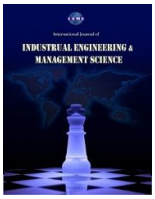
University of Hormozgan

International Journal of Industrial Engineering & Management Science

journal homepage: www.ijiems.com



Using Trended Regression Trees for Computing Unbiased Makespan of PERT Networks

Mohamad Ali Movafaghpour [a]

[a] *Industrial Engineering Department, Jundi-Shapur University of Technology, Dezful, Iran*

ARTICLE INFO

Article history:
Received: 2021-11-10
Received in revised form: 2022-04-02
Accepted: 2021-06-01

Keywords:
Project Planning
Stochastic Activity Duration
Stochastic Longest Path
Machin Learning
Trended Regression Tree

ABSTRACT

Scheduling a network of activities is a common task in project management. This paper deals with estimating the Project Completion Time in stochastic networks. Traditionally, one of the well-known tools to schedule projects consisting of activities with stochastic activity durations is Project Evaluation and Review Technique (PERT). Since PERT is known to be over-optimistic, practitioners prefer to use simulation-based methods such as Monte Carlo simulation. PERT assumes rough estimations for the mean and variance of the merge events. The completion time of merge events follows the rules of an important order stochastic known as sample maximum. We review analytic bounds and estimators for the mean and variance of sample maximum/minimum. Since no one of the currently available bounds and estimators is fit for the general case, first we develop two estimates for the mean value of sample max/min and two upper bounds for the variance of sample max/min. Then we developed a novel method named as Trended Regression Trees (TRT) to find more accurate estimates for mean and variance of sample maximum/minimum. Having more precise estimates for merge events makes it possible to estimate the project completion time with higher accuracy. Computational results are presented that confirm significant accuracy improvement in estimating the mean and variance of project completion time. We examined huge cross-validation to find the most reliable Trended Regression Trees.

1. Introduction

In project management, a project consists of a set of activities and a set of binary relations expressing which tasks cannot start before which preceding tasks are completed. We can illustrate this set of tasks and precedence relations as a network. This network is a directed acyclic graph. With such an activities network the problem is to find critical activities, and determining the optimal starting times of activities, to minimize the project makespan.

In spite of projects with deterministic durations, in many real-world projects, activity durations can be represented by stochastic numbers, where the duration of some or all of the activities are modeled as probabilistic numbers with known central tendency and dispersion indices or even stochastic distributions. Traditionally, the shortest path or Project Completion Time (PCT) in such a probabilistic network is found with the Project Evaluation and Review Technique (PERT). PERT provides inaccurate information about project makespan. Usually, this inaccuracy is significant enough to make such estimates as not applicable. As a result of this systematic error, many improvements since the introduction of PERT in 1959 by Malcolm et al. (1959) have been developed. However, in spite of its biased results and the many improvements, PERT is presented in most textbooks on Project Management. This is due, perhaps, to its simple and informative application.

To be exact, PERT is observed to be optimistic; i.e. most of the procedures designed for analysis purpose like as the PERT resulted in an estimation of the earliest completion time that is both optimistic and biased (Pontrandolfo, 2000). This situation arises from the effect called “merge event bias”. In general, the maximum of a set of random variables is a new random variable different from each of them and therefore it has its mean and variance different from each of them.

The results generated by PERT for mean value of PCT are almost known to underestimate the values observed in action. One of the most important causes of this problem usually is reported as merge event bias. In this paper, we review the recent findings in statistics for modeling the mean value and variance of the merge event and propose novel formulas for mean value and variance of the event merging non-identically distributed activities. Simulations made to compare the results generated by our

* Corresponding author.
E-mail address: movafaghpour@jsu.ac.ir

developed algorithm versus PERT show that our proposed method can provide more effective estimates for mean value and variance of merge events and consequently PCT.

Other than planning start/finish time of activities, which is our focus problem in this paper, there are other problems regarding project planning for example considering constrained capacity for project resources, time-cost tradeoffs, and etc.. In this paper with no lack of generality for using the findings of this paper, we relax the capacity constraints and activity time-cost tradeoffs.

In the following sections first we review the techniques for finding the shortest path in probabilistic networks; and their shortcomings are discussed in section 2. Some revisions on classic PERT are discussed in section 3. Section 4 reviews some bounds on the mean and variance of a merge event. In section 5 we present our calculated bounds on such statistic and their efficiency is evaluated in section 6.

2. Literature Review

Critical Path Method (CPM) first was proposed by Kelley & Walker (1959) at the DuPont Corporation. Then Malcolm et al. (1959) developed PERT at less than one year later in General Dynamics and the US Navy. Zhen-ting et al. (2010) argue that CPM and PERT are two primary well-known tools and are frequently used for project scheduling. Although we know that most activity durations are stochastic in nature, in contrast with CPM that assumes deterministic activity durations. PERT assumes that each activity duration follows a random distribution.

Enumerating all the potential paths connecting the start node to the end node in the project network and then finding the most probable CP is another idea which was proposed by Chu et al. (2014).

Yao & Chu (2007) attempt to search for discretized probability distribution function for project makespan through enumeration all the potential paths.

Besides stochastic activity durations, resource levels may also be available at stochastic amounts, Chen et al. (2018) and Rostami et al. (2018) considered such problem.

Bordley et al. (2019) investigate the problem of uncertain project deadlines and propose an algorithm called modified-PERT for crashing activities. Sometimes two types of uncertainty are considered, e.g. Golenko-Ginzburg and Gonik (1998) focused on a problem in which each activity is of random duration depending on the resource amounts assigned to that activity; the resources consumption rate is linearly correlated with the resource amount plus a random offset.

PERT enables the project managers and planners in diverse fields to estimate the probability of meeting specified completion dates for example Nekoufar and Movafaghpour (2017) reported a PERT implementation for university curriculum planning. A novel approach to estimate project completion time is using machine learning algorithms for improving the prediction accuracy instead of using statistics. For example Yan and Shang (2019) used relative entropy kernel regression (REKR) and Gaussian margin machines (GMM) for precise forecast values for design time. Also Hajiali, Mosavi, and Shahanaghi (2020) developed a model based on ensemble learning using certain specifications of projects to estimate the project completion time. They reported higher reliability and better robustness in estimating project span. In this paper, we focus on stochastic activity durations and develop a less biased estimate for the project makespan.

3. PERT

Malcolm et al. (1959) suggested the classical PERT mean and variance based on the given expert estimated lower and upper bounds a, and b and most likely value m:

$$E[X]=\frac{a+4m+b}{6} \tag{1}$$

$$V[X]=\left(\frac{b-a}{6}\right)^2. \tag{2}$$

In a PERT activity network, durations of activities are assumed to be independent of each other and follow a distribution called Beta distribution (Hajdu & Bokor, 2016)) following the Probability Density Function (PDF) as below:

$$f(x)=\begin{cases}\dfrac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}x^{p-1}(1-x)^{q-1} & if\ 0\le x\le 1,\\ 0 & otherwise.\end{cases} \tag{3}$$

Both p and q must be positive real numbers. For the mean value of such random variable X we have:

$$E[X]=\left(\frac{p}{p+q}\right), Var[X]=\left(\frac{pq}{(p+q)^2(p+q+1)}\right) \tag{4}$$

Classic PERT assumes a modified Beta distribution on activity durations. Such that each activity duration x_i has a mode and is confined between a pair of bounds:

$$a\le x_i\le b,\ \text{Mode}(x_i)=m. \tag{5}$$

Therefore PDF (3) is revised as:

$$f(x)=\begin{cases}\dfrac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\dfrac{(x-a)^{p-1}(b-x)^{q-1}}{(b-a)^{(p+q)-1}} & if\ a\le x\le b,\\ 0 & otherwise.\end{cases} \tag{6}$$

Where both p and q be are real numbers. We can compute the mean, variance and distribution mode of modified Beta distribution using its probability distribution function as:

$$E[X]=a+(b-a)\left(\frac{p}{p+q}\right)=\frac{pb+qa}{p+q}=\frac{a+(p+q-2)m+b}{p+q} \tag{7}$$

$$Var[X] = (b-a)^2\left(\frac{pq}{(p+q)^2(p+q+1)}\right) \quad (8)$$

Although Malcolm et al. (1959) assumed beta distribution for the activity durations. This was criticized by many researchers, and several new distribution types were introduced. e.g. Lee et al. (2013) after analyzing two sample activity networks concluded that if one assumes that Project Completion Time (PCT)s are normally distributed, PERT may lead to an approximately 10% to 30% more optimistic result. They recommended that PCT is not always normally distributed and its distribution and variability are affected by the distribution and variability of activity durations. Their best fitted distribution for most activity distribution was found to be normal and the difference in completion probabilities between Normal PDF and best fit (known to be Lognormal) PDF for Weibull, Lognormal, Exponential, was 3%, 8% and 11% respectively.

Finally Hajdu & Bokor (2016) after testing several distributions for activity durations concluded that instead of selecting the proper activity duration distributions, project experts have to devote more effort to precisely determining the activity durations. Because from a practical point of view, using different activity duration distributions does not result in significant differences. They found that the precision of the three-point estimation plays a much more important role in determining the project makespan.

Burgelman & Vanhoucke (2019) focused on Markovian PERT networks in which activity durations are exponentially distributed random variables. Although their assumed activity durations followed the same distribution, but they noticed that no general analytic solution would exist to derive project makespan distributions. They assessed the performance of several methods addressed in the literature both from the theoretical and managerial point of view and introduced a new benchmark dataset to generate Markovian PERT networks.

Hernandez-Bastida & Fernandez-Sanchez (2018) evaluated some extreme distributions with different assumed probability distribution functions satisfying position matching with three estimates (a, b, and m) given by experts. They showed that if the additional information is incorporated (in a form that expert indicates the probability of an interval), estimates for mean and variance differ very significantly. And finally, they concluded that no additional assumption could be incorporated into the distribution of activities unless it is fully justified.

Despite its ease of use arose from its long history, the PERT method is still an active research field; therefore the original PERT mean and variance (1-2) are still highlighted in popular operations research textbooks (see, e.g., Hillier and Lieberman, 2010; Winston, 2004).

MacCrimmon and Ryavec (1964) analyzed the assumptions used in PERT and highlighted four possible sources of error. Among them, the two most important are: the approximation formula for the mean and the variance. They also showed that the calculated mean for project makespan will always be optimistically an underestimate, but the calculated variance may be biased in either direction.

Kim et al. (2014) recommended using 5th, 10th, 90th, and 95th percentiles instead of common percentiles P0 and P100 used in classical PERT. They evaluated the accuracy of a wide range of PERT mean-variance estimation formulas and proposed new sets of weights for each set of three estimates P5, P50, and P95. The important assumption that they ignored was the way these estimates are gathered. Since the estimates are gathered by reference to experts' opinion; usually it is easier and therefore more accurate to ask experts for estimating the mode of the distribution than the median (i.e. P50 in their notation) than estimating P5, P50, and P95 percentiles.

Hajdu & Bokor (2016) argued that the precision in estimating the three-point plays a much more important role in determining the mean and variance of activity durations. Sackey and Kim (2019) proposed heuristic estimating relations for mean value and variance of activity durations based on three given estimates as:

$$E[X] = 0.025\,a + 0.95\,m + 0.025\,b\ ,\ Var[X] = [(b-a)/6.43]^2 \quad (9)$$

Their proposed expressions had no analytical proof for yielding better results than classic PERT but they conducted a simulation on five case studies and showed that their proposed modified PERT gives a better mean error rate than the classic PERT.

Herrerias-Velasco et al. (2011) provided a more accurate expression for estimating the PERT variance. They developed a method to find parameters of beta distribution in which their estimated beta distribution has the same distribution mode as expressed by experts while the original PERT mean expression was verified. Also, they concluded that, in general, their proposed estimate for mean was very accurate. Then it was verified by (Ballesteros-Perez, 2017).

$$Var[X] = \left(\frac{5}{7} + \frac{16}{7}\times\frac{(b-m)(m-a)}{(b-a)^2}\right)\times\left(\frac{b-a}{6}\right)^2. \quad (10)$$

Ballesteros-Perez (2017) suggest a sequential merging process named as M-PERT which is a consecutively reduction technique to convert a network of activities into a new single merged activity with calculated mean and variance for duration. The most important shortcoming of his proposed procedure is disregarding merge event bias. His proposed formulas for estimating mean and variance for every resulted merge activity lacks the precision for estimating the sample maximum statistic. Merge event bias was observed within the first days of implementing PERT in the U.S. Air Force projects, in which the consistent and systematic errors was evidencing in project forecasts (Ballesteros-Perez, 2017). Elmaghraby (2005) noted that the fallacy of PERT is replacing the random durations by their mean values to determine the project makespan. It is not "practical" to replace the randomness with an equivalent certain value in the form of Expected value if uncertainty is inherently still available in the activity durations.

Despite the simplicity and widespread adoption of the original PERT and its alternative extensions, neglecting the merge event bias, significantly underestimated the duration average and overestimated the duration variance of real-life projects. Therefore, its simulation-based rival which yields more accurate solutions was proposed by van Slyke (1963) only less than five years after PERT was born. Monte Carlo method relays on extensive sampling runs and yields criticality index of each activity, therefore some researchers tried to infer statistical distributions for completion time of project or individual activities, e.g. Milian (2008) introduced the

idea of a positively correlated random variable called as control variable for estimating the probability distribution of the project duration. Using the Monte Carlo method requires software implementations, therefore it has not yet received wider application "on the field" by practitioners and project managers (Avlijas, 2018).

Since uncertainty typically does not allow a treatment based on averages (Elmaghraby, 2005), we propose a novel estimate for mean and variance of merge event.

For instance, consider two activities *A* and *B*. Both of them sink to a merge event denoted as *P*. Suppose that the durations of activities *A* and *B* denoted as T_A and T_B respectively, are a random variable. Therefore, the completion time of both activities *A* and *B* denoted by T_P is random variable. Formally, T_P can be defined by the expression as:

$$T_P = \max\{T_A, T_B\}. \quad (11)$$

It follows that:

$$E[T_P] = E[\max\{T_A, T_B\}]. \quad (12)$$

According to the traditional PERT method E[TP] is approximated as:

$$\kappa = \arg\max\{E[T_A], E[T_B]\}, \quad (13)$$

$$E[T_P] = \mu_\kappa, \quad (14)$$

$$Var[T_P] = \sigma_\kappa^2. \quad (15)$$

This approximation results it well-known shortcomings of PERT in underestimation of $E[T_P]$ and overestimating $\mathrm{Var}[T_P]$. In fact, the gap between the exact value of $E[T_P]$ and the estimated value μ_κ presented by PERT, increases with the number of parallel activities sinking to the merge event (Pontrandolfo, 2000).

In statistics, T_P is known as a statistic called 'sample maximum'. If all T_A, T_B, T_C,… are Identically Independently Distributed (IID) random variables, there are various upper/lower bounds recommended for mean value and variance of T_P by several researchers.

The earliest bounds on sample max are published by Gumbel (1954) and Hartley and David (1954) when considering an arbitrary population mean 0 and variance 1:

$$E[X_{n:n}] \le \frac{n-1}{\sqrt{2n-1}}. \quad (16)$$

If the probability distribution is symmetric about zero, they proposed the following bound:

$$E[X_{n:n}] \le \frac{n}{\sqrt{2(2n-1)}}\sqrt{1-\frac{1}{\binom{2n-2}{n-1}}}. \quad (17)$$

Balakrishnan and Rao (1998) quoted the universal bound for the mean of the sample max from any continuous distribution with mean μ and variance σ^2 as:

$$E[X_{n:n}] \le \mu + \frac{n-1}{\sqrt{2n-1}}\sigma, \quad n = 2,3,\ldots \quad . \quad (18)$$

And in case of continuous symmetric distribution a tighter universal bound as:

$$E[X_{n:n}] \le \mu + \frac{n\sigma}{\sqrt{2(2n-1)}}\sqrt{1-\frac{1}{\binom{2n-2}{n-1}}}, n = 2,3,\ldots \quad (19)$$

Moriguti (1951) recommends the following upper and lower bounds on the variance of the sample max of a symmetric population:

$$\lambda_n\sigma^2 \le Var[X_{n:n}] < \frac{1}{2}n\sigma^2, \lambda_n = \frac{\pi}{2^n}\left[1+O\left(\frac{1}{n}\right)\right] \quad (٢٠)$$

Where σ2 corresponds to variance of the population. He also provides a relation for coefficient of variation of sample max:

$$\frac{\sqrt{V(X_{n:n})}}{E[X_{n:n}]} \approx \frac{1}{2^{\frac{n}{2}-1}} \quad (21)$$

In which $X_{n:n}$ is the sample maximum and n is the sample size.

Papadatos (1995) proved an upper bound for sample max drawn of a set of n iid random variables X_i as:

$$Var[X_{n:n}] \le n\sigma^2 \quad (22)$$

Arnold and Groeneveld (1979) proposed the following bounds on the mean value and variance of the i'th order statistics denoted as $X_{i:n}$.

$$\mu - \sigma\sqrt{\frac{n-i}{i}} \le E[X_{i:n}] \le \mu + \sigma\sqrt{\frac{i-1}{n-i+1}}, 1 \le i \le n. \quad (23)$$

Rychlik (2008) proved the following bound:

$$0 \le Var[X_{i:n}] \le \max\left\{\frac{n}{i}, \frac{n}{n+1-i}\right\}\sigma, 1 \le i \le n.$$

In which X_i is an arbitrary dependent identically distributed random variable with a finite variance σ and $X_{i:n}$ denotes the i'th order statistic of the sample with homogeneous variance and mean value.

The above bound when i = 1 or n is presented for a finite population with size N (Balakrishnan et al., 2003) as:

$$Var[X_{i:n}] \le n\frac{N-n}{N-1}\sigma, i = 1 \text{ or } n. \quad (25)$$

Bertsimas et al. (2006) prooved the optimality of the following tight upper bound on the expected value of the highest order statistic $X_{n:n}$:

$$\sup E[X_{n:n}] = \mu + \sigma\sqrt{n-1}. \quad (26)$$

In which each random variable X_i follows an arbitrary distribution θ with identical mean–variance pair μ and σ^2. If each variable X_i has individual mean–variance pair μ_i and σ^2_i they proposed the following two closed-form upper bounds on the expected value of the highest order statistic:

$$\sup E[X_{n:n}] = \frac{1}{2}\left(\sum_{i=1}^{n}\left[\mu_i + \sqrt{\left(\mu_i - \max_{1\le i\le n}\left\{\mu_i + \frac{n-2}{2\sqrt{n-1}}\sigma_i\right\}\right)^2 + \sigma_i^2}\right] + (2-n)\left[\max_{1\le i\le n}\left\{\mu_i + \frac{n-2}{2\sqrt{n-1}}\sigma_i\right\}\right]\right) \quad (27)$$

$$\sup E[X_{n:n}] = \frac{1}{2}\left(\sum_{i=1}^{n}\left[\mu_i + \sqrt{\left(\mu_i - \min_{1\le i\le n}\left\{\mu_i + \frac{n-2}{2\sqrt{n-1}}\sigma_i\right\}\right)^2 + \sigma_i^2}\right] + (2-n)\left[\min_{1\le i\le n}\left\{\mu_i + \frac{n-2}{2\sqrt{n-1}}\sigma_i\right\}\right]\right) \quad (28)$$

Aven (1985) proved two bounds on the expected maximum (minimum) for non-identically distributed variables:

$$E[X_{n:n}] \le \frac{1}{n}\sum_i E[X_i] + \sqrt{n-1}\times \min_j\left\{\sqrt{\frac{\sum_i Var[X_i]}{n} + Var[X_j] + \frac{1}{n}\sum_i (E[X_i])^2 - \frac{1}{n^2}\left(\sum_i E[X_i]\right)^2}\right\} \quad (29)$$

$$E[X_{n:n}] \le \max_i\{E[X_i]\} + \sqrt{n-1}\min_j\left\{\sqrt{\frac{\sum_i VAR[X_i]}{n} + Var[X_j]}\right\}, \quad (30)$$

$$E[X_{1:n}] \ge \frac{1}{n}\sum_i E[X_i] - \sqrt{n-1}\times \min_j\left\{\sqrt{\frac{\sum_i Var[X_i]}{n} + Var[X_j] + \frac{1}{n}\sum_i (E[X_i])^2 - \frac{1}{n^2}\left(\sum_i E[X_i]\right)^2}\right\} \quad (31)$$

$$E[X_{1:n}] \ge \min_i\{E[X_i]\} - \sqrt{n-1}\min_j\left\{\sqrt{\frac{\sum_i VAR[X_i]}{n} + Var[X_j]}\right\}. \quad (32)$$

Several useful bounds on mean and variance of sample maximum and sample minimum are reviewed by Arnold et al. (2008), Balakrishnan and Rao (1998) and Rychlik (1998). We refer the interested readers to those papers for more detailed results about order statistics.

4. Developed Methods

4.1. New Upper Bounds on Mean and Variance of Sample Max

All the above reviewed papers provide bounds for sample max drawn from a single population; except Bertsimas et al. (2006) who provide bounds on mean value of sample max drawn from different populations (Eq. 18-19). In this section we use the following notation to present novel bounds and estimates of mean and variance of sample max/min.

Table 1. Notation used

Statistic	Point Estimator	Upperbound
$E[X_{n:n}]$	$\hat{E}[X_{n:n}]$	-
$SD[X_{n:n}]$	$\hat{SD}[X_{n:n}]$	$UB_{SD[Xn]}$
$E[X_{1:n}]$	$\hat{E}[X_{1:n}]$	-
$SD[X_{1:n}]$	$\hat{SD}[X_{1:n}]$	$UB_{SD[X1]}$

Based on numerous experiments, we introduce the following upper bounds on variance for sample max/min. Every random variable X_i is assumed to follow an arbitrary distribution with mean and variance, μ_i and σi^2 respectively.

$$UB_{SD[Xn]} = \max_{i\in ActiveSet_U}\{\sigma_i\} \qquad UB_{SD[X1]} = \max_{i\in ActiveSet_L}\{\sigma_i\} \quad (33)$$

Active Set_U is the set of effective variables, i.e. all random variables (following beta, triangular, or uniform distribution and satisfying: $a_i < x_i < b_i$) having their upper limit b_i greater than the lower limit a_i of the random variable with the biggest mean value, and Active Set_L is the set of effective variables; in other words, the set of all random variables having their lower limit a_i smaller than the upper limit b_i of the random variable with the smallest mean value.

We introduce the following point estimators for mean value and variance of sample max and sample min:

$$\hat{E}[X_{mn}] = \max_{i \in ActiveSet_U} \{\mu_i + \sigma_i\} \quad \hat{E}[X_{1n}] = \min_{i \in ActiveSet_L} \{\mu_i - \sigma_i\} \qquad (34)$$

4.2. Regression Tree (RT)

Regression trees preserve the advantages of both regression analysis and decision trees. Regression analysis as a statistical procedure that tries to fit a function to a set of observations in order to model the relationship between a dependent with one (/some) independent variable(s). In tree regression, constructing tree-like structures is inherited from decision trees and therefore all observations are clustered by repeated splits on the instances of each parent node, creating if-then-else rules (Gkioulekas & Papageorgiou, 2021). Such models are computed quickly and simply interpreted. Since RT is recently introduced in operations research as a stochastic data mining tool, some optimization improvements are investigated by researchers to enhance its performance such as Bertsimas et al. (2021).

For example, the regression tree for data points presented in Table 2 is illustrated in Figure 1 which can estimate the output variable y with a MSE less than 17.7.

In order to estimate the mean value and Standard Deviation (SD) of sample maximum/minimum as a dependent variable, a set of features is to be defined based on the mean and variance of the duration of activities sinking into the merge event. Heuristically a set of features is selected as showed in Table 3. The same features would be selected for constructing RT3 and RT4 with some tiny modification by replacing maximum with minimum operator. The random variable with biggest mean is denoted by index κ and the random variable with the second biggest mean is denoted by index η. Among all activities sinking into a merge event, a subset of activities called as 'Active set' or 'effective activities' is selected.

Table 2. Sample data for illustrative regression tree

input		Output
x_1	x_2	Y
2	12	2
4	12	2
6	12	2
8	20	3
10	25	5
12	26	20
16	60	100
18	65	100
20	73	100
22	70	100
24	50	70
26	52	60
28	52	50
32	51	18
34	54	4
36	55	3
38	56	2

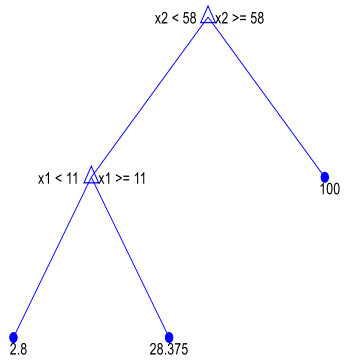


Figure 1. Regression Tree for data shown in Table 2

4.3. Trended Regression Tree (TRT)

Traditional RT's stop with a binary tree that at each of its terminal leaf c, a value of $\overline{y}_c$ represents the average output variable for all the instances assigned to that leaf. This kind of averaging the output variables $y_i, (i \in c)$ may ignore some useful information behind the values of input variables $x_{i,m}, (i \in c)$. Therefore, we prefer to develop a linear regression function at each terminal leaf. Specifically, after the full size RT is produced, we perform a correlation analysis and then a linear independency analysis. We perform a correlation analysis to avoid over fitting regression line, eventually less correlated variables with the output variable are ignored. i.e. the input variables that have an absolute value of correlation coefficient with response variable less than 65th percentile of all correlation coefficients are ignored.

Linear independency analysis is also performed at each leaf to distinguish the independent variable set based on the instances assigned to each leaf. This analysis is done because some leaf may have fewer instances than required for regression. Making a linear independency analysis is done by performing a Gauss-Jordan elimination to find Reduced Row Echelon form. By selecting independent columns, a linear regression model is fitted for each node. Fig.2 depicts a sample Trended Regression Tree (TRT).

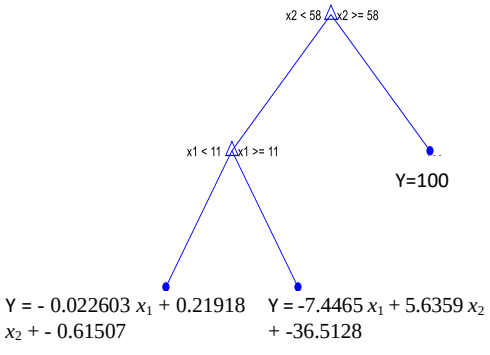


Figure 2. Trended Regression Tree for data shown in Table 2

5. Experimental Results

In order to analyze the efficiency of proposed estimates for mean and variance of project completion time three training datasets of random activities following triangular/beta/uniform distributions with three-point estimates (*a*, *b*, and *m*) are generated; *a* is selected with a random distribution as uniform(80, 100); *b* is selected such that the distance between a and b follows a random uniform distribution between [0, 0.2**a*], and *m* (for triangular and beta distributions) is selected with a uniform distribution between a and b. In order to calculate the observed mean value and Standard Deviation (SD) of sample max/min, 2000000 observations were made for each merge event. Three other validation dataset are generated to test the validity of the results generated by regression trees and trended regression trees.

Four regression trees RT1, RT2, RT3, and RT4 are constructed for estimating mean value of sample maximum, SD of sample maximum, mean value of sample minimum, and SD of sample minimum respectively. Four Trended Regression Trees TRT1, TRT2, TRT3, and TRT4 are also constructed for the same reason. All the experiments are codded in MATLAB R2018a. Table 3 summarizes the notation used for bounds and estimates in the literature compared with proposed ones.

Table 3. Selected features of merging activities for establishing RT 1 and RT 2

Type	Definition	Formula	Symbol	RT 1	RT 2
Input	Number of effective activities merging	n	x_1	●	●
	Gap between two biggest means	$\mu_\kappa - \mu_\eta$	x_2	●	●
	SD Ratio of two variables with biggest means	$\frac{\sigma_\eta}{\sigma_\kappa}$	x_3	●	●
	Relative gap between PERT *Var* & Max *Var* to the range of SD's	$\frac{\max_{i\in ActiveSet_U}\{\sigma_i\}-\sigma_\kappa}{\max_{i\in ActiveSet_U}\{\sigma_i\}-\min_{i\in ActiveSet_U}\{\sigma_i\}}$	x_4	●	●
	Relative gap between two biggest means to the sum of their SD.	$\frac{\mu_\kappa-\mu_\eta}{\sigma_\kappa+\sigma_\eta}$	x_5	●	●
	SD of the activity with the biggest mean	σ_κ	x_6	●	●
	SD of the activity with the second biggest mean	σ_η	x_7	●	●
	Ratio of SD of the activity with the second biggest mean to gap between two biggest means	$\frac{\sigma_\eta}{\mu_\kappa-\mu_\eta}$	x_8	●	●
	gap between two biggest means plus the sum of their SD.	$\mu_\kappa-\mu_\eta+\sigma_\kappa+\sigma_\eta$	x_9	●	●
	gap between two biggest means plus the gap of their SD.	$\mu_\kappa-\mu_\eta+\sigma_\kappa-\sigma_\eta$	x_{10}	●	●
	Average of means of effective activities	$\bar{\mu}=\frac{1}{n}\sum_{i\in ActiveSet_U}\mu_i$	x_{11}	●	●
	SD of means of effective activities	$\sqrt{\frac{1}{n}\sum_{i\in ActiveSet_U}\left(\mu_i-\bar{\mu}\right)^2}$	x_{12}	●	●
	Range of means of effective activities	$\max_{i\in ActiveSet_U}\{\mu_i\}-\min_{i\in ActiveSet_U}\{\mu_i\}$	x_{13}	●	●
	Mean of SD's of effective activities	$\bar{\sigma}=\frac{1}{n}\sum_{i\in ActiveSet_U}\sigma_i$	x_{14}	●	●
	SD of SD's of effective activities	$\sqrt{\frac{1}{n}\sum_{i\in ActiveSet_U}\left(\sigma_i-\bar{\sigma}\right)^2}$	x_{15}	●	●
	Range of means of effective activities	$\max_{i\in ActiveSet_U}\{\sigma_i\}-\min_{i\in ActiveSet_U}\{\sigma_i\}$	x_{16}	●	●
	Relative 3 sigma head of two variables with biggest means	$\frac{\mu_\kappa+3\sigma_\kappa}{\mu_\eta+3\sigma_\eta}$	x_{17}		●
	Gap between 3 sigma head of two variables with biggest means	$\mu_\kappa+3\sigma_\kappa-(\mu_\eta+3\sigma_\eta)$	x_{18}		●
	Coefficient of variation of the variable with the biggest mean	$\frac{\mu_\kappa}{\sigma_\kappa}$	x_{19}		●
	Coefficient of variation of variable with the second biggest mean	$\frac{\mu_\eta}{\sigma_\eta}$	x_{20}		●

● : used as input variable. ■ : used as output variable.

Table 4. Output features of merging activities for establishing RT 1 and RT 2

Type	Definition	Formula	Symbol	RT 1	RT 2
Output	Relative gap between expected value of sample max and the biggest mean	$\frac{E[X_{n:n}]-\mu_{\kappa}}{\sigma_{\kappa}}$	y_1	▪	
	Ratio of sample max SD to the SD of the variable with the biggest mean	$\frac{\sqrt{Var[X_{n:n}]}}{\sigma_{\kappa}}$	y_2		▪

Table 5. The number of instances grouped by the number of effective activities for calculating sample max in each dataset

	Distribution	Number of Effective Activities										Total
		1	2	3	4	5	6	7	8	9	10	
DATASET1	Tri.	**655**	**1664**	**1547**	**1551**	**1384**	**1025**	**655**	**337**	**149**	**33**	9000
DATASET2	Beta	704	1605	1651	1481	1345	1029	695	333	126	31	9000
DATASET3	Unif.	661	1627	1624	1549	1324	1027	637	341	171	39	9000
DATASET4	Tri.	278	1014	2362	3715	3978	3099	1928	1101	429	96	18000
DATASET5	Beta	229	1116	2446	3707	3923	3005	2007	1066	410	91	18000
DATASET6	Unif.	263	1003	2408	3708	3965	3015	2008	1055	455	120	18000

Table 6. Notation used for numerical experiments

Symbol	Reference	Eq. No.	Symbol	Reference	Eq. No.
RT1	RT1		**RT3**	RT3	
Mu$_{max\_PERT}$	PERT	(14)	**Mu**$_{min\_PERT}$	PERT	(14)
Mu$_{max\_BalaA}$	Balakrishnan and Rao (1998)	(18)	**Mu**$_{min\_AGu}$	Arnold and Groeneveld (1979)	(23)
Mu$_{max\_BalaB}$	Balakrishnan and Rao (1998)	(19)	**Mu**$_{min\_AGl}$	Arnold and Groeneveld (1979)	(23)
Mu$_{max\_AGu}$	Arnold and Groeneveld (1979)	(23)	**Mu**$_{min\_AGu2}$	Arnold and Groeneveld (1979)	(23)
Mu$_{max\_AGu2}$	Arnold and Groeneveld (1979)	(23)	**Mu**$_{min\_AvenA}$	Aven (1985)	(31)
Mu$_{max\_AGl}$	Arnold and Groeneveld (1979)	(23)	**Mu**$_{min\_AvenB}$	Aven (1985)	(32)
Mu$_{max\_AvenA}$	Aven (1985)	(29)	$E[X_{1:n}]$		(34)
Mu$_{max\_AvenB}$	Aven (1985)	(30)			
Mu$_{max\_Bu}$	Bertsimas et al. (2006)	(27)			
Mu$_{max\_Bl}$	Bertsimas et al. (2006)	(28)			
$E[X_{n:n}]$		(34)			
RT2	RT2		**RT4**	RT4	
SD$_{max\_PERT}$	PERT	(15)	**SD**$_{min\_PERT}$	PERT	(15)
SD$_{max\_MorigU}$	Moriguti (1951)	(20)	**SD**$_{min\_MorigU}$	Moriguti (1951)	(20)
SD$_{max\_MorigL}$	Moriguti (1951)	(20)	**SD**$_{min\_MorigL}$	Moriguti (1951)	(20)
UB$_{SD[Xn]}$		(33)	**UB**$_{SD[X1]}$		(33)

5.1. Regression Tree vs. others

As it can be seen in Fig.2 the proposed RT1 estimates the mean value of sample max with the least relative Mean Absolute Error (MAE). Although PERT has a small error as compared with other bounds reviewed in the literature, but it is outperformed in competition with RT1. The other and more important advantage of RT1 over PERT for estimating sample max is declared when we focus on its trend over increasing the number of activities. When the number of activities increased, PERT commits a significantly increasing error (Fig.5).

Fig.4 illustrates relative MAE of the proposed RT2 and other estimates for SD of sample max. As it can be seen RT2 prepares the

least relative MAE. Although PERT has a small error as compared with other bounds, but it loses the competition if compared with RT2. Error rate of RT2 has no significant trend over increasing the number of activities as compared with PERT (Fig.6). Increasing trend of relative MAE for PERT and RT1 and RT2, seems to be suppressed into a deceasing trend; this decrement may be caused by decreasing appeared in the at this point. Since the activity durations are random numbers, after producing equal number in each group of activity number, the number of effective activities is calculated based on generated random numbers.

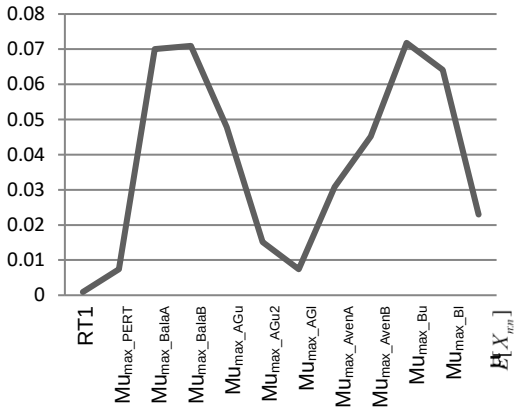


Figure 3. Relative MAE for several methods estimating mean of sample maximum

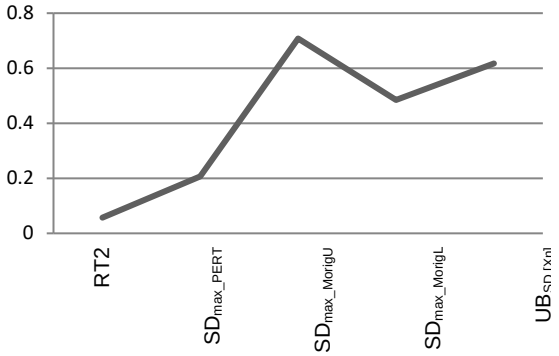


Figure 4. Relative MAE for several methods estimating SD of sample maximum

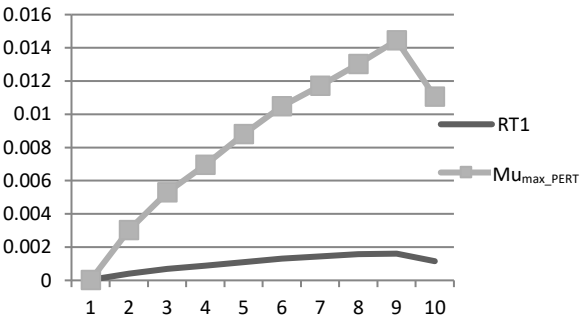


Figure 5. Relative MAE for the results generated by RT1 and PERT for estimating mean of sample maximum versus the number of effective activities

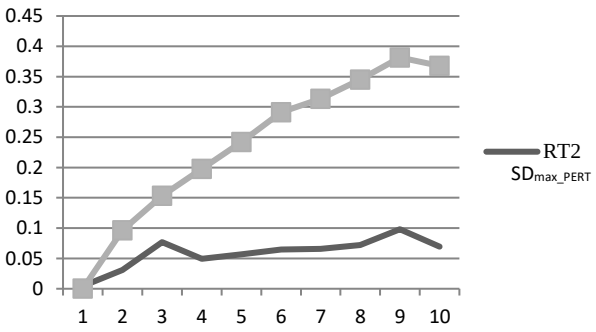


Figure 6. Relative MAE for the results generated by RT2 and PERT for estimating SD of sample maximum versus the number of effective activities

Relative MAE is calculated as:

$$\text{Relative MAE} = \frac{|observed\ vlue\ by\ simulation - estimated\ vlue|}{observed\ vlue\ by\ simulation}$$

RT3 provides good estimates for mean value of sample minimum as compared with other reviewed bounds (Fig.7) and has no significant trend by increasing the number of activities as compared with its best rival, PERT (Fig.9). RT4 provides better estimates for SD of sample minimum as compared with other reviewed bounds (Fig.8) and has no significant trend by increasing the number of activities as compared with its best rival, PERT (Fig.10).

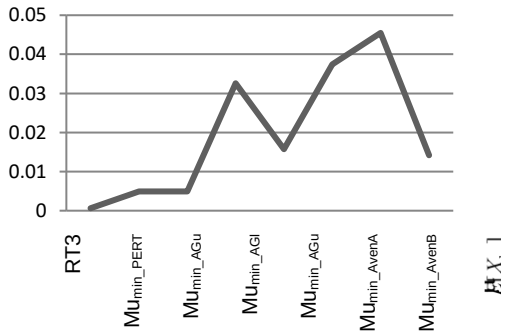


Figure 7. Relative MAE for several methods estimating mean of sample minimum

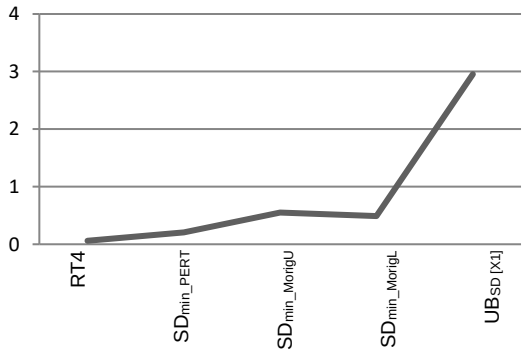


Figure 8. Relative MAE for several methods estimating SD of sample minimum

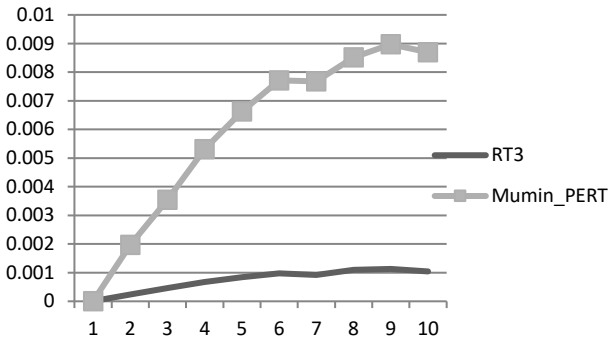


Figure 9. Relative MAE for the results generated by RT3 and PERT for estimating mean of sample minimum versus the number of effective activities

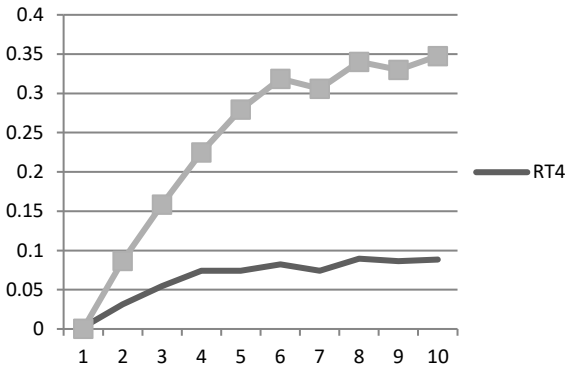


Figure ١٠. Relative MAE for the results generated by RT4 and PERT for estimating SD of sample minimum versus the number of effective activities

In this paper we recommended two heuristic upper bounds for mean and SD of sample Max/Min, denoted as $E[X_{1:n}], E[X_{n:n}]$,$UB_{SD[X1]}$ and $UB_{SD[Xn]}$. The success rates for these bounds in bounding their respective statistic is depicted in Fig.11. Although these bounds have large offset with the observed value as shown previously in Fig.4 and 8, but they are capable of providing a strong bound with almost 100% success rate. Among the other bounds reviewed in the literature, Mu_{max_AGu}, Mu_{max_AvenA}, Mu_{max_AvenB}, Mu_{max_Bu}, Mu_{max_Bl} provide a 100% success rate for upper bounding mean of sample max; and Mu_{min_AGl}, Mu_{min_AvenA}, Mu_{min_AvenB} provide a 100% success rate for lower bounding mean of sample min.

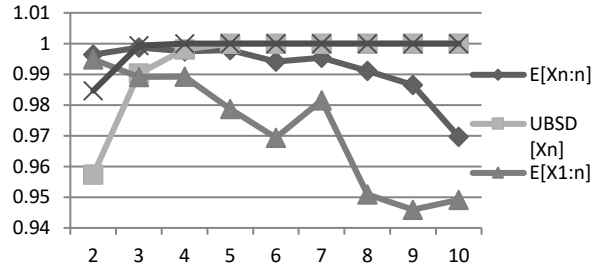
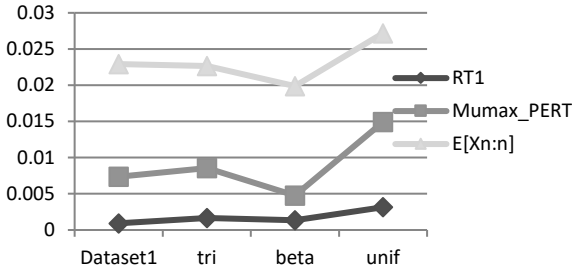
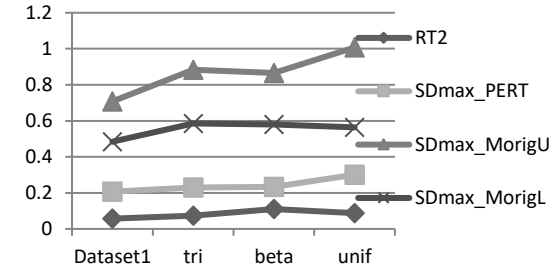


Figure 11. Success rate for several upper bounds developed

Developed regression trees are generated based on 9000 sample each one with 2000000 simulation runs. In order to validate the accuracy of RT1, RT2, RT3, and RT4, three new datasets numbered as 4-6 of random activities were generated following triangular (tri), beta and uniform (uinf) distributions. Each set contains 18000 sample projects consisting of a single merge event. In order to calculate observed statistics, 2000000 simulation runs are actuated. Relative MAE of all developed estimates is summarized in Fig.12-13.

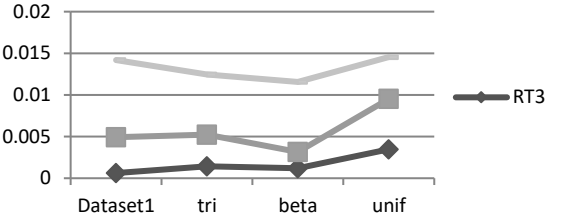


(a)

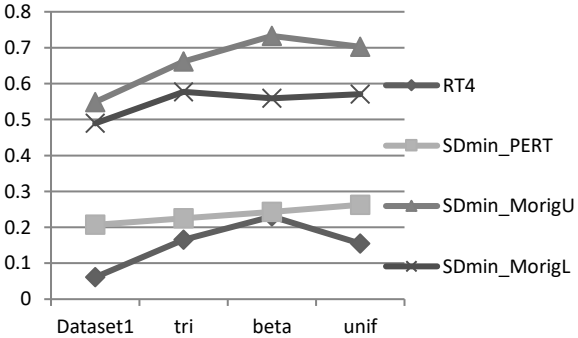


(b)

Figure ١2. Relative MAE of TR1 and RT2 compared with other methods across training and validation data



(a)



(b)

Figure 13. Relative MAE of TR3 and RT4 compared with other methods across training and validation data

As it is seen, in all cases the regression trees build based on primal 9000 random activities with triangular distribution, outperforms other best rivals even when dealing with new 18000 sample projects with other activity distributions. This proves the efficiency of using developed regression trees for every new instance problem.

5.2. Trended Regression Tree vs. Regression Tree and Cross-Valiation

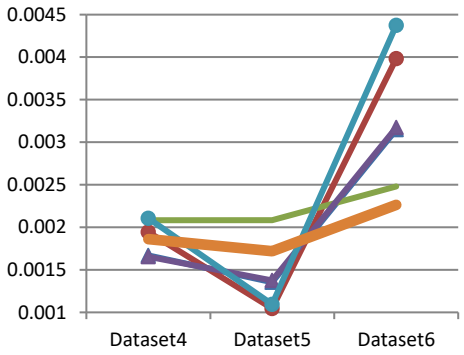
After proofing the superiority of Regression Trees over PERT and other estimates in predicting sample max/min statistics, it's time to make the same comparison between RT and TRT and selecting which dataset to use as training dataset for gaining better results.

As it was mentioned earlier, four Regression Trees are made for calculating:

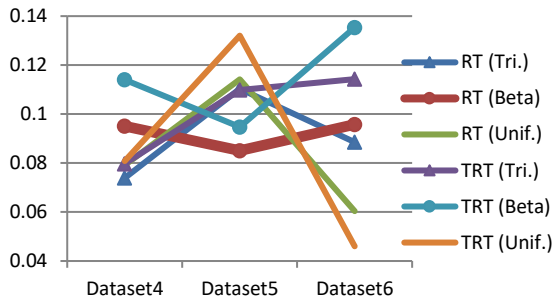
- Mean value of sample max
- Standard deviation of sample max
- Mean value of sample min
- Standard deviation of sample min.

Four other Trended Regression Trees are also made to calculate the same statistics. Generating RT and TRT is repeated three times for each of three training datasets of different probability distributions. The performance of all these 24 RT's and TRT's is evaluated when dealing with activities of different probability distribution. The results are depicted in Fig.14-15. The most reliable results of RT and TRT are defined as: having no extra oscillations when dealing with unseen data as compared with its performance when dealt with training data. In Fig.14-15 the performance chart of each most reliable RT or TRT is highlighted in bold. Based on our experimentations:

- Mean value of sample max is to be calculated with a TRT trained by a dataset following a uniform probability distribution.
- Standard deviation of sample max is to be calculated with a RT trained by a dataset following a beta probability distribution.
- Mean value of sample min is to be calculated with a TRT trained by a dataset following a triangular probability distribution.
- Standard deviation of sample min is to be calculated with a RT trained by a dataset following a beta probability distribution.

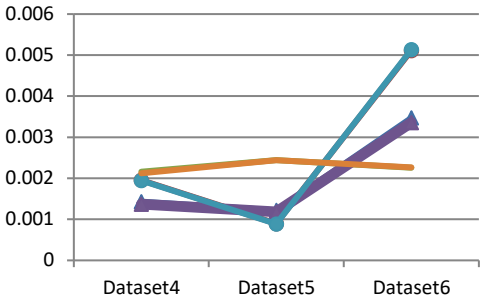


(a)

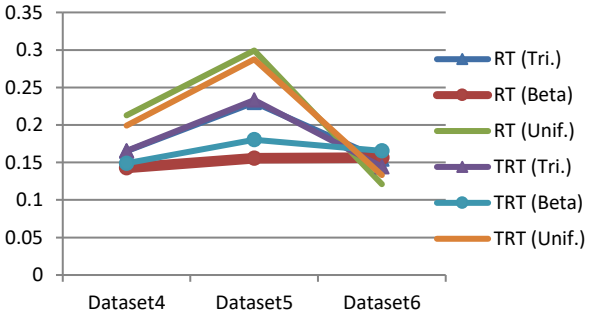


(b)

Figure 14. MAE of several RT and TRT which are made based on training data with Triangular (Tri.), Beta (Beta), and Uniform (Unif.) distribution when dealing with unseen data with different distributions to predict:
(a) mean value of sample maximum, and (b) standard deviation of sample maximum



(a)



(b)

Figure 15. MAE of several RT and TRT which are made based on training data with Triangular (Tri.), Beta (Beta), and Uniform (Unif.) distribution when dealing with unseen data with different distributions to predict:
(a) mean value of sample minimum, and (b) standard deviation of sample minimum

6. Concluding Remarks

In this paper, we introduced the notion of Trended Regression Trees (TRT) for the first time. Estimating mean and SD of sample max and sample min with TRT yields promising performance measures. These findings could be used to enhance PERT to

calculate more precise results for project makespan. Also we recommended four bounds for these statistics.

Trended Regression Trees are more promising than every other method for interpolating complex functions. For example we tried to make our estimates using genetic programming, Artificial Neural Networks or clustering, but no one of the results was comparable with RT's or TRT's.

One of the most important results of using TRT's in estimating values of some function is its clear box nature; i.e. by not trying to map an unknown function, TRT just assigns the new data to a cluster of previously observed data. Finally, the TRT is expressed in form of some if-else conditional expressions and a regression equation. Therefore, it is recommended as a easily understood tool for moelling systems. TRT can explain complex systems by local linear equations. Future potential applicaions of TRT are extended to disigning expert systems, desition rules extraction, case based resoning, … .

Nowadays, risk analysis of real world projects is actuated on powerful computers by running million runs of Monte Carlo simulation. By using TRT it is possible to refine the results of PERT; and eventually more precise results of scheduling probabilistic projects would be generated on handheld devices or even mobile cellphones. It will make it easier and perhaps requiring less time to report criticality index of activities as good as simplifying scenario analysis and several other calculations for real world probabilistic projects on the go. i.e. we can develop simple android applications for handheld devices with less hardware capabilities in wich can provide the same accuracy of expensive commercial software packages for planning with stochastic activity durations.

References

[1] Arnold, Barry C.; N. Balakrishnan, H. N. Nagaraja. (2008). A First Course in Order Statistics, Classics in applied mathematics; 54, SIAM, Philadelphia, USA.

[2] Arnold, B.C., & Groeneveld, R.A. (1979). Bounds on Expectations of Linear Systematic Statistics Based on Dependent Samples. Annals of Statistics, 7, 220-223. Erratum in Ann. Statist. 8, 1401.

[3] Aven, T. (1985). Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables. Journal of Applied Probability. 22 (3), 723-728.

[4] Avlijas, G. (2018). Examining the value of Monte Carlo simulation for project time management. Journal of Sustainable Business and Management Solutions in Emerging Economies, pp. 1–11, DOI: 10.7595/management.fon.2018.0004.

[5] Balakrishnan, N. and C. R. Rao (1998). Order Statistics: An Introduction, , in: Order Statistics: Theory & Methods (Handbook of Statistics 16), edited by: N. Balakrishnan and C.R. Rao, Volume 16, Pages 105-145, Elsevier Science, North-Holland.

[6] Balakrishnan, N., Charalambides, C., Papadatos, N. (2003). Bounds on expectation of order statistics from a finite population. Journal of Statistical Planning and Inference 113, 569–588.

[7] Ballesteros-Perez, P. (2017). M-PERT: Manual Project-Duration Estimation Technique for Teaching Scheduling Basics. Journal of Construction Engineering and Management, 143(9): 04017063-1:13.

[8] Bertsimas, D., Dunn, J., Wang, Y. (2021). Near-optimal Nonlinear Regression Trees, Operations Research Letters, 49(2), Pages 201-206.

[9] Bertsimas, D., Natarajan, K., Teo, C.-P. (2006). Tight Bounds On Expected Order Statistics, Probability in the Engineering and Informational Sciences, 20(4), 667 - 686.

[10] Bordley, R. F., Keisler, J. M., Logan, T. M. (2019). Managing Projects with Uncertain Deadlines, European Journal of Operational Research, 274, 291-302.

[11] Burgelman, J., Vanhoucke, M. (2019). Computing project makespan distributions: Markovian PERT networks revisited. Computers and Operations Research, 103, 123-133.

[12] Chu, W.-M., Chang, K.-Y., Lu, C.-Y., Hsu, C.-H., Liu, C.-H., Hsiao, Y.-C. (2014). A New Approach to Determine the Critical Path in Stochastic Activity Network, Mathematical Problems in Engineering, , 2014, 547627, 1-13.

[13] Elmaghraby, S. E. (2005). On the fallacy of averages in project risk management, European Journal of Operational Research, 165, 307–313

[14] Gkioulekas, I., Papageorgiou, L.G. (2021). Tree regression models using statistical testing and mixed integer programming, Computers & Industrial Engineering, 153, 107059.

[15] Golenko-Ginzburg, D., Gonik, A. (1998). A heuristic for network project scheduling with random activity durations depending on the resource allocation. International Journal of Production Economics, 55, 149-162.

[16] Hajdu, M., Bokor, O. (2016). Sensitivity analysis in PERT networks: Does activity duration distribution matter?. Automation in Construction, 65, 1–8.

[17] Hajiali, M., Mosavi, M.R., and Shahanaghi, K. (2020). A new decision support system at estimation of project completion time considering the combination of artificial intelligence methods based on earn value management framework, International Journal of Industrial Engineering: Theory, Application and Practice 27(1), 1-12.

[18] Hernandez-Bastida, A., Fernandez-Sanchez, M. P. (2018). How adding new information modifies the estimation of the mean and the variance in PERT: a maximum entropy distribution approach, Annals of Operations Research, 274, 291-308.

[19] Herrerıas-Velasco, J.M., Herrerıas-Pleguezuelo, R., van Dorp, J.R. (2011). Revisiting the PERT mean and variance. European Journal of Operational Research, 210, 448–451.

[20] Hillier, F.S., Lieberman, G.J. (2010). Introduction to Operations Research, 9th Edition, Mc Graw-Hill, New York, NY.

[21] Kelley, J.E., Walker, M. (1959). Critical-Path Planning and Scheduling. Proceedings of the Eastern Joint Computer Conference, IRE-AIEE-ACM '59 (Eastern).

[22]Kim, S.D., Hammond, R.K., Bickel, J.E. (2014). Improved Mean and Variance Estimating Formulas for PERT Analyses, IEEE Transactions n Engineering Management, 61(2), 362-369.

[23]Lee, D.-E., Arditi, D., Son, C.-B. (2013). The Probability Distribution of Project Completion Times in Simulation-based Scheduling, KSCE Journal of Civil Engineering, 17(4), 638-645.

[24] MacCrimmon, K. R., Ryavec, C. A. (1964). An analytical study of the PERT assumptions, Operations Research, 12(1), 16–37.

[25] Malcolm, D.J., Roseboom, J.H., Clark, C.E. and Fazar, W. (1959). Application of a technique for research and development program evaluation. Operations Research, 7: 646-669.
[26] Milian, Z. (2008). Monte Carlo Simulation with Exact Analysis for Stochastic PERT Networks. Proceedings of the 25th International Symposium on Automation and Robotics in Construction, Vilnius, Lithuania, 598-603.
[27] Moriguti, S. (1951). Extremal properties of extreme value distributions. Annals of Mathematical Statistics, 22, 523–536.
[28] Nekoufar, M., Movafaghpour, M.A. (2017). Assessing the Effectiveness of a Single Curriculum for a Group of Students with Different Mathematics Literacy. Journal of Interpolation and Approximation in Scientific Computing, 2, 38-48.
[29] Papadatos, N. (1995). Maximum Variance of Order Statistics, Maximum Variance of Order Statistics, Annals of the Institute of Statistical Mathematics, 47(1), 185-193.
[30] Pontrandolfo, P. (2000). Project duration in stochastic networks by the PERT-path technique. International Journal of Project Management, 18, 215-222.
[31] Rostami, S., Creemers, S., & Leus, R. (2018). New strategies for stochastic resource-constrained project scheduling. Journal of Scheduling, 21, 349-365.
[32] Rychlik, T. (1998). Bounds for expectations of L-estimates, in: Order Statistics: Theory & Methods (Handbook of Statistics 16), edited by: N. Balakrishnan and C.R. Rao, Volume 16, 105-145, Elsevier Science, North-Holland.
[33] Rychlik, T. (2008). Extreme variances of order statistics in dependent samples. Statistics and Probability Letters, 78, 1577–1582.
[34] Sackey, S. Kim, B.-S. (2019). Schedule Risk Analysis using a Proposed Modified Variance and Mean of the Original Program Evaluation and Review Technique Model, KSCE Journal of Civil Engineering, , 23, 1484-1492.
[35] van Slyke, R. M. (1963). Monte Carlo Methods and the PERT Problem. Operations Research, 11(5), 839-860.
[36] Winston, W.L., 2004. Operations Research, Applications and Algorithms. Duxbury Press, Pacific Grove.
[37] Yan, H.-S., and Shang, Z.G. (2019). Product design time forecast using relative entropy kernel regression. International Journal of Industrial Engineering: Theory, Application and Practice, 26(3), 343-360.
[38] Yao, M.-J., Chu, W.-M. (2007). New approximation algorithm for obtaining the probability distribution function for project completion time. Computers and Mathematics with Applications, 54, 282–295.
[39] Zhen-ting, H., Xuan, Z., Xiang-Xing, K. (2010). A new analytical algorithm for computing probability distribution of project completion time. Journal of Central South University of Technology, 17: 1006−1010.