stages.

First stage:

Let M_{11}, M_{12}, M_{13} and M_{14} be the specific (estimated) values of the continuous distribution for this stage (the moment estimation method may be seen in section IV). . Let N_1 be the number of outcomes desired for this stage. Define $x_1 = (x_1(1),...,x_1(N_1))$ as the outcome vector and $p_1=(p_1(1),...,p_1(N_1))$ as the probability vector.

The mathematical expressions for the central moments of the discrete distribution are:

- Average: $m'_{11} = x_1\, p_1 = \sum_{j=1}^{N_1} x_1(j)p_1(j)$
- Second, third and fourth moments (*k*= 2,3,4):

$$m'_{1k} = (x_1 - m'_{11})^k\, p_1 = \sum_{j=1}^{N_1}(x_1(j)-m'_{11})^k\, p_1(j)$$

The resulting quadratic optimisation model is:

$$\begin{aligned} \min \quad & \sum_{k=1}^{4} w_k(m'_{1k}-M_{1k})^2 \\ & \sum_{j=1}^{N_1} p_1(j) = 1 \\ & m'_{11} = \sum_{j=1}^{N_1} x_1(j)p_1(j) \\ & m'_{1k} = \sum_{j=1}^{N_1}(x_1(j)-m'_{11})^k\, p_1(j) \\ & \qquad\qquad k=2,3,4 \\ & p_1 \geq 0 \end{aligned}$$

(3)

In order to avoid unit effects, the weights for the different moments are defined as $w_k = \frac{w'_k}{M_{1k}{}^2}$, where w'_k represents the relevance given to the moment of order *k*.

To develop the goal programming model new deviation variables must be defined. Let n_{1k} and d_{1k} be the positive and negative (surplus and slack) deviation variables for moment *k*, *k*=1,2,3,4, defined as usual in goal programming.

So, the alternative goal programming problem is:

$$
\begin{aligned}
\min \quad & \sum_{k=1}^{4} w_k (n_{1k} + d_{1k}) \\
& \sum_{j=1}^{N_1} p_1(j) = 1 \\
& m'_{11} = \sum_{j=1}^{N_1} x_1(j)\, p_1(j) \\
& m'_{1k} = \sum_{j=1}^{N_1} (x_1(j) - m'_{11})^k\, p_1(j) \\
& \qquad k = 2,3,4 \\
& m'_{1k} - M_{1k} + n_{1k} - d_{1k} = 0 \qquad k = 1,2,3,4 \\
& p_1 \geq 0 \qquad n_{1k}, d_{1k} \geq 0 \quad k = 1,2,3,4
\end{aligned}
\tag{4}
$$

In this case, the weights for the different moments are defined as $w_k = \dfrac{w'_k}{|M_{1k}|}$, with the same meaning of w'_k that in the first model.

Second and successive stages:

These stages are different because there is more than one distribution, there are N_{t-1} different conditional distributions. The model will be presented for the second stage.

Let N_2 be the number of outcomes desired for stage 2, and define $x_2 = (x_2(1), \ldots, x_2(N_2))$ to be the outcomes for any conditional distribution, because the stochastic decision model assumes the same outcomes with different probabilities. We also define

$$p_{2h} = (p_{2h}(1), \ldots, p_{2h}(N_2)), \qquad h = 1, \ldots, N_1$$

as the conditional probability vectors associated to each one of the outcomes obtained in the previous stage (see figure 1).

The moments considered are also the four central moments, but they are moments conditioned to previous values because of the inter-period dependence. Next we consider the conditional expected value and variance, but this is not considered for the third and fourth moment. In the following section an estimation method for conditional moments is explained.

Because the stochastic model assumes the same outcomes in the different conditional distributions, only one optimisation problem has to be posed for all the conditional distributions in the stage.

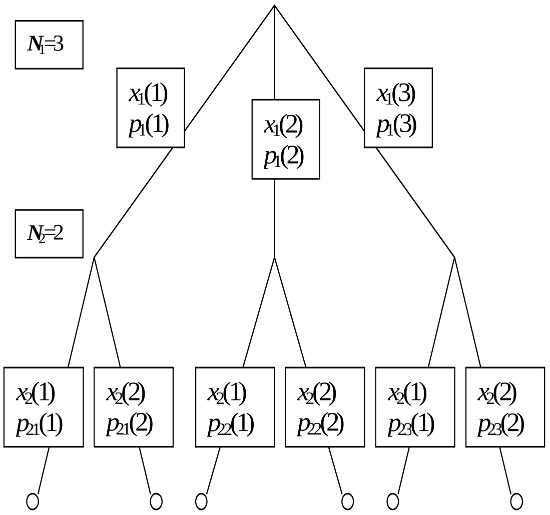

Fig.1. Scenario Tree

Let M_{2h1}, M_{2h2}, M_{2h3} and M_{2h4} be the four conditional specific (estimated) values for the stage 2 supposed the h-th scenario in stage 1 (h=1,...,N_1) (see section IV for the estimation method).

The quadratic optimisation model for second stage is the following:

$$
\begin{aligned}
\min \quad & \sum_{h=1}^{N_1} \sum_{k=1}^{4} w_{hk} (m'_{2hk} - M_{2hk})^2 \\
& \sum_{j=1}^{N_2} p_{2h}(j) = 1 \qquad h = 1, \ldots, N_1 \\
& m'_{2h1} = \sum_{j=1}^{N_2} x_2(j)\, p_{2h}(j) \qquad h = 1, \ldots, N_1 \\
& m'_{2hk} = \sum_{j=1}^{N_2} (x_2(j) - m'_{2h1})^k\, p_{2h}(j) \\
& \qquad h = 1, \ldots, N_1 \qquad k = 2,3,4 \\
& p_{2h} \geq 0 \qquad h = 1, \ldots, N_1
\end{aligned}
\tag{5}
$$

The alternative goal programming problem, with the usual definition of positive and negative deviation variables for every distribution and every moment considered, is:

$$
\begin{aligned}
\min \quad & \sum_{h=1}^{N_1}\sum_{k=1}^{4} w_{hk}(n_{2hk}+d_{2hk}) \\
& \sum_{j=1}^{N_2} p_{2h}(j)=1 \qquad h=1,\dots,N_1 \\
& m'_{2h1}=\sum_{j=1}^{N_2} x_2(j)\,p_{2h}(j) \qquad h=1,\dots,N_1 \\
& m'_{2hk}=\sum_{j=1}^{N_2} (x_2(j)-m'_{2h1})^k\,p_{2h}(j) \\
& \qquad h=1,\dots,N_1 \qquad k=2,3,4 \\
& m'_{2hk}-M_{2hk}+n_{2hk}-d_{2hk}=0 \\
& \qquad h=1,\dots,N_1 \qquad k=1,2,3,4 \\
& p_{2h}\ge 0,\, n_{2hk},\, d_{2hk}\ge 0 \quad h=1,\dots,N_1 \\
& \qquad k=1,2,3,4
\end{aligned} \tag{6}
$$

In order to avoid unit effects and measure the relevance of every scenario (every conditional distribution), the moments weights are defined as $w_{hk}=p_1(h)\dfrac{w'_{hk}}{M_{2k}{}^2}$, in the quadratic model, and $w_{hk}=p_1(h)\dfrac{w'_{hk}}{|M_{2k}|}$ in the goal programming model.

Similarly, for successive stages, one problem is solved for each stage, taking into account previous values in the tree.

IV. GENERATING UNIVARIATE TREES: CONDITIONAL MOMENT ESTIMATION AND TEMPORAL STATISTICAL ANALYSIS

In this section we develop the moment estimation procedure when the structure of the tree is previously defined. We focus on the estimation of the moments of the conditional distributions.

The past history induces different conditional probabilities and different values of specified moments. A question to pose is how long does the past influence, i.e., how many different previous stages (conditional discrete distributions) must be considered in each stage? This question connects with moment estimation and previous statistical analysis will provide the answer.

Due to the tree structure, we divide the estimation problem in three different problems: the first stage, the second one, and the problem for third and later stages.

For the estimation procedure, we have daily time series for around thirty years. Let $\{Y_t\}_{t\in T}$ be the historical time series of inflows in a measuring point.

First stage: In this case, no previous history must be considered, because there is not past information in the tree. So, moments are estimated from the historical data, selecting the observations corresponding to this stage, by classical methods. Let T_1 be the index set corresponding to observations in first stage and $\{Y_t\}_{t\in T_1}$ these collected data.

- Average: $M_1=\dfrac{1}{|T_1|}\sum_{i=1}^{|T_1|} Y_i$
- Central moments:

$$
M_k=\frac{1}{|T_1|-k+1}\sum_{i=1}^{|T_1|}(Y_i-M_1)^k \qquad k=2,3,4. \tag{7}
$$

Second stage: In this case, the distributions considered must be conditional distributions of previous values, i.e., conditioned to the first stage value in the scenario to generate because of time dependence. Nevertheless, we have considered no time dependence in the third and fourth moments (like in [8] and after expert judgement), so they are estimated in the same way as in the first stage.

For the conditional average, simple linear regression is used to estimate it because of the high value of first autocorrelation in the time series. Let $\{Y_t\}_{t\in T_2}$ be the data corresponding to second stage, then the linear regression line of $\{Y_t\}_{t\in T_2}$ over $\{Y_t\}_{t\in T_1}$ is obtained. So, the expected value for each one of the N_1 conditional distributions, M_{2h1}, h=1,...,$N_{1,}$ is estimated from the regression line taking the independent variable the value of the outcome in first stage, i.e., $x_1(h)$.

For the conditional variance, data of the first stage are classified in increasing order and divided into N_1 subsets whose sizes refer to the probabilities obtained for that stage, $p_1(h)$. For each of these subsets, data of the second stage are identified by year. They are then sequenced in the time series and classified in twin subsets (see figure 2). Finally, the variance of these twin sets is estimated.

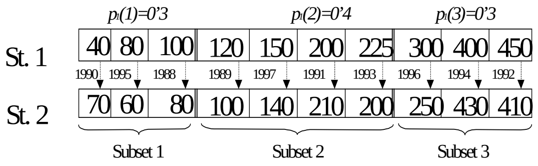

Fig.2. Subsets to estimate conditional variance in stage 2

Third and later stages: We shall start by justifying why we consider the third stage and later stages similar. This is an empirical conclusion obtained after statistical analysis.

We have used daily time series for around thirty years of some inflows in some different measuring points of Spanish basin rivers. Every one of these sequences has been studied by time series analysis.

Furthermore, we were interested in obtaining conclusions independent of period lengths. So, this statistical study has been repeated with weekly and monthly time series, obtaining similar conclusions with each of these analyses.

The Box-Jenkins method has been applied to the different time series and ARIMA models with seasonal factors obtained. In every model, it may be observed that the influence of third and prior periods of the current period in the same year is not relevant when values of the two last periods are included.

So, the statistical analysis conclusion is that the only information that needs to be considered to obtain conditional distributions is that contained in the last two stages before the current one. And this conclusion implies that the moment estimation method will be equal for the third and for later stages.

Analogously to the second stage, third and fourth moments will be supposed independent of previous values and will be estimated by equation (7). Conditional variance will be estimated using the same procedure described in second stage.

Hence, the only difference in the process estimation with the second stage is found in the conditional average estimation. The expected value conditioned to past history in the tree will be obtained by multiple linear regression. The regression line used is such that two independent variables are the values in the two last stages before the current stage. The outcomes obtained in the scenario for these last two stages will be the input values of these variables to estimate the conditional expected value in the current scenario and current stage.

V. GENERATING MULTIVARIATE TREES: SPATIAL STATISTICAL ANALYSIS

The final aim of this paper is to provide a procedure to obtain a scenario tree for a stochastic programming model for hydrotermal coordination. The uncertainty is modelled for different measuring points, so the model works over a multivariate tree.

A spatial statistical analysis is needed in order to establish the relationship between the inflow's behaviour in the different measuring points of the same and different basins.

We started with a cluster analysis including all of the time series. The analysis result, how it was expected, shows a similar behaviour into the measuring points from the same basin, obtaining a different cluster for every one. Hence, different scenario trees will be generated, one tree for each basin.

Furthermore, the correlation into each cluster is very high, so the following procedure is applied.

Firstly, one point in each basin is selected (the measuring point with higher correlation with the other points in the same basin).

Afterwards, one univariate tree is generated for this representing point, following the procedure explained in the previous sections.

Finally, the multivariate tree is completed by simple linear regression, i.e., outcomes for each of the other basin's points are obtained by its regression line over the representing point, taking the independent variable the value of the corresponding outcome.

VI. EXAMPLE

An example for an univariate tree representing inflows in a Spanish basin is presented to clarify the procedure and realise the differences obtained with both models ((1) and (2)).

In this example, the tree has been generated for three stages (months) to show the results in a suitable way. Actually, yearly trees for twelve stages are being used in the true coordination hydrotermal model.

Data to develop the scenario tree correspond to 28 years and one measuring point. The weights assigned to the different moments (w'_i) are 2, 1, 0.5 and 0.25 for first, second, third and fourth moments, respectively. These proposed values show the relevance considered for the moments, i.e., decreasing importance with the order of the moment.

Firstly, we define the tree structure by mean of the number of outcomes desired in each stage. Let N_1=3, N_2=3 and N_3=2. This structure with more outcomes in first periods is selected, because it's usual to consider more important closer stages than later stages, hence a more exhaustive description is desired at the beginning of the tree.

The outcomes obtained with the quadratic model are presented in the next table:

TABLE I. OUTCOMES MODEL (1)

Stage 1	1902	776	139
Stage 2	2277	991	368
Stage 3	1049	295	

Table II shows the probabilities obtained in every stage with this model (Stages 1 and 2 must be read in horizontal way, conditional distributions of stage 3 must be read in vertical way).

TABLE II. PROBABILITIES MODEL (1)

St1	.04			.22			.74		
St2	.11	.89		.09	.41	.50	.05	.16	.79
St3	.98	.08		.98	.08	.06	1	.11	.07
	.02	.92		.02	.92	.94		.89	.93

For the goal programming model, the results obtained are presented in tables III and IV. These tables must be read in the same way that previously for the first model.

TABLE III. OUTCOMES MODEL (2)

Stage 1	1898	596	58
Stage 2	2277	1052	402
Stage 3	957	195	

TABLE IV. PROBABILITIES MODEL (2)

St1	.04			.39			.57		
St2		1		.06	.32	.62	.07		.93
St3		.20		1	.25	.12	1		.16
		.80			.75	.88			.84

The obtained scenario trees are enough similar with small differences. Both models achieve a perfect adjustment to the moments in the first stage, but not in the next stages. This gap occurs because the model decision requires the same outcomes for all the conditional distributions. Currently, we are modifying the decision model to allow different outcomes in order to achieve a better adjustment of the scenario tree.

VII. CONCLUSIONS

The presented procedure to obtain multivariate scenario trees for inflows in a hydro system allows develop stochastic programming models for hydrotermal co-ordination. The multivariate scenario tree obtained approximates the unknown continuous distributions of natural inflows via their moments, so statistical properties of the discrete approximations and the underlying continuous distributions are similar. The procedure is being applied to validate it to some Spanish hydro subsystems with suitable results.

VIII. REFERENCES

[1] R. C. Clemen, Making Hard Decisions: An Introduction to Decision Analysis. PWS-Kent Publishing Co., Boston, 1991

[2] A .C. Miller and T. R. Rice, "Discrete Approximations of Probability Distributions," Management Science, 29, 1983, pp. 352-362

[3] D. R. Carino, T. Kent, D. H. Myers, S. Stacy, M. Sylvanus, A. L. Turner, K. Watanabe and W.T. Ziemba, " The Russel-Yasuda Kasai Model: An Asset/Liability Model for a Japanese Insurance Company Using Multistage Stochastic Programming," Interfaces, 24, 1994, pp. 29-49.

[4] J. M. Mulvey, "Generating Scenarios for the Towers Perrin Investment System," Interfaces, 26, 1996, pp1-13.

[5] G. Consigli and M. A. H. Dempster, "Dynamic Stochastic Programming for Asset/Liability Management," Working paper 04/96 of Finance Research Group, Judge Institute of Management Studies, University of Cambridge, UK, 1996

[6] Z. Chen, G. Consigli, M. A. H. Dempster and N. Hicks-Pedrón, "Towards Sequential Sampling Algorithms for Dynamic Portfolio Management," Working paper 01/97 of Finance Research Group, Judge Institute of Management Studies, University of Cambridge, UK, 1997

[7] S. Zenios, "Asset/Liability Management under Uncertainty for Fixed-income securities," Annals of Operations Research, 59, 1995, pp. 77-97.

[8] K. Hoyland and S.W. Wallace, "Generating Scenario Trees for Multistage Problems," Working paper of Department of Economics and Technology Management, Norwegian University of Science and Technology, Norway, 1998.