AN INTEGRATED REGRESSION-BOOTSTRAP SAMPLING SCHEDULING METHOD FOR PROBABILISTIC DURATION RANGE ESTIMATION OF CONSTRUCTION PROJECTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÖZGÜR BARIŞ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
CIVIL ENGINEERING

SEPTEMBER 2019

Approval of the thesis:

AN INTEGRATED REGRESSION-BOOTSTRAP SAMPLING SCHEDULING METHOD FOR PROBABILISTIC DURATION RANGE ESTIMATION OF CONSTRUCTION PROJECTS

submitted by **ÖZGÜR BARIŞ** in partial fulfillment of the requirements for the degree of **Master of Science in Civil Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** ________________

Prof. Dr. Ahmet Türer
Head of Department, **Civil Engineering** ________________

Prof. Dr. Rıfat Sönmez
Supervisor, **Civil Engineering, METU** ________________

Examining Committee Members:

Assist. Prof. Dr. Aslı Akçamete Güngör
Civil Engineering, METU ________________

Prof. Dr. Rıfat Sönmez
Civil Engineering, METU ________________

Assist. Prof. Dr. Onur Behzat Tokdemir
Civil Engineering, METU ________________

Assist. Prof. Dr. Güzide Atasoy Özcan
Civil Engineering, METU ________________

Assist. Prof. Dr. Saman Aminbakhsh
Civil Engineering, Atılım University ________________

Date: 05.09.2019

Name, Surname: Özgür Barış

Signature:

ABSTRACT

AN INTEGRATED REGRESSION-BOOTSTRAP SAMPLING SCHEDULING METHOD FOR PROBABILISTIC DURATION RANGE ESTIMATION OF CONSTRUCTION PROJECTS

Barış, Özgür
Master of Science, Civil Engineering
Supervisor: Prof. Dr. Rıfat Sönmez

September 2019, 90 pages

The estimation of project duration is important in construction projects, since it directly affects the costs and the success of the projects. Schedule risks can be determined along with the duration estimations in order to assess the possibility of delay penalties and additional costs. Therefore, the accuracy of the estimation of project duration range is very crucial in construction industry to assess the schedule risks. This thesis presents an integrated approach of non-parametric bootstrap sampling method and regression analysis in order to determine the project duration ranges with their probabilities for construction projects. This approach combines both the regression and probabilistic methods in order to provide the range estimation of construction project's duration with its corresponding probability. The non-parametric bootstrap sampling method, when integrated with the regression analysis, has advantages for range estimation purposes when compared to classical simulation methods such as Monte Carlo Simulation method or probabilistic methods such as Program Evaluation and Review Technique (PERT), since it requires no assumptions regarding the probability distributions and correlations of the input data. Moreover, the proposed integrated regression-bootstrap sampling scheduling method is expected to provide more accurate results than the traditional methods due to regression analysis, which is used to take the effecting factors of the productivity into account and non-parametric bootstrap sampling method, which increases the sample size by resampling the original sample without requiring any assumptions of distributions and correlations of the activities. To expose the advantages and accuracy of the new integrated regression-bootstrap sampling scheduling method, the new method is compared with the traditional probabilistic scheduling methods through two case studies. The comparisons reveal that the proposed method presents a practical non-parametric approach that provides adequate project duration range for realistic evaluation of schedule risks of construction projects.

Keywords: Probabilistic Risk Analysis, Bootstrap Method, Regression Analysis, Construction Duration, Range Estimation

ÖZ

İNŞAAT PROJELERİNİN OLASILIKSAL SÜRE ARALIĞININ BELİRLENMESİ İÇİN ENTEGRE REGRESYON-BOOTSTRAP ÖRNEKLEME PROGRAMLAMA YÖNTEMİ

Barış, Özgür
Yüksek Lisans, İnşaat Mühendisliği
Tez Danışmanı: Prof. Dr. Rıfat Sönmez

Eylül 2019, 90 sayfa

İnşaat projelerinde toplam sürenin tahmini, maliyetleri ve projenin başarısını doğrudan etkilediği için önemlidir. Ayrıca, gecikme riskleri de, gecikme cezalarının ve ek maliyetlerin olasılığını da değerlendirmek için süre tahminleriyle birlikte belirlenebilir. Bu nedenle, süre aralığı tahmininin doğruluğu inşaat sektöründe çok önemlidir. Bu tez, proje sürelerinin inşaat projeleri için olasılıkları ile belirlenmesi amacıyla parametrik olmayan bootstrap örnekleme yöntemini ve regresyon analizini entegre bir yaklaşım sunmaktadır. Bu yaklaşım, inşaat projesinin süresinin aralık tahminini ve bu aralık değerlerine karşılık gelen olasılığı sağlamak için hem regresyon hem de olasılıksal yöntemleri birleştirmektedir. Parametrik olmayan bootstrap örnekleme yöntemi, regresyon analizine entegre edildiğinde verilerin olasılık dağılımları ve korelasyonları hakkında varsayım gerektirmediği için, Monte Carlo Simülasyon yöntemi gibi klasik simülasyon yöntemleriyle veya Program Değerlendirme ve Gözden Geçirme Tekniği (PERT) gibi olasılıksal yöntemlerle karşılaştırıldığında aralık tahmini için avantajlara sahiptir. Dahası, önerilen entegre regresyon-bootstrap örnekleme planlama yönteminin, verimliliği etkileyen faktörleri hesaba katmak için kullanılan regresyon analizi ve faaliyetlerin dağılım varsayımları ile korelasyonları gerekmeksizin orijinal numuneyi yeniden örnekleyerek numune büyüklüğünü arttıran parametrik olmayan bootstrap örnekleme yöntemi ile geleneksel yöntemlerden daha doğru sonuçlar vermesi beklenmektedir. Yeni entegre regresyon-bootstrap örnekleme zamanlama yönteminin avantajlarını ve doğruluğunu ortaya çıkarmak için, yeni yöntem iki örnek olay incelemesiyle geleneksel zamanlama risk değerlendirme modelleriyle karşılaştırılmıştır. Karşılaştırmalar, önerilen yöntemin, inşaat projelerinin zamanlama risklerinin gerçekçi bir şekilde değerlendirilmesi için yeterli proje süresi aralığını sağlayan, parametrik olmayan pratik bir yaklaşım sunduğunu ortaya koymaktadır.

Anahtar Kelimeler: Olasılıksal Risk Analizi, Bootstrap Metodu, Regresyon Analizi, İnşaat Süreleri, Aralık Tahmini

Dedicated to my beloved family

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my supervisor Prof. Dr. Rıfat Sönmez for his brilliant guidance and comments, criticism and encouragement throughout this study.

I would like to thank the examining committee members, for their invaluable ideas and comments. I would also like to gratefully thank to Res. Asst. Murat Altun for his significant support, guidance and motivation throughout the study.

I would like to also present special thanks to my colleagues and managers at the companies I have worked and currently been working for their understandings and tolerations.

I wish to thank my friends Mert Susuzlu, Uğurcan Ertaymaz, Elif Karakoçak, Hasan Sezai Sevindik, Namık Erkal, Barış Özel, Eftun Türkşanlı, Aysu Özden and Cansu Coşkun for their invaluable helps, great friendships and continuous supports.

I am deeply grateful to my mother Nesrin Barış, my father Beşir Barış, my brother Onur Barış and his wife Zerin Barış for keeping me strong with their continuous supports and positive energy throughout my life.

Finally, I would like to thank Dilşen Kuzucuoğlu for her toleration and support throughout the whole process. She always helps me to overcome the problems and makes me feel strong all the time.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

All construction projects have a certain budget and deadline. Contractors must complete the projects within the given budget and timeline in order not to make any losses. Therefore, they should complete their cost and duration estimations throughout the tendering stage in order not to make a loss. However, sometimes, in particular in design-build projects the scope of the project and drawings may not be clear at the tendering stage, so that it is very challenging to perform a detailed estimate for cost and duration (Sonmez, 2008). Hence, conceptual cost and duration estimations should be performed at the tendering stage of construction projects, in order to evaluate the cost and schedule risks and make decision to bid or not to bid. On the other hand, competition is getting harder when compared to past. Therefore, the accuracies of duration and cost estimations are getting more important than ever before. The contractors should perform their cost and duration estimations as accurately as possible to compete with the opponents, evaluate the opportunities of the new projects and minimize the risks of penalties such as delay penalty in case a project is awarded.

There are two components of costs, which are direct and indirect costs. The details of the cost and cost components will not be explained in detail, since the topic of this thesis is based on the range estimation of project duration. However, it is necessary to mention that cost also depends on total project duration as shown in Figure 1.1 for emphasizing the importance of the total time estimations of construction projects. As it can be clearly seen in the Figure 1.1, duration estimation is very important. The accuracy of the estimation of the construction project’s duration affects costs and inaccurate duration estimates may lead to low profits or even losses.

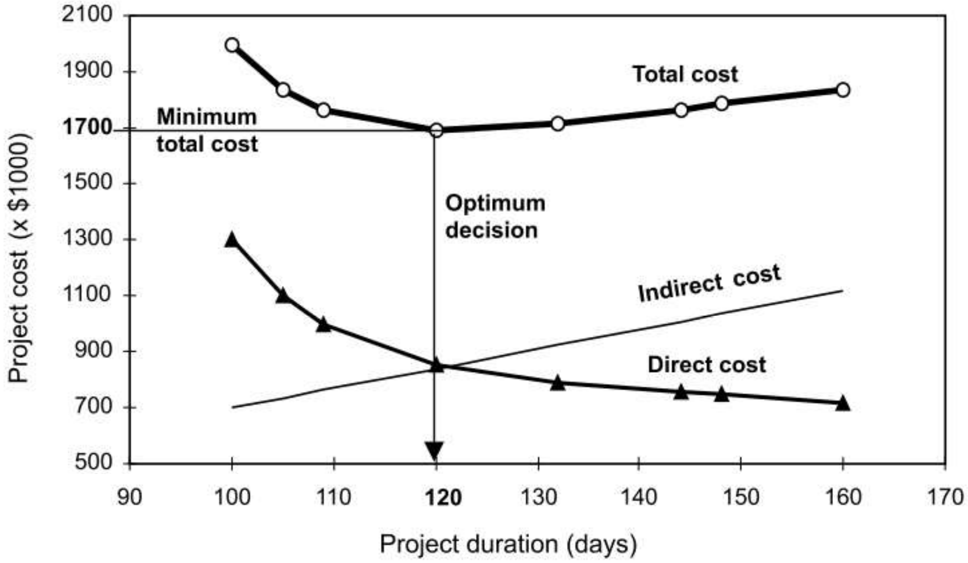


Figure 1.1. Time-cost relationship example adapted from Hegazy (1999)

In addition to affecting costs, duration estimation has one more important role. Duration estimations should show the contractor if the project is going to be delayed or not and if the project is going to be delayed, the probability of the delay should also be known. However, it is a challenging task for decision makers to understand the probability of delay and the risk of the extra costs depending on duration with a single point estimation. Therefore, it is important for decision makers or project managers to have a range estimation of project duration with its probabilities. This thesis will provide a robust model for decision makers in this manner. The classical scheduling methods will be mentioned in the next section.

1.1. Scheduling Methods

There are several methods for project scheduling. Some of them provides point estimations, whereas the other ones are capable of providing range estimates. In addition, some of those methods are deterministic. On the other hand, some of them are probabilistic methods. Deterministic methods and point estimations are not sufficient for decision makers or project managers to comment on the prediction of the duration due to the lack of probability and a range estimation. Probabilistic methods, such as Project Evaluation and Review Technique (PERT) can provide the durations for the desired probability levels with the assumption of normality and without making any simulations. On the other side of the medallion, simulation methods provide range estimations with their corresponding probabilities by simulating the schedule. The project duration range estimates enable decision makers to easily comment on the expected schedule and foresee the delay risk with its probability. It becomes also possible to take precautions beforehand, since the delay risk becomes clear by estimating the duration with simulation methods.

Although it may seem as probabilistic methods are always more sophisticated and accurate than deterministic methods, one should be careful about the assumptions in those methods. Main drawbacks of the existing probabilistic methods are the assumptions. The main assumptions are the statistical distributions of the activities' and the whole project's durations, which may directly affect the estimation result. It is not simple at all to identify the distribution types exactly. Similarly, there are usually correlations among durations of activities which are neglected in the general PERT method and Monte Carlo Simulation methods.

When the most commonly used methods in the previous studies are investigated, critical path method (CPM), PERT and Monte Carlo Simulation method come out to be the most widely used techniques. CPM is the most popular deterministic method. Although it is a deterministic method, it is widely used since its development, due to its easiness to perform (Lee & Arditi, 2006). CPM is also the underlying technique of PERT. PERT is a probabilistic method and it also provides the range estimate values with the duration estimations of the activities and project in addition to CPM. Moreover, it is also possible to compute the probability of project completion with the assumption of normality of total project duration. On the other hand, it is generally assumed that distributions of activities follow beta distribution in PERT (Lee, 2005), whereas the distribution of the whole project is considered to follow normal distribution. Monte Carlo Simulation method is also a probabilistic method and requires user to define the statistical distribution types of activity durations such as normal, beta, triangular, uniform distributions, etc., which is different than PERT.

Monte Carlo is capable of providing the range estimations together with their corresponding probabilities by simulating the project schedules as much as required by the user. As mentioned above, the results may considerably change according to defined distribution types in Monte Carlo Simulation method. Moreover, the activities in construction projects may not always follow beta distribution and the total durations of construction projects may not always follow normal distribution, which are the assumptions of PERT.

In addition to the above-mentioned drawbacks, there is also another issue regarding the past data. It is generally hard to have sufficient data for accurate prediction purposes to get prepared for the tenders of the new upcoming projects, since the records and data of the completed projects are generally not saved and stored well in the database of the contractors. Therefore, it gets also harder to analyze the past data and conduct a duration estimation with small data sets.

In accordance to the foregoing explanations, it is apparent that there is a lack of a robust model for practical and realistic evaluation of construction schedule risks, which does not require the user to define the distribution types of activities, correlations between durations of activities and can eliminate the disadvantage of having a small data set while providing the duration range estimation of construction projects with the corresponding probabilities. This gap will be filled by developing an integrated method, which combines non-parametric bootstrap method with linear regression analysis. As a result, a new method, which is integrated regression-bootstrap scheduling method for construction projects, will be presented in this thesis. There will be no need to determine the statistical distributions and correlations of the activity durations before conducting the range estimates and the disadvantage of having a small data set will be eliminated with the usage of bootstrap. In addition, the productivities of the activities will be used as the input for the analysis to evaluate the construction schedule risks in a more realistic manner. Therefore, the linear regression analysis is used to include the relations between the affecting factors and the productivities of each activity. So that, the variations in the activity durations can be explained and it makes the method more realistic in terms of predicting the project durations and their range.

The rest of the thesis is organized as follows. Literature review of productivity, probabilistic estimation of project duration methods and bootstrap method will be presented in chapter 2. The new methodology and the proposed tool integrating the non-parametric bootstrap and linear regression analysis will be explained in chapter 3. Chapter 4 will involve two case studies that demonstrate the proposed method and the tool and also the comparison of the results obtained by the proposed method with PERT and Monte Carlo Simulation. The conclusion and final remarks will be mentioned in chapter 5.

CHAPTER 2

LITERATURE REVIEW

In this chapter, previous research on productivity, probabilistic estimation of project duration methods and bootstrap method will be introduced, since this study is based on these topics. Project evaluation and review technique (PERT) and Monte Carlo Simulation will also be investigated due to the fact that they have been widely used in most of the previous researches, which address probabilistic estimation of project duration and also both of these two methods are used for comparison.

2.1. Productivity

Productivity is one of the major issues in construction. It can directly affect the duration and cost of a construction project. Hence, it always draws attention in literature. In some of the studies, researchers are trying to develop models in order to estimate the productivity of some activities in a project, whereas in some other studies researchers are trying to find the factors that affect the productivity. In this study, the productivities of each activity are used along with their affecting factors in order to include the variations within the activities.

Productivity in construction industry may be considered as labor productivity, since construction industry is labor-intensive. Therefore, the previous researches introduced below are based on labor productivity only.

Sonmez and Rowings (1998) developed a model for productivity forecasting, using regression analysis and neural network method. In their study, the model is developed to estimate the productivity of concrete pouring, formwork and concrete finishing activities. The data were obtained from eight building projects completed in Iowa, USA using the database of a contractor. It should also be noted that, only the data of seven projects were used for developing the model, because the data of the remaining project were planned to be used for testing the accuracy of the model by calculating the mean squared error (MSE) and mean absolute percent error (MAPE) of the results calculated by the developed model. The factors were also taken from the same database according to available data. For instance, quantities completed, job type, crew size, percent overtime, percent laborer, temperature, humidity, precipitation and concrete pump information were available in the database of the contractor and used as the factors affecting the productivity. Then the researchers used all the factors that could affect productivity of activities based on the availability of the data. After this, regression models were built in a parsimonious manner. Unnecessary parameters were dropped from the model by backward elimination method in each step until the model consisted of adequate data with reasonable accuracy when compared to models obtained in previous steps. Neural network models were included to improve the parameter selection of regression analysis. The methodology of this study can be seen in Figure 2.1.

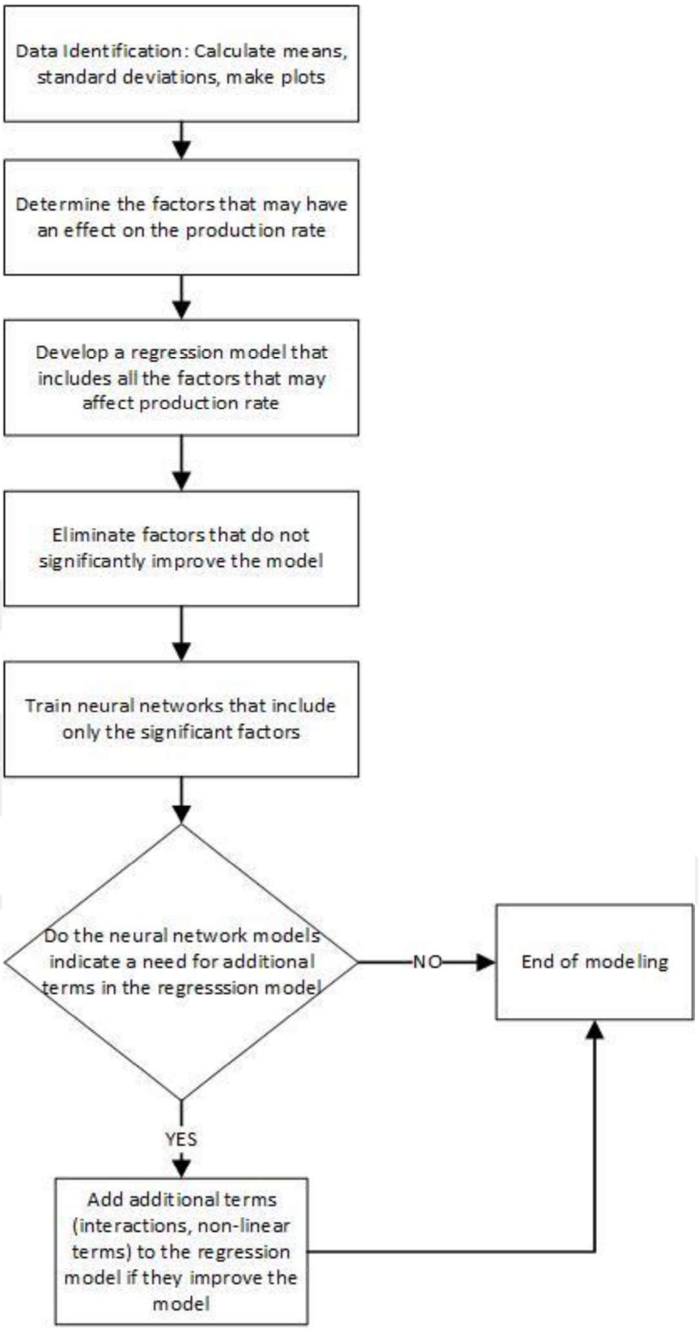

Data Identification: Calculate means, standard deviations, make plots
Determine the factors that may have an effect on the production rate
Develop a regression model that includes all the factors that may affect production rate
Eliminate factors that do not significantly improve the model
Train neural networks that include only the significant factors
Do the neural network models indicate a need for additional terms in the regresssion model
NO
End of modeling
YES
Add additional terms (interactions, non-linear terms) to the regression model if they improve the model

Figure 2.1. The Methodology of Sonmez & Rowings (1998)

AbouRizk, Knowles, and Hermann (2001) aimed at building up a neural network model to predict the productivity multiplier of pipe installation activity in industrial projects, for estimating the labor production rates in industrial projects. As stated in the paper, normally, the labor production rates were determined at the beginning of the projects as an average value and then those productivity values have been modified according to that specific project for instance considering the skill of labors or weather conditions. That was the starting point of the researchers and their idea was to come up with a model that computes the multiplier value in order to modify the labor productivity for obtaining the accurate production rates. To be able to develop such a model, factors affecting the productivity were determined first. For this purpose, literature review and a survey (questionnaire) were used and the factors were determined. Then the database of a company was used to gather the data of previously completed projects and interviews were also conducted with the field personnel for being sure that all available data were covered and gathered. After completing the data collection stage, neural network model was implemented using the collected data to get the productivity multiplier value. The results obtained from the model were validated by comparing them with the values of estimators and actual multiplier values. The validation has shown that the developed neural network model was accurate enough and provided a robust solution when compared to values estimated by estimators, since there was no subjectivity in neural network model.

Park (2006) has attempted to provide a conceptual framework to develop a model for estimating the productivity. In this regard, a general regression model has been provided. Besides that, the factors affecting the productivity were identified through extensive literature review. Those factors were scheduled overtime, change orders, materials management, weather and human factors and all of them were explained in a more detailed manner. No quantitative historical data were collected and no validation has been conducted as well. This research may be considered as a kind of general review rather than a new finding.

Kazaz and Ulubeyli (2007) investigated the factors that have an effect on productivity in Turkey. In order to identify those factors, questionnaires and interviews were conducted with the engineers. Then the obtained qualitative data were converted to quantitative data with statistical methods. First, the frequencies of the answers were calculated and then the relative importance indexes were computed. As the result, money related issues came out to be the most important factors, since they were found to be the most effective factors on productivity.

Al-Zwainy, Abdulmajeed, and Aljumaily (2013) developed a multivariable linear regression model in their research in order to predict the labor productivity of marble finishing works in construction projects taking place in Iraq. Before developing the model, the data were collected through interviews and direct observation at site. The factors affecting the productivity were chosen based on the interviews conducted with the engineers and project managers. On the other hand, the historical productivity data were collected through observations at ten different project sites in different regions of Iraq. After finalizing the data collection process, the linear regression model was developed using the collected data in order to predict the productivity value. The model was validated by conducting statistical tests such as probable error, standard error and test of significance. In addition to these statistical tests, the model was also validated using the data obtained from a completed real project, which was not used during the development stage of the model. The results of the model and validations showed that the developed model provides accurate results. The proposed linear regression model was also compared to the results of a neural network. The proposed regression model also gives slightly better results when compared to neural network.

Nasirzadeh and Nojedehi (2013) aimed to provide a system dynamics (SD) based approach to model labor productivity that accounts for all the influencing factors and their inter-relations. First, the factors were identified by conducting literature review and interviews with experts. Then the model was constructed first in a qualitative manner with showing the inter-relations between the factors and effects of the factors. After that, the model was converted to a quantitative model. The developed model was tested using a previously completed housing project for estimating the productivity. The model provided satisfactory results, however authors also concluded that more completed projects were needed for the validation of the model.

Li, Chow, Zhu, and Lin (2016) aimed to evaluate the effect of high temperature on productivity of rebar workers in China. For this purpose, they have also implemented a regression model in order to see the effects of high temperature quantitatively. However, for the data collection stage, the adopted approaches were literature review, interviews with experts in order to get their opinions and observation at site. Weather data was collected throughout the observations. The main purpose of the observations was to identify how many hours the workers are really working. It means that the idle times and the indirect times were not considered and only the direct time was included in the model in order to purely compute the productivity. The observations were made in two different reinforced concrete building construction sites and the weather data was collected in these two sites as well. The other factors were determined with the help of literature review and expert opinions. After all these data collection processes, the collected data were used to implement the regression model. So that, the effect of high temperature on rebar workers have been revealed by conducting the regression model. However, there was no validation method.

As a more recent study, Usukhbayar and Choi (2018) studied the effects of climatic factors on the construction productivity in Mongolia. In order to evaluate the effects of climatic factors, they investigated the concreting activity. The climatic data were collected from the climate station located in Ulaanbaatar, Mongolia. The historical productivity values have been collected through the surveys. After that, regression model was built and regression analysis was conducted in order to see the effects of climatic factors. The validation of the model and collected data were conducted using statistical definitions and formulae.

The objectives of the above studies can be divided into two general categories, which are developing a model that predicts the productivity in construction projects due to the importance of productivity in estimating the total project duration and investigating the affecting factors of productivity. Sonmez and Rowings (1998), AbouRizk et al. (2001), Park (2006), Al-Zwainy et al. (2013) and Nasirzadeh and Nojedehi (2013) developed a model for predicting the productivity, whereas Kazaz and Ulubeyli (2007), Li et al. (2016) and Usukhbayar and Choi (2018) investigated the affecting factors of productivity.

To sum up, Nasirzadeh and Nojedehi (2013) conducted a system dynamics based approach, which was developed in 1960. AbouRizk et al. (2001) implemented neural network model and Sonmez and Rowings (1998) also used neural network model but together with regression model. On the other hand; Park (2006) and Al-Zwainy et al. (2013) implemented regression models to predict the productivity. As seen, neural network models and regression models are the most common methods for estimation. In this thesis, also regression models will be adopted to model the relation between the productivity and influencing factors. Rather than finding a multiplier as stated in the research of AbouRizk et al. (2001), the model developed in this thesis will directly calculate the productivity according to input. Then using those productivity values, total project duration range and corresponding probabilities will be calculated. The data of one of the case studies demonstrated in this thesis will be the same as the data used in the study of Sonmez and Rowings (1998). Apart from these, all of the above-mentioned studies are limited to the gathered data. This lack will be eliminated using bootstrap in this thesis. For instance, if there are only 10 data available in the database, they may be resampled even 1000 times, each resample having again 10 data. Besides, the studies only provide a point estimate for the productivity and do not provide a range estimate for the productivity, or project duration. On the other hand, the purpose of Kazaz and Ulubeyli (2007) was just to identify the factors affecting the construction labor productivity. So that, they have just conducted interviews and questionnaires in order to determine affecting factors. On the other hand, Li et al. (2016) and Usukhbayar and Choi (2018) have also implemented a regression model additionally, in opposition to the study of Kazaz and Ulubeyli (2007), since they wanted to find out the effect of the temperature and climate quantitatively. If the idea would be only to find the affecting factors without considering their effects, only questionnaires and interviews might be satisfactory. Nonetheless, to see the effects of the factors quantitatively, one should implement a model such as regression model. In addition, it is also clear that the most common method is regression model for finding the effect of the factors on productivity, since regression model is easy to implement and its result is easy to evaluate. Therefore, this thesis study also adapts regression models rather than another modeling method. However, the validation is missing in the studies mentioned above.

In all the studies related to productivity, first of all historical data were collected in order to develop a model to predict the productivity or in order to determine the effects of the factors on productivity. The factors were determined through surveys/questionnaires, literature review and interviews mainly. The historical productivity data were obtained generally through observations and database of the contractors. To develop a model that predicts the productivity of a future project with known information of factors of that future project (input), one should use the historical data. The past performances and completed projects are going to underlie the model for estimations. The most common technique to model the productivity is regression analysis, since it is relatively easy to implement and provides accurate solutions. Regression models developed based on the historical data will also be used in this thesis to find the relations between the factors and productivity values of activities.

2.2. Probabilistic Estimation of Project Durations

The studies focused on probabilistic estimation of project durations will be investigated below, since it is important to understand the previous techniques in order to overcome and enhance them. So that, although PERT and Monte Carlo Simulation are not used while developing the proposed method in this thesis, previous studies used PERT and Monte Carlo Simulation method will be mentioned. Additionally, both traditional methods will be used for comparison in chapter 4.

Ergin, Balas, and Keyder (1995) came up with the idea to develop a new method to predict the duration and cost of the projects. The main advantages and disadvantages of PERT and Monte Carlo simulation were explained. Then they suggested using

PERT with triangular distribution instead of beta distribution (standard PERT). Then their idea of performing PERT with triangular distribution was validated using a completed offshore platform project. The results of proposed method, standard PERT and Monte Carlo simulations were compared with each other and real project data. The results have shown that the proposed model, which is PERT with triangular distribution, and Monte Carlo Simulation have provided similar results. In addition, their accuracies were better than standard PERT, since their results were closer to actual duration and cost of the completed offshore platform project, which used as the case study to validate the proposed technique. The researchers have also concluded that this has also shown the importance of probability distribution while performing PERT, since PERT came up to be sensitive to probability distribution types.

Isidore and Back (2002) established a new technique that integrates the range estimation of project cost and duration. Monte Carlo Simulation method was implemented together with regression model. First, the project cost and duration were separately found as range estimates with corresponding probabilities using Monte Carlo Simulation. Then these two were combined using the regression model in order to find the relation between the cost and duration. So that, if one could determine the desired level of accuracy of duration, then the corresponding cost value could also be computed or vice versa. However, it was found that linear regression model was not adequate based on the coefficient of determination (R^2 value) and so that polynomial regression analysis was conducted, since it provided a higher R^2 value, which means that polynomial regression model could predict the result accurately when compared to the linear model. Although it was stated that this technique was applied to several example projects and the results were satisfactory, the validation method was not shown in the study. However, the following outcomes could be understood as the contribution of this research. Taking the duration and cost estimation of -for instance-95% accuracy level for a project separately may mislead the decision makers. If the duration and cost estimations are required of the same project, they should be calculated in an integrated manner. So that, the duration can be selected for a desired level of accuracy and then corresponding cost value can be found using regression model or vice versa. Estimating the cost and duration simultaneously in this way could provide results that are more accurate than the results obtained separately.

Öztaş and Ökmen (2005) developed a model to take risks of the activities into account in order to find the total project duration range with its probability considering the risk factors. The data were introduced to Monte Carlo Simulation for obtaining the results. The proposed method has been validated using a completed real project. As the result of the case study, the model provided realistic findings. However, there might subjective inputs exist, since the risks and their effects were determined based on the engineering judgement when the historical data were not available.

Barraza (2011) aimed to develop a method in order to allocate the total project time contingency to activities. For this purpose, Monte Carlo method was used, since it provides range estimations with corresponding probabilities. In this paper, probabilistic project duration estimation was also explained and time contingency was formulated. The proposed method was used in a bridge project in order to show how to perform the proposed model. At the end, the author has concluded that project managers might benefit from this method, since it did not directly give the total project time contingency but the time contingencies of each activity. The allocation of the time contingency to each activity could allow project managers to see the delay risks while executing the project, since they would know the buffer times of each activity instead of the buffer time of the whole project.

Nguyen, Phan, and Tang (2013) have developed a method to predict the duration range of the multistory building projects with their probabilities. They have used Monte Carlo simulation method to iterate the possibilities for obtaining the range estimation. First, the construction sequence and critical activities were determined through questionnaires with experts. As mentioned above, one should determine the statistical distribution (probability distribution) types of each activity before continuing with Monte Carlo. The authors conducted another questionnaire with experts in order to determine the distribution function. In addition, they have also conducted a goodness of fit test for the result of the second survey while determining the probability distribution. Then they applied Monte Carlo Simulation to the available data and they validated their model using a completed real project, which was also a multistory building. The results have shown that the proposed method was reasonably accurate, since the duration of the real project was within the range of mean ±1 standard deviation. However, one should also note that the proposed model was limited to multistory building projects.

Liu (2013) has explained PERT in detail. In addition, the differences and similarities between PERT and CPM were also investigated. Then the duration of an example project was calculated using PERT. Although there is no new finding in this study, it is a good summary to understand the concepts of PERT and CPM.

Hajdu and Bokor (2014) and Hajdu and Bokor (2016) have examined the effect of the different activity duration distributions such as beta, uniform, triangular and lognormal distributions, on project duration using PERT. They have also included the three-point (optimistic time, most likely time and pessimistic time) estimation difference in their study for further comparison. Both artificially created projects and real projects were used for examination of the effects. The outcomes of both studies were that activity duration distributions had no significant effect on project duration when compared to inaccurate (± 10%) three-point estimation. Therefore, the accuracy of the optimistic time, most likely time and pessimistic time estimations were more important than activity distributions. In addition, one should also note that the significance of probability distribution without considering the accuracy of three-point estimation has not been explained in detail.

Kong, Zhang, Li, Zheng, and Guan (2015) have shown the advantages of Monte Carlo Simulation over CPM and PERT. First, the differences have been explained qualitatively. Then, in addition, a case example was used to show the difference quantitatively. As the final remark, the authors concluded that Monte Carlo provided a more realistic result and therefore, it could be more helpful in risk management, since it was capable of listing the possible project completion dates with their corresponding probabilities. The study summarizes differences between Monte Carlo, CPM and PERT.

Karabulut (2017) also compared the Monte Carlo with PERT. A case example was used for comparison. The duration of the case example was first computed using PERT and then it was again found using Monte Carlo Simulation this time. Then the results of both methods were compared with each other. The author has concluded that Monte Carlo Simulation is an iterative method, which considers all possible outcomes with respect to pre-determined probability distribution functions of the activities. Therefore, Monte Carlo simulation provided a more realistic result when compared to PERT as the conclusion of this study.

Lee, Lee, and Alleman (2018) built up a model to estimate the duration of ultra-critical coal fired power plant projects. PERT and Monte Carlo were used for this model. The pessimistic, most likely and optimistic times were determined according to opinions of experts. After determining the expected time values and sequences of each activity, MS Project was used to find the critical path. After completing these stages, Monte Carlo was used to simulate the schedule 1000 times to find the 85% probability range of the duration using only activities on critical path (critical activities). The outcome of the proposed model was compared with the durations of four completed ultra-critical coal fired power plant projects. The results have shown that the model was accurate, since the durations of the four completed ultra-critical coal fired power plant projects remained in the of 85% probability range. In this study, there were some subjectivities. First, PERT was conducted using beta distribution and Monte Carlo Simulation was performed using triangular distribution. In addition, the authors have referred to the opinions of the experts for the pessimistic, most likely and optimistic durations of the activities. These issues make the model subjective, since the results may change significantly with different distribution types and different estimations of pessimistic, most likely and optimistic durations due to fact that those are the foundation of the proposed model.

Arunmohan and Lakshmi (2018) computed the duration of a project using PERT and Monte Carlo Simulation separately, in order to compare the outcomes with each other. However, the results were not compared with the duration of a completed real project. The authors have concluded that PERT provided optimistic result, since the duration computed using PERT was shorter than the duration calculated using Monte Carlo Simulation. In addition, Monte Carlo also provided the probabilities of the computed duration range. When compared, there was no possibility to complete the project within the estimated duration using PERT according to Monte Carlo Simulation results.

Elaiwi (2018) has explained the concepts of PERT and CPM including their advantages and differences. As the main difference of the paper of Liu (2013), Elaiwi (2018) scheduled a yacht construction project using both PERT and CPM techniques. So that, it was possible to see which technique provides more optimistic results. As the result of this research, PERT came out to be more optimistic when compared to CPM, since the project duration obtained using PERT was shorter than the duration obtained using CPM.

Silvianita, Aprillia, Mulyadi, Citrosiswoyo, and Suntoyo (2018) calculated the duration of a graving dock construction project both using CPM and PERT methods. In addition, the authors have also computed the probability of completing the project within the calculated time of CPM, using PERT. First, the network has been built and the critical path has been identified using CPM. Then the duration of the whole project was calculated again using CPM, which is equal to summation of the durations of the critical activities. Then the total duration of the same project was again calculated using PERT and the critical path identified before. The results of the CPM and PERT were 204 days and 210 days, respectively. After that, assuming the duration of the project was normally distributed, the probability of completing the project within 204 days was computed using PERT with the help of z-Table. The probability was approximately 80%. The outcome of this study was that CPM is a deterministic method, whereas PERT can additionally provide the probability of completing a project within a specified time. If normal distribution is assumed for the total duration of the project, z-value is calculated and z-Table is used to find out the probability in PERT.

CPM, PERT and Monte Carlo methods are the most commonly used methods for project scheduling. CPM is a deterministic method, whereas PERT and Monte Carlo Simulation are probabilistic methods. CPM provide the results as point estimation. On the other hand, it is possible to find the range estimation with corresponding probabilities by PERT and Monte Carlo Simulation. However, Monte Carlo Simulation iterates the project schedule as many times as requested by the user in order to include the many probabilities. PERT also takes advantage of CPM for identifying the critical path in order to use the critical activities to compute the project duration. In PERT, it is assumed that activities follow beta distribution and the total project duration follows normal distribution. On the other hand, one should have determined the statistical distributions of the activities such as triangular distribution, normal distribution, and beta distribution etc. before continuing with Monte Carlo Simulation method. However, non-parametric bootstrap method does not require the user to determine the statistical distributions and correlations of the activities at all. Additionally, sample size is also increased using the bootstrap method, which leads to more realistic results when compared to existing methods. These are the main advantages of non-parametric bootstrap method over PERT and Monte Carlo Simulation. In addition, it is also possible to find the range estimation with corresponding probabilities using bootstrap.

As already introduced in the previous studies mentioned above, it is possible to perform PERT and Monte Carlo in various types of construction projects such as graving dock, offshore platform, multistory buildings and power plant projects in order to estimate the durations. Monte Carlo has been also used to take risks of the activities into account and to allocate the total project time contingency to activities. PERT and Monte Carlo may be used together in a project. The outcomes of PERT and CPM become the inputs of Monte Carlo for simulating. This way, range estimations can be obtained with their probabilities. In this thesis, CPM will be used for determining the critical path and then the non-parametric bootstrap will be performed to obtain the range estimation of the projects with corresponding probabilities, since non-parametric bootstrap does not require any assumption regarding the distributions and correlations of the activities. Therefore, a robust model will be obtained with the usage of bootstrap.

2.3. Bootstrap Sampling Method

Bootstrap sampling method is used for resampling when the data are scarce (i.e. when available data set is small). It is helpful for reproducing the data in hand, especially if there are small data sets with limited number of members. With the usage of bootstrap, the available data are resampled; therefore, the sample estimates for statistical parameters such as standard deviation are expected to be more accurate due to increased sample size. Most significant feature and importance of bootstrap is that it does not require any assumption regarding the probability distributions of inputs such as normal, beta, uniform or triangular distributions. It provides great advantage while modeling for estimation purposes, since those models are developed using the bootstrapped resamples obtained by resampling the available historical data. So that, the developed models come out to be based on the data obtained from previous real projects. Bootstrap technique will be explained in a more detailed manner in chapter 3. The previous studies that have used bootstrap method within the construction management context are investigated below.

Sonmez (2008) performed bootstrap method together with regression analysis in order to predict the construction project costs. The aim of the study was to integrate the bootstrap and regression methods in order to estimate the range of the costs of the construction projects with their corresponding probabilities. The historical data were collected from a contractor and contained information of 20 completed projects. According to that data, there were 11 cost items available that determine the final cost of the whole project. Leave-one-out cross validation method is applied for the validation of the proposed method. One data point is used for testing the result and all the other data points are used for training the model. Then the data point, which was used for testing, is included for training and one of the other data points, which was used for training, is used for testing until all the data points are used for testing for once. So that, the prediction is made for each 20 projects. The data of the 19 projects, which are being used for the training of the model were resampled 1000 times, so that there were 1000 data sets each containing 19 project data. At the beginning, estimators determined the factors affecting the cost items. After all these, regression analysis was performed for every and each 1000 data set containing 19 project data, that also contained 11 cost items in order to find the relation between the cost and independent parameters, which are the affecting factors. As the summary, there were 1000 regression models for each cost item. So that, the range estimates of costs of each cost item of 19 projects were computed along with their probabilities as shown in Table 2.1 below. The data of the remaining project was used as the input to the developed model in order to compare the outcome of the model and the real cost of that project. In addition, mean absolute percent error (MAPE) was also computed to see the difference between the real cost and the result of the model. The real cost of the tested project lied in the range of 90% probability level of the proposed model and MAPE was small enough to be assumed that it was acceptable. So that, the model was found to be providing solutions accurate enough.

Table 2.1. Outcomes found by using regression models together with bootstrap method in Sonmez (2008)

Cost Items	**Probability Level**		
	5%	**50%**	**95%**
Site development	762.743	1.189.386	1.741.066
Foundations and slab on grade	654.092	896.173	1.134.172
Structure	2.290.073	2.659.794	3.329.636
Enclosure	314.165	1.398.060	1.973.105
Interior finishes	4.019.519	4.860.802	5.800.377
Equipment and special construction	232.315	295.305	452.018
Conveying systems	700.761	739.380	804.205
Mechanical	1.613.006	2.177.185	3.163.362
Fire protection	321.291	366.711	426.857
Electrical	1.333.400	1.586.136	1.841.553
General conditions	2.060.346	2.431.586	3.021.058
Total project cost	**16.900.782**	**18.593.888**	**21.153.484**

Tsai and Li (2008) adopted bootstrap method to reproduce the small data set to be used as input to artificial neural network model for pilot run modeling of a conductor supplier's manufacturing system. The historical database was provided by the same supplier and contained 44 data points. The researchers have selected 10 of them randomly to be used at model developing stage and the non-selected data points were used for validation after developing the model. So that, the writers were able to compare the results of the proposed model with the real data. The selected data were resampled using bootstrap procedure 100 times. There were totally 100 resamples obtained by bootstrap. Then they were used to train the artificial neural network and the result is obtained. In addition, only the original 10 data were also used to train the artificial neural network in order to obtain another result for comparison. Those 10 data for input were increased with the increment of five also in order to see the effect of the increase in the number of the "primitive" data on the results. However, as stated above, not all of the historical data were used in all cases. Some of them were left for the validation. Finally, after comparisons, there were two different outcomes of this research. One of them was that the training data sets, which were obtained by bootstrap method, provided more accurate results when compared to the results obtained by training “primitive” data. The other outcome was that the increase in the number of “primitive” data in the training data set did not provide convergence of the results when compared to the results obtained by training the initially selected data. So that, the usage of bootstrap was very helpful for obtaining more accurate results.

Sonmez (2011) have aimed to predict the range of the construction costs with their corresponding probabilities. However, in this research, neural network model was used instead of regression model. Similarly, the historical data collected from a contractor were resampled using bootstrap method. Then the resamples were used as the inputs to train the neural network model. In addition, same validation method was applied as it was applied in the previous study and the obtained results were similar with the previous findings in the end (Sonmez, 2008). The advantages of bootstrap have been shown one more time in virtue of this study, since the range estimates and their probabilities have been computed easily without any assumptions regarding probability of the costs or the correlation among the cost items through bootstrap method in both of the studies (Sonmez, 2008, and 2011).

Hashemi, Mousavi, and Mojtahedi (2011) applied bootstrap as a contributive tool for risk analysis of bridge construction projects. They gathered the risk data through questionnaires and interviews with the experts. Then the collected data were resampled using bootstrap and the resamples were used to conduct the risk analysis using interval risk score technique. This proposed method was validated with a case study, which was a bridge construction project from Iran. As the results of the validation, it was found out that this hybrid method, which was interval risk score together with bootstrap, provided a more precise solution when compared to conducting the interval risk score technique with original sample only (i.e. without taking advantage of bootstrap). The standard deviation of the developed model has come out to be less than the standard deviation found using original sample only. The researchers have also concluded that bootstrap is very useful, where the data size is too small.

Hashemi, Mousavi, Tavakkoli-Moghaddam, and Gholipour (2013) also used bootstrap for risk assessment as Hashemi et al. (2011) for port management projects. The linguistic terms of occurrences probabilities and impacts of risk factors provided by decision makers were converted into numerical values first and then the most important risk factors were determined. So that, they were resampled 1000 times using bootstrap method and confidence intervals were obtained. A case study, which is a real port project in Iran, was used to validate the proposed method. The results were compared with the traditional methods and it was found out that the standard deviation obtained by using bootstrap was much less, which makes bootstrap a more precise method. Another outcome mentioned by the researchers was that bootstrap also provided time saving, since it would be time consuming if all the data were collected from the experts. As the concept and logic, this research was similar to the previous study mentioned above (Hashemi et al., 2011).

Valášková, Spuchľáková, and Adamko (2015) and Dallah (2012) have shown in their studies how to conduct bootstrap method generally. In addition, Dallah (2012) have also shown to use bootstrap together with regression models, which forms the main body of this thesis. The research of Sonmez (2008) is also very helpful in this manner, since he have also combined bootstrap method with regression analysis and additionally he came up with the range estimates of project costs. In this thesis, the objective is also to find the range estimates but of the duration instead.

As a more recent research, Gardner, Gransberg, and Rueda (2017) compared the point estimate and range estimate of the highway construction costs. Bootstrap method was used to obtain the range estimate and additionally, neural network was also used as the common technique for computing both point and range estimating. Historical data were collected from Montana DOT. The collected data were obtained from 189 completed projects in total. Eighty percent (80%) of them were used to train the artificial neural network model, whereas the remaining 38 project data were used to test the accuracy of the outcomes of the models. Besides, the most effective factors of construction cost were found by conducting interviews and literature review. The collected historical data were resampled 100 times using bootstrap. Then, neural network was applied to each resample and the results were ordered from least to greatest in order to find the probabilities of the corresponding costs. On the other hand, point estimates were also computed only using neural network for the sake of result comparison. As mentioned above, the results were compared using the remaining 38 completed project data, which were not used at the training process of the neural network, in order to validate the proposed model. The point estimates were compared with the real costs by calculating the mean average percentage error (MAPE), whereas the range estimates were compared with the real costs in a different way. It was checked whether the real cost fell within the estimated range computed by using bootstrap together with neural network. As the outcome of validation, point estimates did not provide such unsatisfactory results. However, for the point estimates, as the usual problem, it is not easy to find the contingency amount, whereas the range estimates directly provide corresponding probabilities of the results. So that, one can easily find out how much risk was taken for a selected cost value. In addition, the real costs of 35 of the 38 projects fell within the range estimated by the developed model, which might be assumed to provide satisfactory results. However, the accuracy might be improved by including more data in original sample and/or using more resamples produced by bootstrap.

Sonmez (2008, 2011) and Gardner et al. (2017) used bootstrap as a contributive tool for predicting the range estimates of costs of construction projects, whereas Hashemi et al. (2011) and Hashemi et al. (2013) used bootstrap for risk analysis in construction projects and port management projects, respectively. On the other hand, Tsai and Li (2008) adapted this method for pilot run modeling. Bootstrap has many advantages such as it does not require any assumptions regarding the probability distribution such as normal, beta or uniform probability distributions and it also improves the accuracy while reducing the standard deviation and with the usage of bootstrap, the correlations between the inputs and outputs are also taken into account. The main purpose of bootstrap is to resample the available data and find the probability distribution.

However, it has a wide range of usage area. It was used in construction projects, port management and serial manufacturing facility for range estimation as explained above. This thesis study also adapts bootstrap technique for resampling and range estimation of construction project duration instead of project cost. In this manner, bootstrap will be used for a different objective rather than calculating the range estimation of costs, when compared to above mentioned previous studies.

The literature review section was introduced above. To sum up, there exist many studies focused on productivity and scheduling of the projects, which will together form the basis of this thesis. Regression models, CPM and probabilistic methods were the most common techniques applied in previous studies. They are widely used and investigated for many years. However, there are some lacks in the previous researches, such as assumptions of probability distributions, insufficient historical data and point estimation instead of range estimation. This thesis study will provide improvement to the previous studies in terms of these lacks. It will contribute to the body of knowledge by providing a robust model, which performs probabilistic range estimation of project durations together with their probabilities for risk analysis.

In the next chapter, the objective and methodologies of the thesis will be given and the tool will be explained and it will be shown how to use it in detail.

CHAPTER 3

INTEGRATED REGRESSION-BOOTSTRAP SAMPLING SCHEDULING METHOD

3.1. Objective and Methodologies of the Thesis

The objective of this thesis is to develop a regression-bootstrap scheduling method for practical and realistic evaluation of construction schedule risks. The bootstrap method enables a non-parametric estimation without requiring any assumptions about the probabilities and correlations of activity durations. Bootstrap method is also expected to lead to realistic estimates for project duration ranges due to increased sample size. In addition, regression analysis will enable the variations in the activity durations by including the relations between the affecting factors and the productivities of each activity. Therefore, regression analysis is also expected to lead to realistic estimates for project duration ranges due to the inclusion of the variations of each activity. The comparison of the proposed, integrated regression-bootstrap sampling scheduling method and the traditional methods will be provided in chapter 4. However, in this chapter, the method will be explained in detail.

3.1.1. Critical Path Method (CPM)

CPM is developed by Kelley and Walker (1959) and it is the most common method used for scheduling. Although it was developed in 1959, it is still popular. CPM is a deterministic way to estimate the project schedule, therefore it does not provide the probability of the project completion time and gives the results as single point estimations. In order to perform CPM, firstly, the activities of the project should be determined. Then the network should be formed according to the sequence of the activities as shown below in Figure 3.1 as a sample network. After that, the total duration is calculated based on the duration of the longest path. Total project duration is equal to the duration of the longest path according to the logic of CPM. This longest path is called critical path and the activities on critical path are called critical activities. Critical activities have no float, which means that if any of the critical activities delays, then the whole project also delays with the same amount. However, CPM may mislead the decision makers especially during schedule risk analysis, due to the incapability of range estimation and probabilistic approach. It is hard for engineers and project managers to comment on the results computed using CPM.

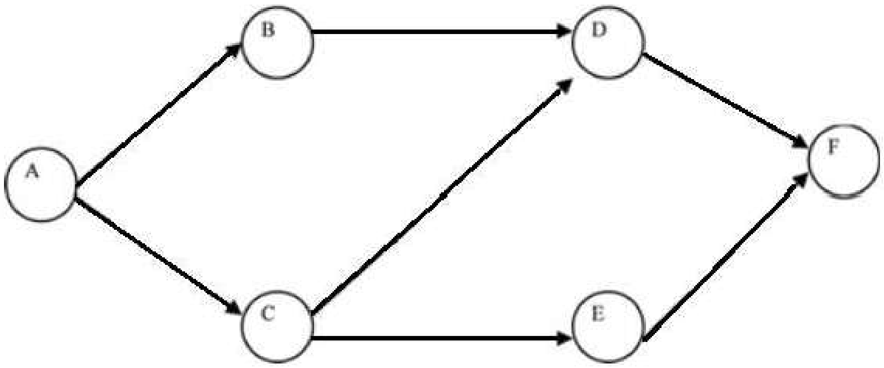
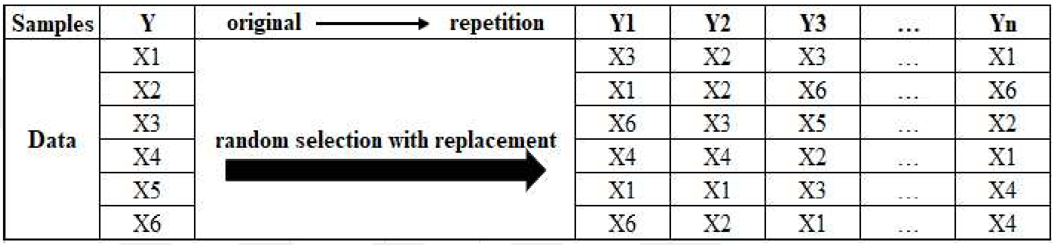


Figure 3.1. Sample network adapted from Lee, Arditi, & Son (2013)

3.1.2. Non-parametric Bootstrap Sampling Method

Bootstrap is a resampling method. Resampling means to create new samples with random selections from the original sample, while all the samples have totally the same number of data points. It is very helpful especially when the available data are scarce.

The bootstrap used in this thesis is named as non-parametric bootstrap method with replacement. As the name implies, it does not require any assumptions regarding the probability distributions of the activity durations such as normal, beta, triangular etc. distribution types, which is a great advantage of non-parametric bootstrap method. In addition, it is also not an easy task to add up the activity durations of different distribution types in order to calculate the total project duration. Although it is possible to add up the different distribution types with simulations, there occurs high range estimation. Non-parametric bootstrap technique eliminates these issues. On the other hand, “with replacement” means that a data (member) can be observed in a sample one time or more than one time and it is also possible that it cannot be observed in one or more samples. It is a completely random process. If n number of data are considered in the original sample, the probability of randomly selecting each data from the original sample is always equal to 1/n while resampling using the bootstrap. The basic principle of resampling the original data using the non-parametric bootstrap with replacement is as illustrated in Figure 3.2 below.



Samples	Y	original → repetition	Y1	Y2	Y3	...	Yn
Data	X1	random selection with replacement	X3	X2	X3	...	X1
	X2		X1	X2	X6	...	X6
	X3		X6	X3	X5	...	X2
	X4		X4	X4	X2	...	X1
	X5		X1	X1	X3	...	X4
	X6		X6	X2	X1	...	X4

Figure 3.2. Basic principle of non-parametric bootstrap with replacement adapted from Valášková et al. (2015)

3.1.3. Regression Models

Regression models are used to unveil the relation between the inputs and output and provide a parametric model. Inputs are called independent variables and output is the dependent variable, since it is determined according to inputs. Regression models can be divided into two different general categories such as linear and non-linear models. Linear models are relatively easier to form and understand. However, in some cases they may become insufficient to explain the relation between the dependent and independent variables. Non-linear models are preferred in those circumstances to obtain equations that are more accurate. The sample formulation of the linear regression models is shown below, where y is the dependent variable, x_1, x_2, x_3, …, x_n are the dependent variables, which are the affecting factors, m_1, m_2, m_3, …, m_n are the coefficients of the affecting factors, b is the constant value and ε is the random error term to consider all the unknown factors that are included in the model. All m and b values are obtained by performing the regression analysis, whereas y and all x values are obtained from the available data and used as inputs in the regression analysis. On the other hand, the expected value of ε is zero. The regression analysis in this case, determines the relation between y and x values.

$$y = m_1x_1 + m_2x_2 + m_3x_3 + \cdots + \ b + \ \varepsilon \tag{3.1}$$

Regression models are also used together with other techniques in previous studies as explained in detail in chapter 2 in order to include the effects of parameters. On the other hand, regression models are incapable of providing range estimates. It should be performed together with a simulation method simultaneously, in order to obtain range estimates while considering the effects of parameters. So that, an integrated and robust method can be developed. Therefore, this thesis aims to integrate non-parametric bootstrap method and regression analysis.

The objective and methodologies of the thesis are covered up to this point. In addition, the methodologies used are explained in a more detailed manner. In the next part, the bootstrap scheduling method developed in this thesis study, will be explained in detail.

3.2. Integrated Regression-Bootstrap Sampling Scheduling Method

The integrated regression-bootstrap sampling scheduling method, which is the objective of this thesis, is developed by combining the afore-mentioned techniques. It integrates the non-parametric bootstrap method and linear regression analysis simultaneously. In addition, it also takes advantage of CPM while calculating the total project duration.

This thesis study proposes a new probabilistic method for scheduling the construction projects and it is named as integrated regression-bootstrap sampling scheduling method. The integrated regression-bootstrap sampling scheduling method is developed in order to conduct the probabilistic schedule risk analyses of construction project durations. Since all the construction projects have a certain budget and deadline and also the duration affects the project costs, it is important to estimate the range of the total construction project durations accurately at the tendering stage for making the go or no-go decisions as explained previously. Incorrect or poor estimations may mislead the decision makers, engineers or project managers, since they can lead to low profits than expected if the contractor decides to bid or lost opportunities if the contractor decides not to bid.

The integrated regression-bootstrap sampling scheduling method is capable of providing the range estimation of the duration of a construction project with its corresponding probabilities. Therefore, it becomes easy for the decision makers to conduct the probabilistic risk analysis, since they are able to see the durations together with their probabilities, which is a great advantage while estimating and making decisions. This advantage comes from the usage of non-parametric bootstrap method with replacement.

The non-parametric bootstrap method is used for resampling; therefore, it provides a range estimation together with the corresponding probabilities. It resamples the original, historical available data and predicts the upcoming/future projects. In addition, it does not require any assumptions or calculations regarding the distribution types of the activities and the projects such as normal, beta or triangular distribution etc. while resampling. So that, it provides a great simplicity for the users, since no information is needed regarding the distribution types. Even though the distribution types of the activities are determined with the statistical methods one by one, it is not easy to add up the durations of different distribution types to find the total project duration. It can be achieved by using simulation methods such as Monte Carlo Simulation; however, there occur high correlations between the activities and these correlations should also be taken into account for a realistic estimation. The non-parametric bootstrap method also eliminates the necessity of calculating and including the correlations between the activities while analyzing the project schedule.

The non-parametric bootstrap method is also useful for analyzing the small data sets, when there are no sufficient data available and the data are scarce, since it resamples the original data set for analysis. Since the non-parametric bootstrap method is used with replacement, the members in the original data set can be observed in a bootstrap sample (resample) one time, more than one time or it is even possible that it cannot be observed in one or more bootstrap samples due to the randomness of the process as explained in section 3.1.2. However, the sizes of the original data set and the bootstrap samples must be equal to each other.

The proposed integrated regression-bootstrap sampling scheduling method uses productivity values of each activity for analysis in order to include the variations within all the activities for obtaining more realistic results. However, the affecting factors of the productivities of each activity should also be included, since the productivity can change according to some factors. Therefore, the proposed integrated regression-bootstrap sampling scheduling method is an integrated method, which integrates the non-parametric bootstrap method with the linear regression analysis. The need for the usage of the non-parametric bootstrap method arises from the advantages explained above. On the other hand, the linear regression analysis is used to include the relations between the affecting factors and the productivities. So that, the variations in the activity durations can be explained and it makes the method more realistic in terms of predicting the project durations and their range. Therefore, the integration of the non-parametric bootstrap method and the regression analysis yields a robust model for accurate and realistic range estimations for probabilistic risk analysis purposes.

When it comes to the traditional methods, one of them is CPM and it is a deterministic method as already explained in the previous section. It provides only a single point estimation without any probability. Therefore, it is difficult for decision makers or engineers to comment on the result obtained by using CPM for evaluating the risks. So that, it is almost impossible to conduct any realistic risk analysis using CPM for the duration of the construction projects, since there may exist many unforeseen difficulties or problems in the construction projects. Therefore, the estimations should come with their probabilities for decision makers to be able to conduct the risk analysis and comment on the results.

PERT is a probabilistic method and it also provides the probability of the estimation. In PERT, it is assumed that the durations of the activities follow the beta distribution (three-point estimation) and the total duration of the project follows the normal distribution in order to provide the probability of the total project estimation. However, these assumptions may not be valid for all the construction projects. Therefore, PERT may mislead the decision makers due to the inaccurate estimations as the result of the wrong assumptions.

On the other hand, Monte Carlo Simulation is a simulation method that requires user to define the statistical distribution types of the durations of the activities before simulating the schedule of the project. It is capable of providing the range estimations of the construction projects with their corresponding probabilities. So that, it is easy for decision makers to see the possible risks of time overrun, since the duration range of the whole project will be available with its probability. However, incorrect assumptions again yield to inaccurate results as it was the case in PERT. The distribution types may be determined with the statistical methods, however adding up the different distribution types while computing the total duration of the project creates high correlations between the activities that should be included in the analysis. However, Monte Carlo Simulation method does not take these correlations into account.

To sum up, the integrated regression-bootstrap sampling scheduling method is proposed to overcome the aforementioned drawbacks of the existing methods to provide more realistic results than the existing methods for the estimation of the total duration and schedule risks of the construction projects. Therefore, it will allow engineers and decision makers to foresee the possible delay risks and comment on the results. So that, the proposed integrated regression-bootstrap sampling scheduling method can be used as a probabilistic risk analysis tool. It is explained in detail, how to use this method, in the upcoming parts of this section. The flowchart of the bootstrap scheduling method is provided in Figure 3.3.

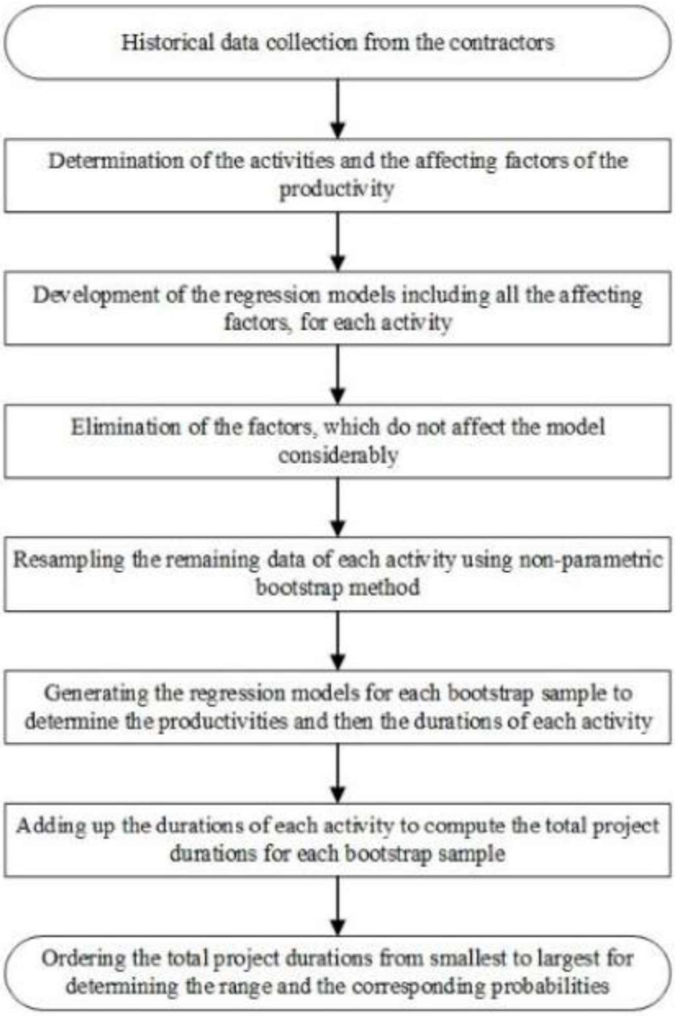

Historical data collection from the contractors
Determination of the activities and the affecting factors of the productivity
Development of the regression models including all the affecting factors, for each activity
Elimination of the factors, which do not affect the model considerably
Resampling the remaining data of each activity using non-parametric bootstrap method
Generating the regression models for each bootstrap sample to determine the productivities and then the durations of each activity
Adding up the durations of each activity to compute the total project durations for each bootstrap sample
Ordering the total project durations from smallest to largest for determining the range and the corresponding probabilities

Figure 3.3. Flowchart of the integrated regression-bootstrap sampling scheduling method

First of all, the historical data should be collected for instance from the contractors for analysis, in order to obtain a realistic result, since artificial data could not present the reality well enough. Therefore, it is important to have the past data of the completed real projects for accurate prediction results, because the accuracy of the estimations also depends on the historical data. After that, the main activities of the project, which will be used in the range estimation of total project duration, such as excavation, concrete pouring and concrete finishing etc. should be identified using the available historical data. After identifying all the activities that will be used in the analysis to estimate the range of the total project durations, the productivity values of each selected activity should be calculated using the Equation 3.2 in order to include the variations within each activity for having a more realistic result at the end, as explained before.

$$Productivity = \frac{Units\ Completed}{Manhours\ Spent} \qquad (3.2)$$

In addition to identifying the activities according to the available data, the affecting factors such as weather conditions, quantities of the activities, crew size and working hours per day etc. of the productivity values of these activities should also be determined according to the available historical data. Even though the historical data do not contain any information regarding the weather conditions such as temperature and humidity, it would be better to obtain these information from the past weather records due to the fact that especially weather temperature may directly affect the productivity as explained in the literature review chapter, since the construction industry is labor intensive. The affecting factors and the activities that will be used for the range estimation may change from one database to another, since they are determined according to the available data in hand. However, as the available data contains more information, the results get more realistic, since more possibilities will be included in the analysis while predicting the range of the total project duration.

After calculating the productivity values of each activity and determining the affecting factors of each productivity value as explained above, linear regression analysis should be performed in order to see the relations between the affecting factors and the productivity values in order to see if the model is parsimonious or not, since parsimonious models perform better (Sonmez & Rowings, 1998).

The linear regression analysis should be conducted for each activity for the same purpose including all the affecting factors and the productivity values in the first model. The obtained model will be in the format as shown in Equation 3.1, where y is the productivity value and x_1, x_2, $x_{3, \ldots,}$ x_n are the affecting factors of the productivities. Besides, m values are the coefficients of the affecting factors and b value is a constant value. These m and b values are computed by conducting linear regression analysis according to the values of productivities and affecting factors and ε is the random error term with an expected value of zero to take all the unknown factors into account that are not included in the model. After that, the factors, which do not considerably affect the model, should be eliminated to obtain a parsimonious model. The factors, which do not affect and reduce the accuracy of the model considerably in its absence, should be eliminated according to two parameters, which are significance level (P value) and coefficient of determination (R^2 value) (Sonmez, 2008). If the significance level of a factor is greater than 0.1 and the elimination of that factor does not significantly reduce the coefficient of determination parameter, which shows the accuracy of the model, then it may be dropped from the model.

The elimination of the factors should be done one by one. For instance, if there are seven affecting factors determined initially, firstly it should be dropped to six factors and again the linear regression analysis should be conducted to exactly see the accuracy of the model and the P values of the remaining affecting factors. If the accuracy of the model does not reduce at a significant level and there still exist one or more factors having P values greater than 0.1, then elimination of the remaining affecting factors should be continued again as one at a time without losing the accuracy of the model significantly and until all the P values of each factor are smaller than 0.1. If the accuracy of the model reduces considerably, then the factor that causes this reduction must be included in the model and if there is another factor having a P value larger than 0.1, it should be dropped from the model and checked again if the accuracy does not reduce significantly.

This process should be applied until obtaining a final model with the factors having P values smaller than 0.1 and with an accuracy level similar to first model, which means parsimonious model. However, it is also important to mention that as the coefficient of determination value gets closer to one, the accuracy of the model increases.

After completing the elimination of the factors of each activity, which do not considerably affect the models, as explained above, the final data sets are obtained for each activity. These remaining data sets are resampled using non-parametric bootstrap method with replacement. The non-parametric bootstrap method is conducted for each activity, so that the bootstrap samples are created for each activity. The non-parametric bootstrap method resamples the original data as many times as required according to the user. The number of the bootstrap samples may be equal to 100 or 1000, it depends on the user. The main rules here are that the numbers of the bootstrap samples of each activity should be equal to each other and the sizes of each bootstrap sample should be equal to the sizes of their original data sets. It means that if the original data set of an activity contains 70 data, then all the bootstrap samples of that activity should also contain 70 data and if the number of bootstrap samples of an activity is 100, then the number of the bootstrap samples of all the remaining activities should also be equal to 100. Other than these important rules, the resampling procedure is completely random. Since it is the non-parametric bootstrap method with replacement, a member of the original data set may appear in a bootstrap sample one time or more than time or even not at all as already shown in Figure 3.2 previously, because with replacement means that the selected members are always kept intact in the original data set and so that the probability of randomly selecting a member is always equal during the resampling process. The main advantage of the non-parametric bootstrap method is that it does not require any assumptions regarding the distribution types of the activities and it manifests itself at this point, since no distributions types are required while completing these processes opposed to Monte Carlo Simulation method. Moreover, it is also useful for analyzing small data sets, since it is capable of resampling the original data as many times as required as explained above in detail.

Linear regression analysis is conducted for each bootstrap sample after finishing the resampling of each activity in order to determine the relations between the productivity values and the affecting factors of each bootstrap sample. Therefore, the number of the regression models of an activity is equal to the number of the bootstrap samples. After finding the models of each bootstrap sample of each activity conducting the linear regression analysis, the new values of the same affecting factors of the predicted project should be entered to the models with the same order as conducting the regression analysis for calculating the productivity values of each bootstrap sample. It is important to enter the new values of the same affecting factors, which belong to the predicted project, in the same order as the first one. Otherwise the model cannot provide the correct results. After computing the productivity values, they are converted to durations using again the data of the predicted project. The required data for converting the productivity to duration is the crew size, quantity and working hours per day of each activity. Here, there is no order such as affecting factors, since the duration will be calculated according to the Equation 3.2. After converting the productivity values to the durations, there are duration values of each activity as many as the number of the bootstrap samples of each activity. For instance, if an activity has 100 bootstrap sample, there also exist 100 duration values and this is also valid for the other activities, since the number of the bootstrap samples of all the activities are equal to each other.

After having the durations of each bootstrap sample of each activity, the total project durations are computed by adding up these duration values of each bootstrap sample of each activity, since the relation between the activities are always finish-to-start and an activity cannot start before its predecessor is completed, so that there are also no lags. Finally, there are total project durations as many as the number of the bootstrap samples, which is 100 if the same example above is considered. Ordering these project durations from smallest to largest value gives the distribution of the total project duration and so that, the range estimation is obtained with its corresponding probability. For instance, if 100 total project durations are considered again, after ordering them from smallest to largest value, the fifth value will be the duration that the project will be completed with 5% probability and the ninety-fifth value will be the duration that the project will be completed with 95% probability. So that, the range between these two duration values becomes the 90% probability range.

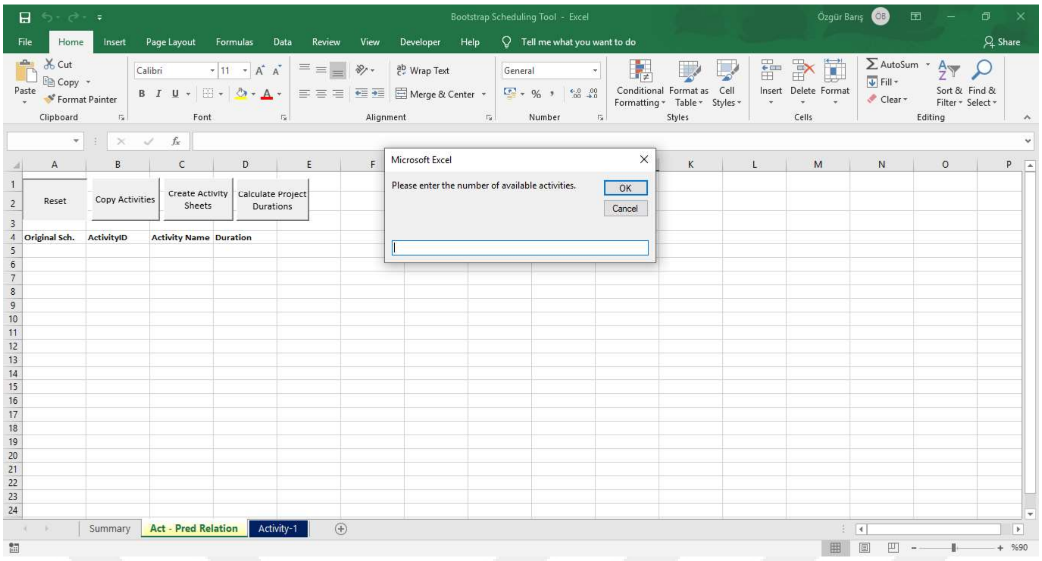
The proposed integrated regression-bootstrap sampling scheduling method is explained in this section. In addition, a tool is developed for performing the integrated regression-bootstrap sampling scheduling method and it will be explained in the next section, so that the method could also be understood better with an example.

3.3. Bootstrap Scheduling Tool

The tool is developed to perform probabilistic risk analysis of project duration. It adapts an integrated method of non-parametric bootstrap and regression analysis. There is no need for any distribution assumption regarding the activity durations and determining the correlation between the activities, because of the advantages of the non-parametric bootstrap method as mentioned before. In addition, it requires user to enter the productivity values of the activities, therefore it also includes the variation of the production rates of the activities, which leading to more realistic results. The predecessor relations between the activities are always finish to start and it is assumed that an activity can start if all of its predecessors are completed.

In the end, the tool provides the range of the total project durations with their corresponding probabilities. It is also important to mention that this tool is developed using MS Excel, therefore it can be used by everyone easily. The usage of this tool is explained below in detail.

After the tool is opened using MS Excel, the first thing to do is to press the reset button, which takes place at the top left-hand corner in “Act – Pred Relation” sheet. After pressing the reset button, there will appear a box as shown in Figure 3.4, which asks for the number of available activities. The user should enter the number of activities available in its database. The upper limit of the activity number is the available memory of the computer, because in the further stages, the tool will create as much new sheets as the number of the activities and the limitation for the number of the sheets in MS Excel is the available memory of the computer.



Microsoft Excel

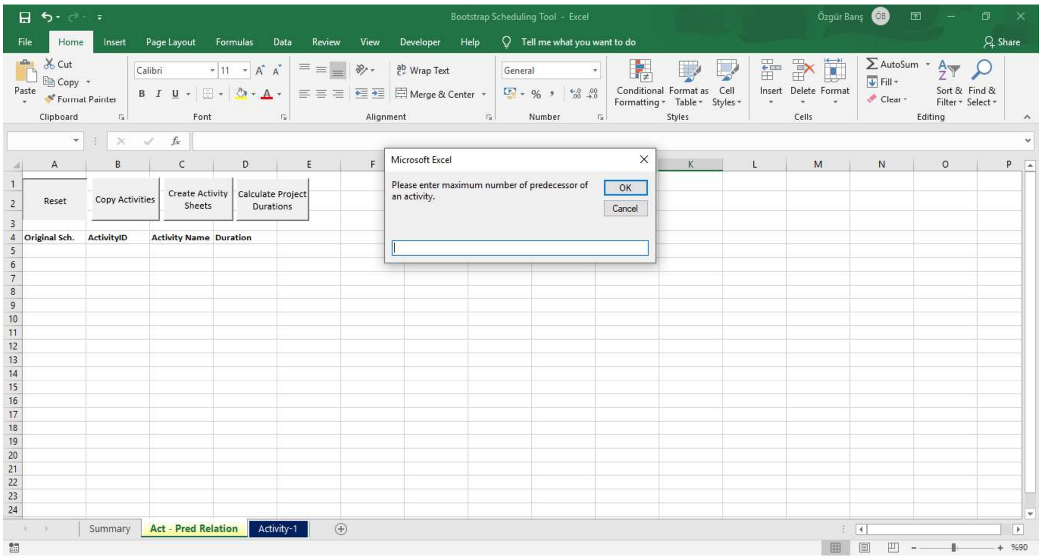
Please enter the number of available activities.

OK

Cancel

Figure 3.4. Dialog box that requires user to enter the activity number

After entering the activity number, it asks for the maximum number of the predecessors of an activity in a new dialog box as shown in Figure 3.5. There are two reasons for asking the maximum predecessor number of an activity. One of them is for the VBA code, since it is really helpful to know the boundaries while coding. The other reason is to highlight the cells, which the user will use for entering the data.



Microsoft Excel

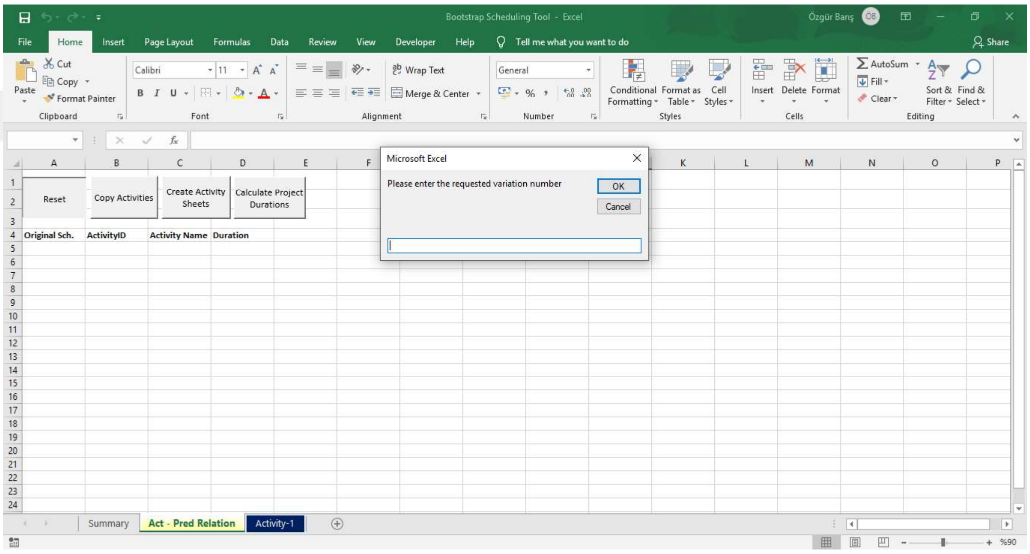
Please enter maximum number of predecessor of an activity.

OK

Cancel

Figure 3.5. Dialog box that requires user to enter the maximum predecessor number of an activity

After entering the maximum predecessor number of an activity, there appears a new dialog box, which asks for the variation number as given in Figure 3.6. Variation means the bootstrap sample. As already explained, the tool will perform non-parametric bootstrap and so that, it will resample the original data set as much as required by the user. In this thesis, the number of bootstrap samples will be 100. However, it is possible to create much more bootstrap sample, but one should note that creating more bootstrap samples will slow down the operating speed of the tool and it will take more time to get the results.

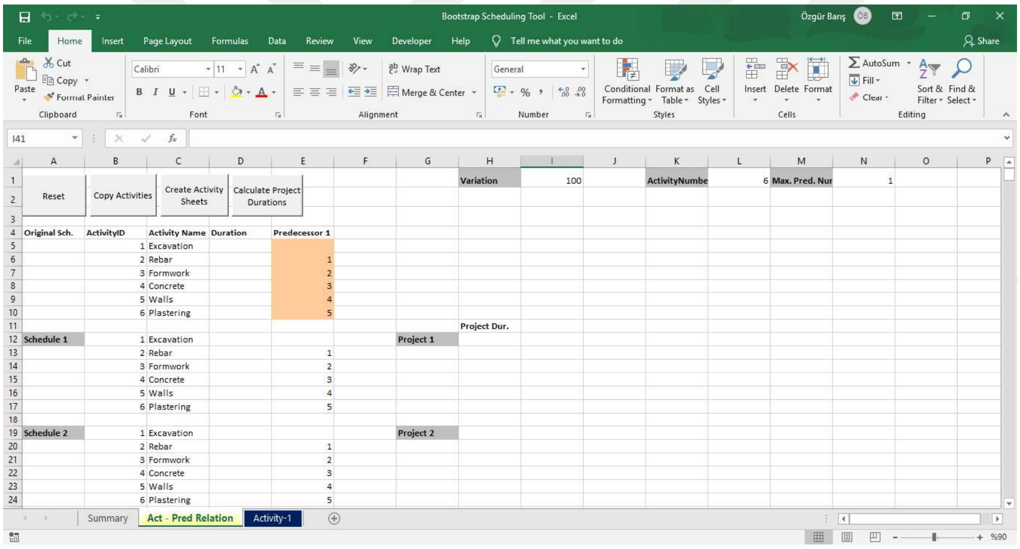


Please enter the requested variation number

Original Sch.	ActivityID	Activity Name	Duration

Figure 3.6. Dialog box that requires user to enter the variation (bootstrap sample) number

After entering the variation number, there will appear a message box that states that the data should be entered to the colored cells. At this stage, only predecessor data are required to be entered to the cells. At this point, the most important thing is to enter the predecessors of the activities as activity IDs instead of activity names. The tool works based on the IDs and in the case that an activity has no predecessor, the predecessor cell of that activity should be left empty as it is illustrated in Figure 3.7. The previously asked numbers are automatically entered to the cells again as given in Figure 3.7. For instance, the required number of bootstrap samples is 100, the activity number is six and the maximum number of the predecessors of an activity is one. It is not a must, but the names of the activities may also be entered under the "Activity Name" heading for the sake of better understanding, so that it will be possible to easily see which activity ID belongs to which activity. It is included just to avoid any confusion for the users. For the cells under the duration heading, the user has nothing to do at this stage and it should be left blank. The cells below the duration heading will be filled in beginning from Schedule 1 part, by the tool itself after the computations. The tool also created the headings such as Schedule 1, Schedule 2, Project 1, Project 2 etc. up to the requested variation (bootstrap sample) number. So that, the last ones are Schedule 100 and Project 100 in this example. Schedule heading stands for the durations of each activity, whereas Project heading is for the total project durations.

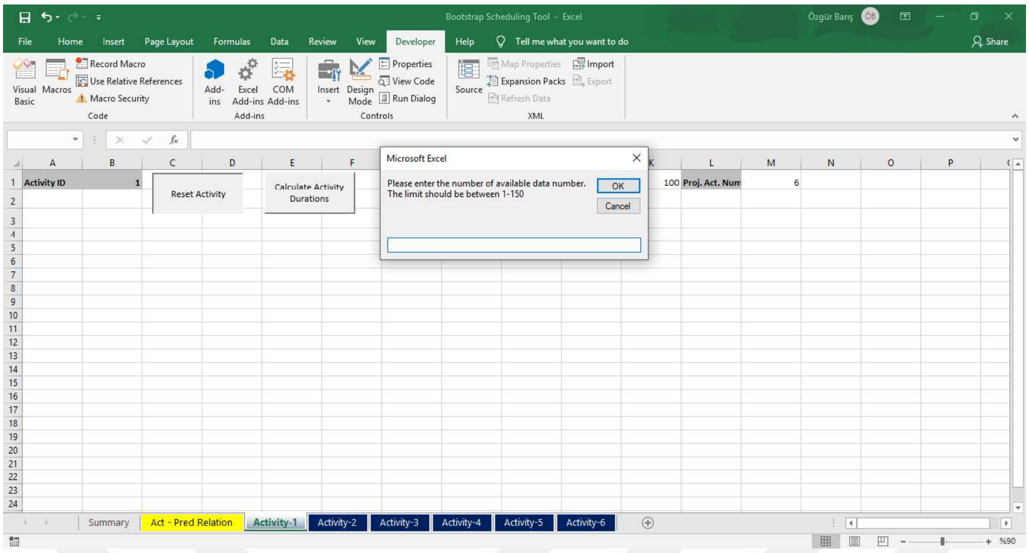


	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1								Variation	100		ActivityNumbe	6	Max. Pred. Nur	1
2														
3														
4	Original Sch.	ActivityID	Activity Name	Duration	Predecessor 1									
5		1	Excavation											
6		2	Rebar		1									
7		3	Formwork		2									
8		4	Concrete		3									
9		5	Walls		4									
10		6	Plastering		5									
11								Project Dur.						
12	Schedule 1	1	Excavation				Project 1							
13		2	Rebar		1									
14		3	Formwork		2									
15		4	Concrete		3									
16		5	Walls		4									
17		6	Plastering		5									
18														
19	Schedule 2	1	Excavation				Project 2							
20		2	Rebar		1									
21		3	Formwork		2									
22		4	Concrete		3									
23		5	Walls		4									
24		6	Plastering		5									

Figure 3.7. The view of the tool after copying the related parts from the original sample

After entering the predecessors and if required, activity names to the cells, "Copy Activities" button should be pressed. It fills the other Schedules as it is in the original sample, up to the variation number, which is 100 again. The copied sections are activity IDs, activity names and predecessors as shown in Figure 3.7. After copying that sections, "Create Activity Sheets" button should be pressed. This button will create as many activity sheets as equal to the activity number entered by the user at the beginning and all the activity sheets will be created automatically by the tool. The requested number of bootstrap sample (variation) and the total activity number of the project are also given in each activity sheet and all are the same. The independent variable number and the project number will be filled by the tool afterwards and those numbers are activity specific. It means that they can change from one activity to another or they may also be the same. There is no rule regarding these two values. So that, the activities are not interdependent in terms of the independent variable number and the project number.

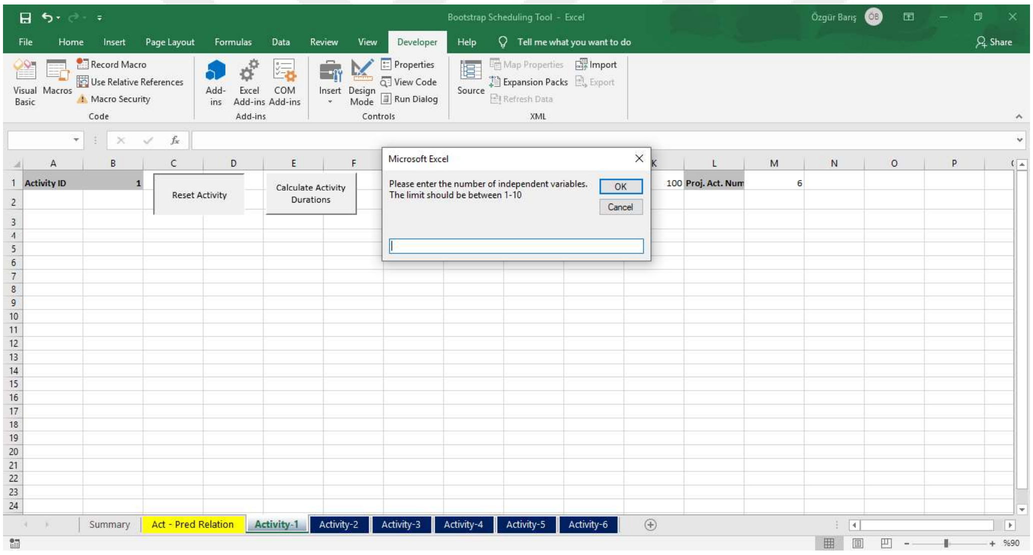
In the activity sheets, firstly, “Reset Activity” button should be pressed in order to start entering the available data. After pressing the “Reset Activity” button, there appears a dialog box asking for the available data number of that activity as shown in Figure 3.8. The user should enter the number of available data of that specific activity. This could also be named as the number of observations of the activity. This number could be different for each activity; therefore, it will be entered for each activity in its own activity sheet. The upper limit of the observations that can be entered to the tool is given as 150. However, it is mostly due to the run time of the tool. As the number of observations and the independent variables of an activity increase, the run time also increases. Therefore, the limit of the data number is stated as 150. On the other hand, after entering the number of available data (observations), there appears another dialog box asking for the number of independent variables this time, as illustrated in Figure 3.9. There is another limitation for the number of the independent variables being ten at most. This limitation is also stated due to the run time concern as mentioned before.



Microsoft Excel

Please enter the number of available data number.
The limit should be between 1-150

Figure 3.8. Dialog box that requires user to enter the available data (observation) number of the activity



Microsoft Excel

Please enter the number of independent variables.
The limit should be between 1-10

Figure 3.9. Dialog box that requires user to enter the independent variable number of the activity

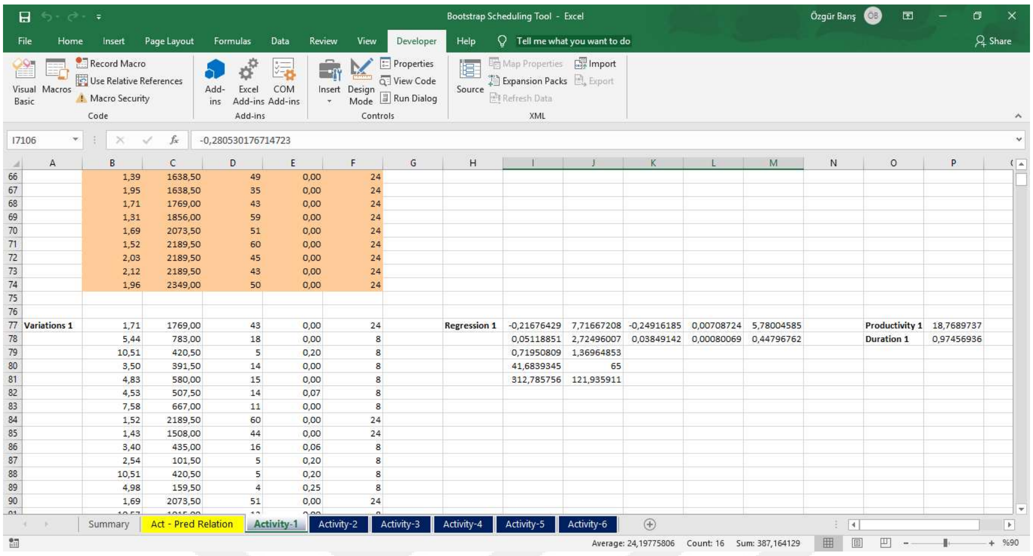
After entering the requested information to the dialog boxes, the requested data should be entered to the colored cells. The data that will be entered to the cells are the productivity values of each observation, the values of the affecting factors of those productivities such as quantity, weather temperature and crew size etc. for finding the relation between the productivity and the affecting factors by using linear regression analysis. The order of the independent variables must be the same for an activity. It means that, same type of independent variables (affecting factors) should be entered to the same column in an activity sheet. For instance, if the weather temperature is the independent variable 1 and humidity is the independent variable 2 for an activity, then this should not be changed while entering the data to the cells of that activity.

In addition, number of workers, working hours per day and quantity of the activity of the predicted project should also be entered by the user in order to convert the productivity value to the duration, since the formulation of the productivity is given as in Equation 3.2. So that, the formulation of duration (days) becomes:

$$Duration\ as\ Days = \frac{Units\ Completed}{Number\ of\ workers * Productivity * Working\ hours\ per\ day} \qquad (3.3)$$

Furthermore, before converting the productivity values to the durations, the tool should calculate the productivity values of the activity of the predicted project using the linear regression formula determined again by the tool itself using the bootstrap samples. Therefore, new values of the independent variables, which belong to the activity of the predicted project, should also be entered by the user in the same order as explained before. So that, in summary, the tool will use the observed data to find the relation between the productivity values and affecting factors by linear regression analysis. Then, it will use this relation and the future values of the independent variables to find the new productivity values. After that, the tool will use these new productivity values and future number of workers, working hours per day and quantity of the activity to calculate the duration for each bootstrap sample step by step. This process is the same for all of the activity sheets created in Excel. Therefore, only one of the activities will be shown as the sample and then continued with the calculation of the project duration.

After entering the available data to the colored cells, “Calculate Activity Durations” button should be pressed. This button runs both the non-parametric bootstrap with replacement and then linear regression analysis for each bootstrap sample in succession. Since the variation number is 100 in this example, this button creates 100 resamples from the original data and runs regression analysis once for each resample totaling 100 times. There will occur some results in activity sheet due to the regression analysis as illustrated in Figure 3.10, after pressing the button.



I7106 | -0,280530176714723

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
66		1,39	1638,50	49	0,00	24										
67		1,95	1638,50	35	0,00	24										
68		1,71	1769,00	43	0,00	24										
69		1,31	1856,00	59	0,00	24										
70		1,69	2073,50	51	0,00	24										
71		1,52	2189,50	60	0,00	24										
72		2,03	2189,50	45	0,00	24										
73		2,12	2189,50	43	0,00	24										
74		1,96	2349,00	50	0,00	24										
75																
76																
77	**Variations 1**	1,71	1769,00	43	0,00	24		**Regression 1**	-0,21676429	7,71667208	-0,24916185	0,00708724	5,78004585		**Productivity 1**	18,7689737
78		5,44	783,00	18	0,00	8			0,05118851	2,72496007	0,03849142	0,00080069	0,44796762		**Duration 1**	0,97456936
79		10,51	420,50	5	0,20	8			0,71950809	1,36964853						
80		3,50	391,50	14	0,00	8			41,6839345	65						
81		4,83	580,00	15	0,00	8			312,785756	121,935911						
82		4,53	507,50	14	0,07	8										
83		7,58	667,00	11	0,00	8										
84		1,52	2189,50	60	0,00	24										
85		1,43	1508,00	44	0,00	24										
86		3,40	435,00	16	0,06	8										
87		2,54	101,50	5	0,20	8										
88		10,51	420,50	5	0,20	8										
89		4,98	159,50	4	0,25	8										
90		1,69	2073,50	51	0,00	24										

Average: 24,19775806 Count: 16 Sum: 387,164129

Figure 3.10. Sample regression results

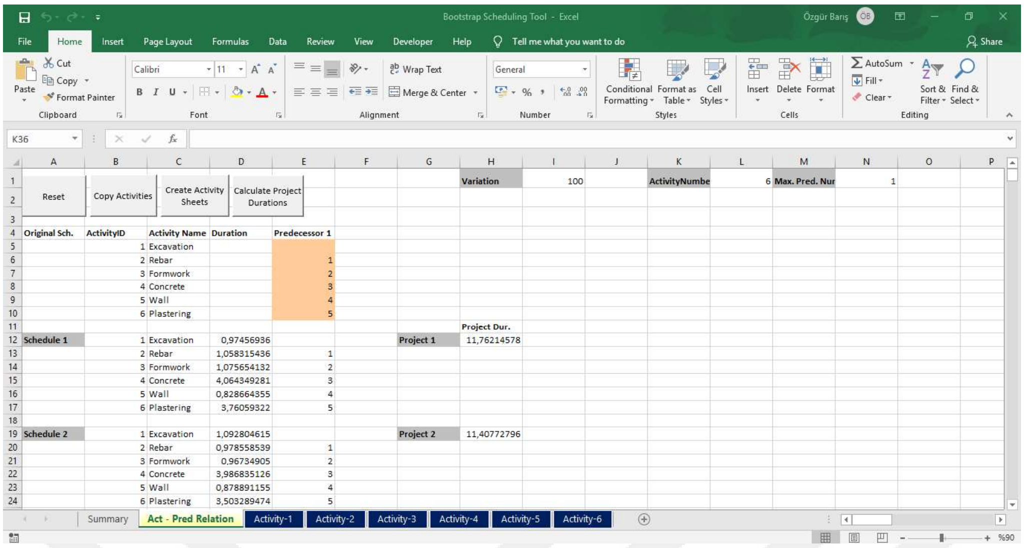
The legend for the appeared numbers is as provided in the Table 3.1. The most important parameters are m values, b value and R^2 (coefficient of determination) value. All m values represent coefficients for affecting factors. The order of the m values is the opposite of the order of the independent variables (affecting factors). For instance, if four independent variables are considered in the tool, the coefficient of the first independent variable is the last m value, which is given as m_1 and the coefficient of the last independent variable is the first m value, which is given as m_n in the Table 3.1. The b value is the constant value and R^2 value is the coefficient of determination, which is important to understand how well the regression model fits to the values (i.e. how well the created regression equation can approach the y value, which is the productivity in this case). R^2 value changes from zero to one. The closer R^2 value to one the better the model fits.

Table 3.1. The legend for the regression parameters

	A	B	C	D	E	F
1	m_n	m_{n-1}	…	m_2	m_1	b
2	se_n	se_{n-1}	…	se_2	se_1	se_b
3	r_2	se_y				
4	F	d_f				
5	SS_{reg}	SS_{resid}				

Using the regression model found by the tool itself, the tool also calculates the new productivity values of each bootstrap resample using the future independent variable values as stated previously. Furthermore, after calculating the new productivity values, it also computes the durations of each bootstrap resample as days and also copies them to the “Act – Pred Relation” sheet as the preparation for the calculation of the total project duration. The process explained in detail for the activity sheet up to this point is also valid for the other activity sheets. The same steps should be applied in each activity sheet and the tool will again automatically copy their durations to the “Act – Pred Relation” sheet, too.

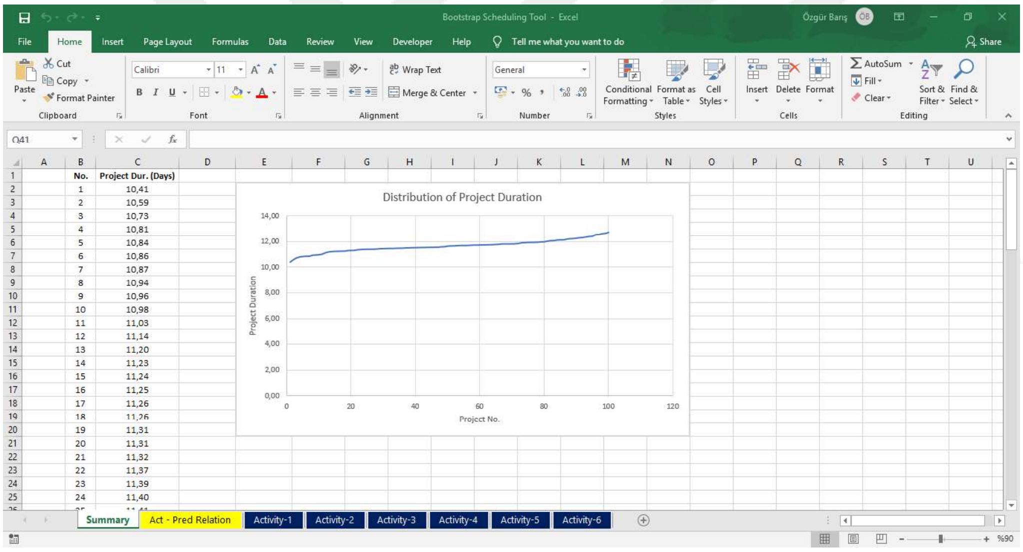
After completing the same steps for other activities, “Calculate Project Durations” button should be pressed in order to have the total durations of all the bootstrap samples. This button calculates the total durations and write them down to the cells next to Project 1, Project 2 etc. accordingly, as shown in Figure 3.11. At the same time, it also copies the durations to the “Summary” sheet, order them from minimum to maximum and so that, a graph showing the distribution of the project durations is drawn as illustrated in Figure 3.12. This distribution shows that there is 80% probability that the project will be finished in 12 days. The graph takes the values up to the 5000th row of the “Summary” sheet.



Variation	100	ActivityNumbe	6	Max. Pred. Nur	1

Original Sch.	ActivityID	Activity Name	Duration	Predecessor 1		
	1	Excavation				
	2	Rebar		1		
	3	Formwork		2		
	4	Concrete		3		
	5	Wall		4		
	6	Plastering		5		
						Project Dur.
Schedule 1	1	Excavation	0,97456936		Project 1	11,76214578
	2	Rebar	1,058315436	1		
	3	Formwork	1,075654132	2		
	4	Concrete	4,064349281	3		
	5	Wall	0,828664355	4		
	6	Plastering	3,76059322	5		
Schedule 2	1	Excavation	1,092804615		Project 2	11,40772796
	2	Rebar	0,978558539	1		
	3	Formwork	0,96734905	2		
	4	Concrete	3,986835126	3		
	5	Wall	0,878891155	4		
	6	Plastering	3,503289474	5		

Figure 3.11. "Act - Pred Relation" sheet after calculating the project durations



No.	Project Dur. (Days)
1	10,41
2	10,59
3	10,73
4	10,81
5	10,84
6	10,86
7	10,87
8	10,94
9	10,96
10	10,98
11	11,03
12	11,14
13	11,20
14	11,23
15	11,24
16	11,25
17	11,26
18	11,26
19	11,31
20	11,31
21	11,32
22	11,37
23	11,39
24	11,40

Figure 3.12. "Summary" sheet after finishing the risk analysis

In this chapter, the tool has been explained and it has been also shown how to use the tool with a sample project together with the related screenshots. In the next chapter, two case studies will be demonstrated using the tool and the results will be compared to the results of PERT and Monte Carlo Simulation methods along with the actual durations.

CHAPTER 4

CASE STUDIES AND COMPARISON OF THE RESULTS

4.1. First Case Study of Construction Project Scheduling using the Integrated Regression-Bootstrap Sampling Scheduling Method

The data used for this case study is the same as the data used in the research of Sonmez and Rowings (1998). Only concrete activities were included in Sonmez and Rowings (1998) for productivity calculations and formwork and pouring activities are used in this case study for scheduling.

The project number in the database is four, each consisting of two activities (formwork and concrete pouring). These four projects belong to building projects of a contractor, which were built in Iowa, USA between the years 1992 and 1994. The data consist of weekly records, since the contractor updated its database on a weekly basis. Therefore, the weekly average values of the affecting factors are used. In addition, the number of weekly data points is 112 for concrete pouring activity and 76 for formwork activity. For the regression models, again the same models are chosen as Sonmez and Rowings (1998), which are RF5 for formwork and RP5 for concrete pouring in that paper. So that, the affecting factors (independent variables) involved in RF5 are quantity (q) and crew size (n), whereas quantity (q), crew size (n), temperature (t), walls over 2,44 m (w = 1 for the walls over 2,44 m and otherwise w = 0), concrete pump (u = 1 when concrete pump is used and otherwise u = 0) and quantity concrete pump interaction term (qu) for including the concrete quantity for which the concrete pump was used are included in RP5.

Leave-one-out cross validation method will be applied while predicting the duration ranges of each project. In leave-one-out cross validation method, one data point is used for testing the result and all the other data points are used for training the model.

Then the data point, which was used for testing, is included for training and one of the other data points, which was used for training, is used for testing until all the data points are used for testing for once. So that, in this case study, it is possible to conduct range estimations totally four times being for different projects at each time, since there are totally four projects available. It means that each time three different projects are used for generating the bootstrap resamples and finding the regression coefficients in order to predict the duration range of the fourth project, which is the remaining one and not used in the analysis for predicting the duration range. For the case study, the project names are 1, 2, 3 and 4. Firstly; projects 2, 3 and 4 are analyzed to predict the duration range of project 1, then projects 1, 3 and 4 are analyzed to predict the duration range of project 2 etc. until the duration ranges of each project are predicted. So that, four predictions have been done in total for the leave-one-out cross validation. In addition, the two activities are linked as finish to start (fs) as shown in Figure 4.1 and the formwork activity is the predecessor of the concrete pouring activity for all the projects.

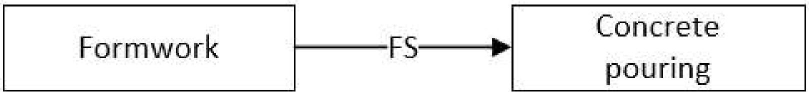


Figure 4.1. The relation between the activities

The number of the bootstrap resamples are 100 for each activity and for each prediction, while conducting all the range estimations. So that, there are also 100 regression models for each activity, since the regression models are formed for each bootstrap resample. The average values of the affecting factors (quantity (q), crew size (n), temperature (t), walls over 2,44 m (w), concrete pump (u) and quantity concrete pump interaction term (qu) of each activity of the predicted project are used with the regression models found by the tool in order to compute the productivity values of each activity of the predicted project. After finding the productivity values, the average crew size, average working hours per day and total quantity of each activity of the predicted project are used to convert the productivity values to the durations (days) as already explained before. Therefore, there exist 100 duration values for each activity. Finally, adding up those duration values yields to total project duration, since the activities are linked as finish to start and concrete pouring cannot start before the formwork has been completed. So that, there are 100 duration values for the whole project. Ordering these values from smallest to largest gives the distribution of the project duration. Since the total number of the duration values are 100, fifth value is the duration that cannot be exceeded with 5% probability and ninety-fifth value is the duration that cannot be exceeded with 95% probability. Therefore, the 90% probability range lies within these two values.

First of all, project 1 is predicted using the proposed tool. The data of the remaining projects are used for generating the bootstrap resamples and regression models for each bootstrap resample. Then, the same procedure is applied for each project to predict the durations as explained above and the distribution graph of each project are provided in Figure 4.2, Figure 4.3, Figure 4.4 and Figure 4.5, respectively.

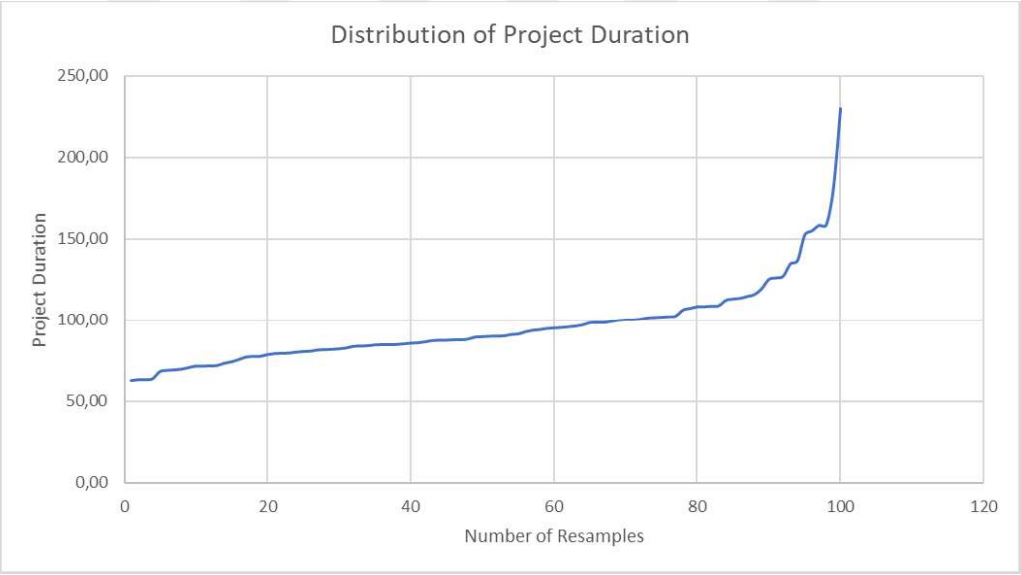


Figure 4.2. Distribution graph of project 1

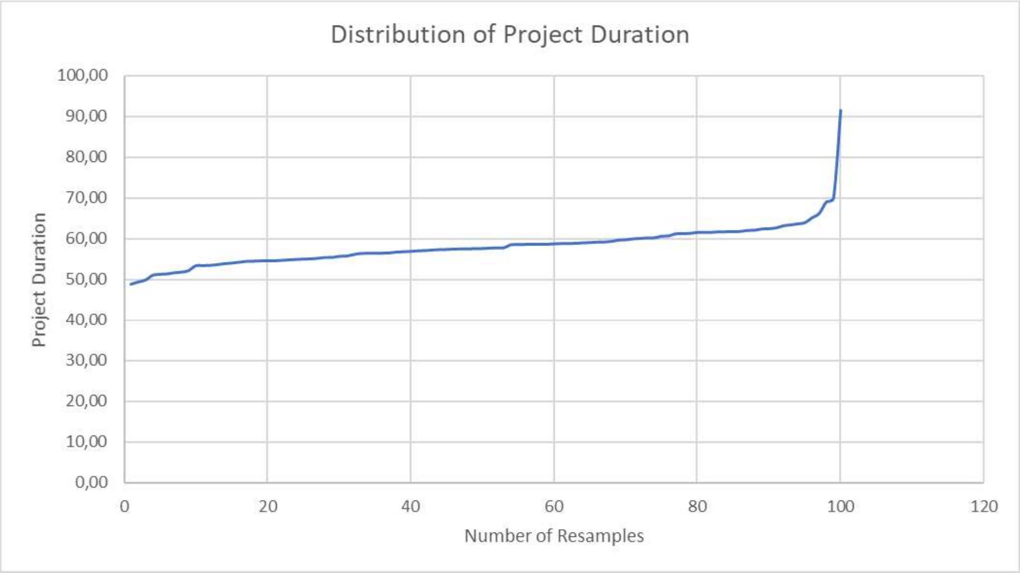


Figure 4.3. Distribution graph of project 2

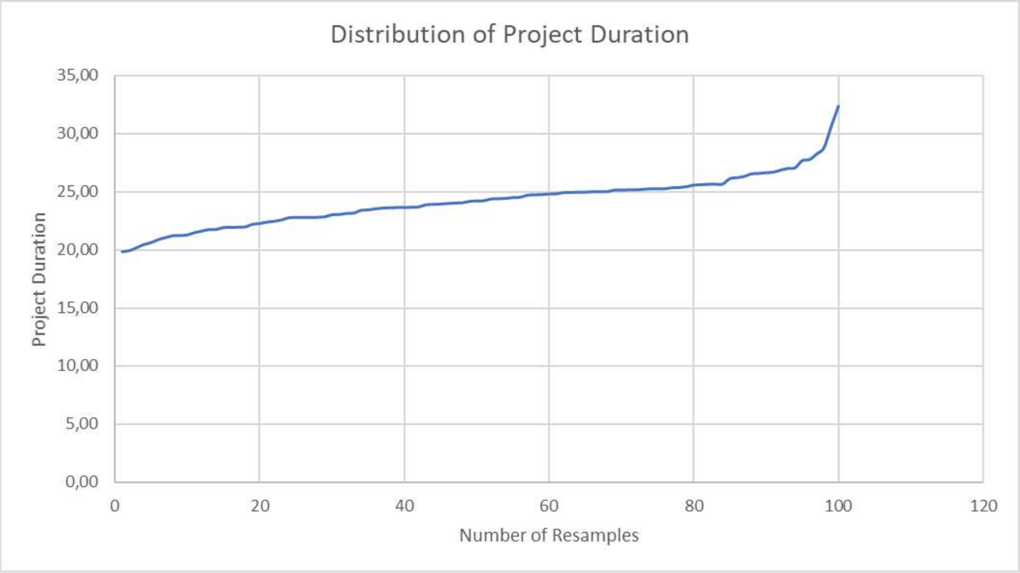


Figure 4.4. Distribution graph of project 3

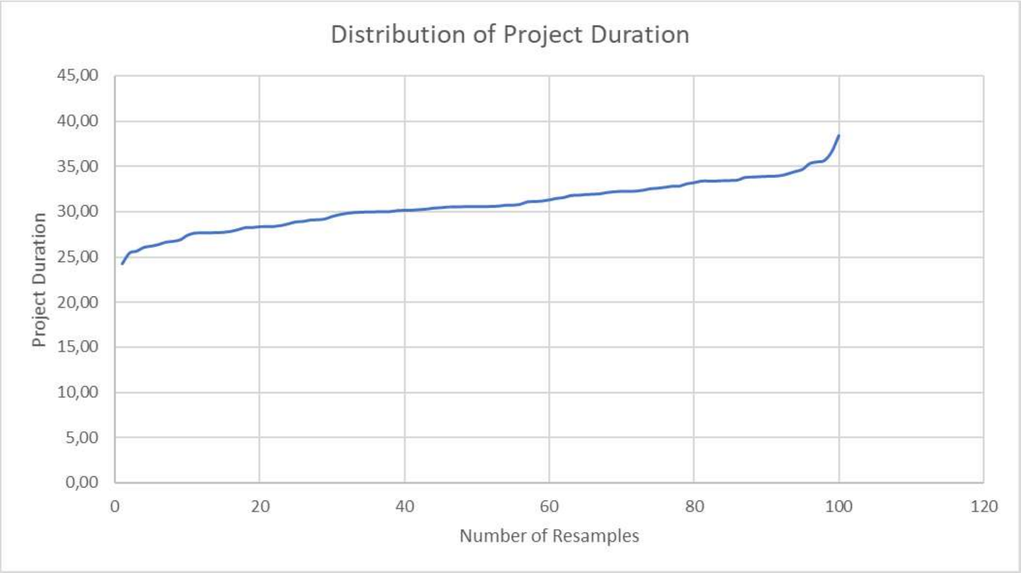


Figure 4.5. Distribution graph of project 4

In addition to the distribution graphs given above, the 5%, 95% and 50% probability values and the actual duration values of each project are provided in the Table 4.1. The differences between actual durations and 50% probability duration values estimated by the bootstrap scheduling tool can also be observed in the same table.

Table 4.1. Differences between the actual durations and range estimations of the tool

	Project 1	**Project 2**	**Project 3**	**Project 4**
Actual Duration (days)	89	69	23	33
%5 Probability Value (days)	69	51	21	26
%95 Probability Value (days)	153	64	28	35
%50 Probability Value (days)	90	58	24	31
Actual - %50 Probability Difference (days)	1	11	1	2

As it can be seen in Table 4.1, the tool predicts the project durations considerably well. Only the actual duration of one of the four projects lies out of the 90% probability range. However, if the whole range estimation of the tool for that project is considered, the actual duration of project 2 also lies within the whole range estimated by the tool as illustrated in Figure 4.3. On the other hand, the actual durations of all the projects are close enough to the estimated durations of the tool corresponding to 50% probability level. This is another indicator of the performance of the proposed method.

In this section, a case study is demonstrated using four different projects and the results are also provided in order to show the performance of the tool individually. However, in the next section, the results of the tool will be compared to the results of Monte Carlo Simulation and PERT using the same four projects.

4.1.1. Comparison with the Existing Methods

The same four projects are used to compare the prediction performance of the bootstrap scheduling tool to the prediction performances of Monte Carlo Simulation and PERT, which are the existing classical methods. 90% probability ranges (the range between the durations of 5% and 95% probability levels) and the durations of 50% probability levels are considered for the comparison along with the actual durations. Firstly, the process and the results of PERT will be explained and then they will be provided for Monte Carlo Simulation method.

4.1.1.1. Analysis with PERT

PERT is a statistical method and developed by U.S. Navy for scheduling project (Salas-Morera, Arauzo-Azofra, García-Hernández, Palomo-Romero, & Ayuso-Muñoz, 2018). It is used together with critical path method (CPM), since the critical path assumption also takes place in PERT. One path with the longest duration also determines the total project duration, as it is the case in CPM. However, the difference is that each activity has its own probability distribution. The probability distribution of the activities is beta distribution in PERT and the graph of PERT-beta distribution is shown in Figure 4.6 below. The duration of each activity in PERT is called expected time and computed using three parameters, which are namely, optimistic time, most likely time and pessimistic time. It is also possible to find the standard deviation and variance. Variance is the square of the standard deviation. The equations of PERT are given below with the detailed explanation.

Total project duration is found by summing up the expected times of the activities on the critical path and assumed to have a normal distribution. Similarly, variance of the whole project is also equal to the sum of variances of the critical activities. Then the standard deviation of the project is calculated by taking the square root of the project variance. So that, the probability of the project to be completed in a pre-determined time can also be found using the z-Table with the assumption of normality. In addition, it is also possible to find the duration corresponding to a pre-determined probability using the z-Table, which is the technique used to calculate the durations with PERT in this thesis.

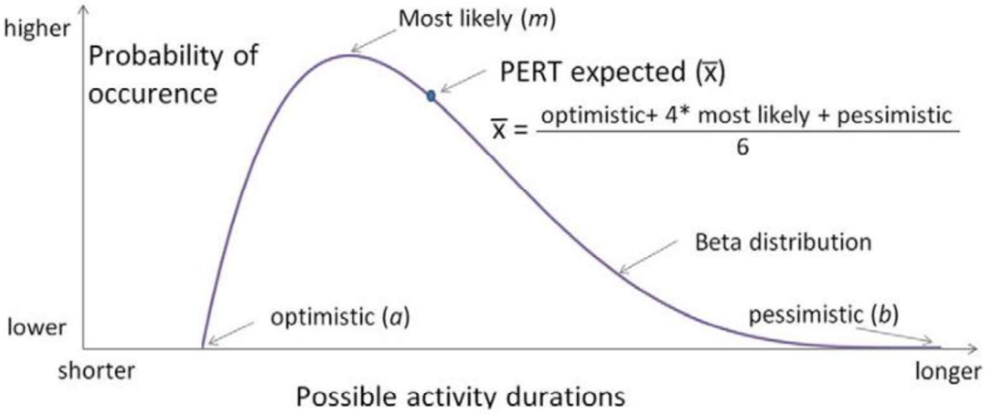


Figure 4.6. PERT beta distribution graph adapted from Hajdu & Bokor (2016)

As mentioned, PERT has a beta distribution, which indicates the three point estimation, namely pessimistic, most likely and optimistic durations (Miklós Hajdu, 2013). In a data set, the worst case can be selected for pessimistic estimation, the best case can be selected for optimistic estimation and average value can be selected for most likely estimation (Miklós Hajdu & Bokor, 2016; Plebankiewicz, Juszczyk, & Malara, 2015). Since the productivity values are used for the prediction of the project durations in order to include the variabilities in the activities, the smallest productivity value is used for the pessimistic estimation, the greatest productivity value is used for the optimistic estimation and the average value of the productivity is used for the most likely estimation. These productivity values belong to the activities of the predicted project and this is applied to each activity. After taking the production values as stated, the total quantity, average crew size and average working hours per day of each activity of the predicted project are also used in order to convert the productivity values to the durations using the Equation 4.1.

$$Duration\ as\ Days = \frac{Total\ Units\ Completed}{Crew\ Size * Productivity * Working\ hours\ per\ day} \tag{4.1}$$

After using this formula, the smallest productivity value gives the pessimistic duration, the greatest productivity value gives the optimistic duration and the average productivity value gives the most likely duration of the activity of the predicted project. After calculating the pessimistic, optimistic and most likely durations of each activity, the below equations, where t_e is expected time, t_o is optimistic time, t_m is most likely time, t_p is pessimistic time and σ is the standard deviation, should be used to calculate the total estimated time and total standard deviation of the whole project. Although the expected time of a project is equal to the summation of the expected times of each activity on critical path, the standard deviation of a project is equal to the square root of the variance of that project and the variance of a project is equal to the summation of the variances of each critical activity. Variances of each activity can be calculated by taking the square of the standard deviations.

$$t_e = \frac{t_o + 4t_m + t_p}{6} \tag{4.2}$$

$$\sigma = \frac{t_p - t_o}{6} \tag{4.3}$$

$$Variance = \sigma^2 \tag{4.4}$$

After finding the total expected duration and the standard deviation values of the whole project, Equation 4.5 should be used for the estimation of the project duration for a desired level of probability, where Z is the number of standard deviations from the mean value and t_s is the project scheduled time. In PERT, the aim here is to find the t_s values of each activity for 5%, 95% and 50% probability levels.

$$Z = \frac{t_s - t_e}{\sigma} \tag{4.5}$$

Z value is obtained from the z-Tables by assuming that the duration of the whole project is normally distributed (Lu & AbouRizk, 2000). Z value should be chosen according to the required level of probability and then the Equation 4.5 should be used to calculate the project scheduled time for that probability level.

The values for the 50% probability level is equal to the mean value, where z is equal to zero. After making the calculations for each activity of each predicted project as stated above, the results of PERT for each project come out to be as given in Table 4.2 considering the relation between the activities as finish to start, formwork activity being the predecessor. So that, the concrete pouring activity cannot start before the formwork activity is completely finished.

Table 4.2. Differences between the actual durations and range estimations of PERT

	Project 1	**Project 2**	**Project 3**	**Project 4**
Actual Duration (days)	89	69	23	33
%5 Probability Value (days)	51	53	19	28
%95 Probability Value (days)	230	112	65	87
%50 Probability Value (days)	140	82	42	57
Actual - %50 Probability Difference (days)	51	13	19	24

After completing the predictions using PERT, Monte Carlo Simulation is used with triangular distribution in order to predict the same projects and compare its results.

4.1.1.2. Analysis with Monte Carlo Simulation

Monte Carlo simulation is a random simulating and iterative method. Its logic may be visualized superficially as illustrated in Figure 4.7, which shows that Monte Carlo simulates the project schedule many times to include all the possible outcomes in its estimation. In addition, it is similar to bootstrap method. The main difference between Monte Carlo Simulation and non-parametric bootstrap is that non-parametric bootstrap does not require any assumptions regarding the probability distributions of the input parameters. In addition, with the usage of non-parametric bootstrap method, the correlations are also taken into account without making any further calculations. Besides these two advantages of the non-parametric bootstrap method over Monte Carlo Simulation, sampling size is also increased by using the bootstrap method as another advantage, so that it is capable of providing more realistic results when compared to Monte Carlo Simulation method. On the other hand, before performing Monte Carlo Simulation, one should determine the probability distributions of the inputs. The pre-determined distributions of the activities in Figure 4.7 seem to be triangular, however it may change according to user. This difference is the main drawback of the Monte Carlo Simulation when compared to non-parametric bootstrap method, since the assumption of probability distribution and correlations may mislead the results. In addition, Monte Carlo also provides results with their corresponding probabilities after performing iterations with pre-determined distributions.

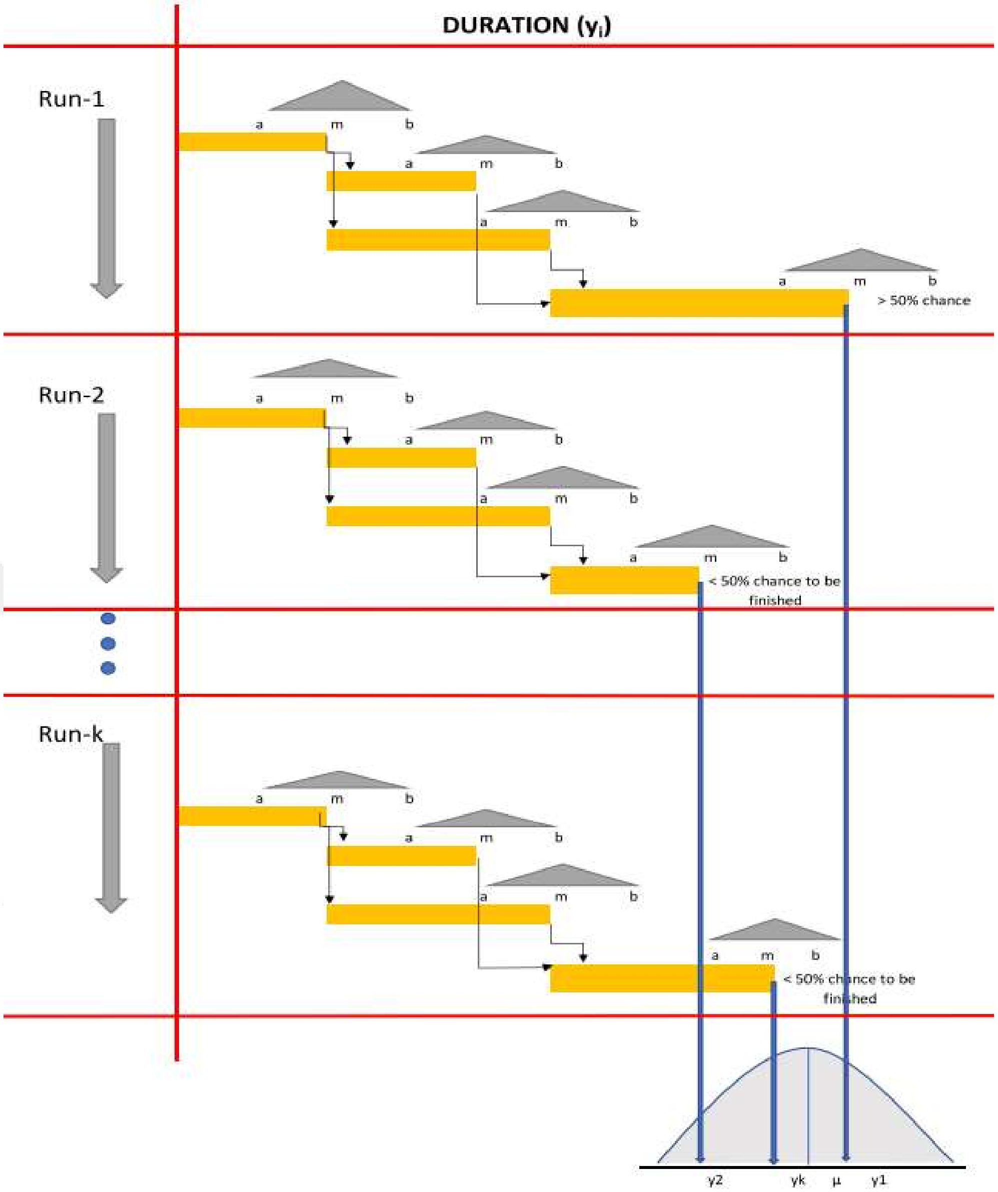
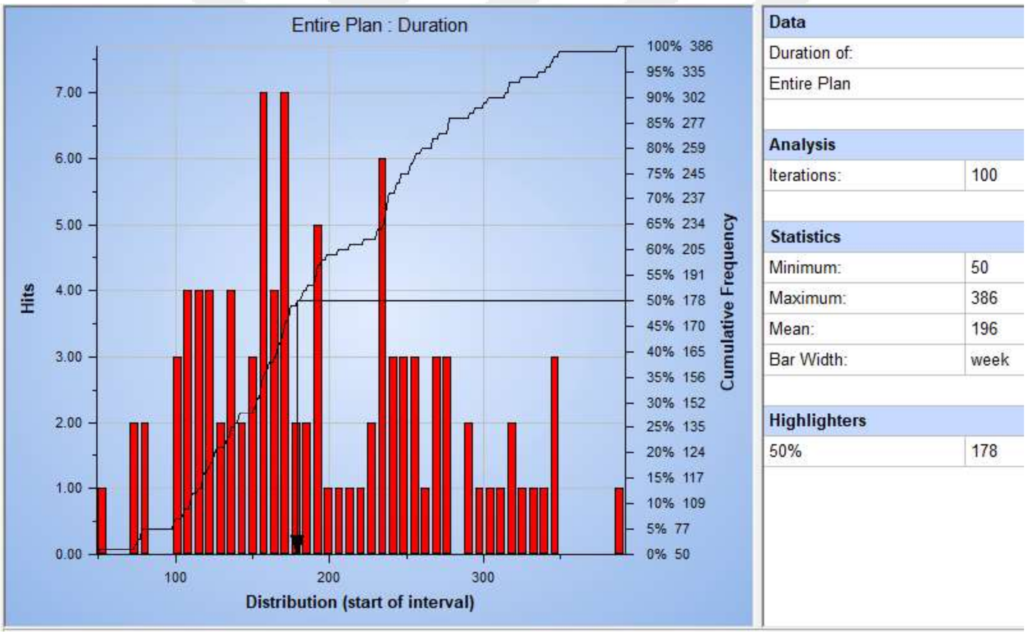


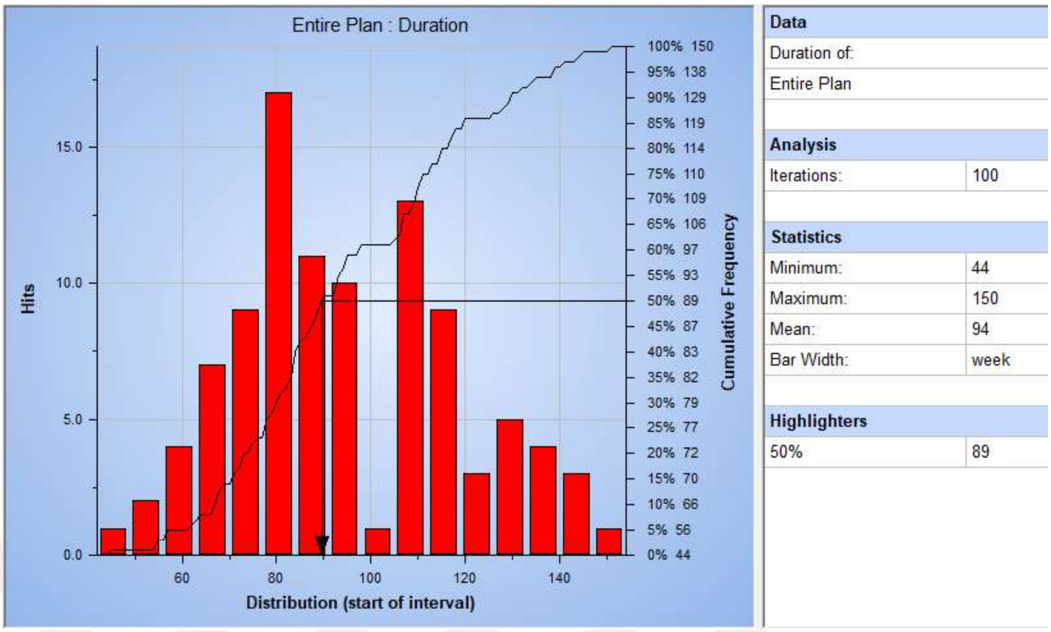
Figure 4.7. Logic of the Monte Carlo simulation adapted from Karabulut (2017)

The triangular distribution is used in Monte Carlo Simulation for the case study in order to predict the same four projects again, since triangular distribution is one of the most widely used distribution types in Monte Carlo Simulation (Barraza, 2011; Kong et al., 2015; D.-E. Lee, 2005; D.-E. Lee et al., 2013; D. Lee & Arditi, 2006; Nguyen et al., 2013; Dorp & Duffey, 1999). However, other distribution types may provide different results. In triangular distribution; pessimistic, most likely and optimistic durations are required for each activity and these durations are taken as the same values calculated for PERT (Dorp & Duffey, 1999). After that, Primavera Risk Analysis software is used to simulate each project schedule. 100 iterations are conducted for Monte Carlo Simulation, which is the same iteration number of the bootstrap scheduling (100 bootstrap resamples are generated while using the tool). Moreover, 7-day calendar is used for each activity of each project to eliminate the holidays occurring on the weekends, since no holidays are considered for the bootstrap scheduling and PERT. Also, the activities are again linked as finish to start and the formwork activity is the predecessor of the concrete pouring activity. After conducting the simulations of each four projects, the distribution graphs of the project 1, project 2, project 3 and project 4 are obtained as illustrated in Figure 4.8, Figure 4.9, Figure 4.10 and Figure 4.11, respectively.



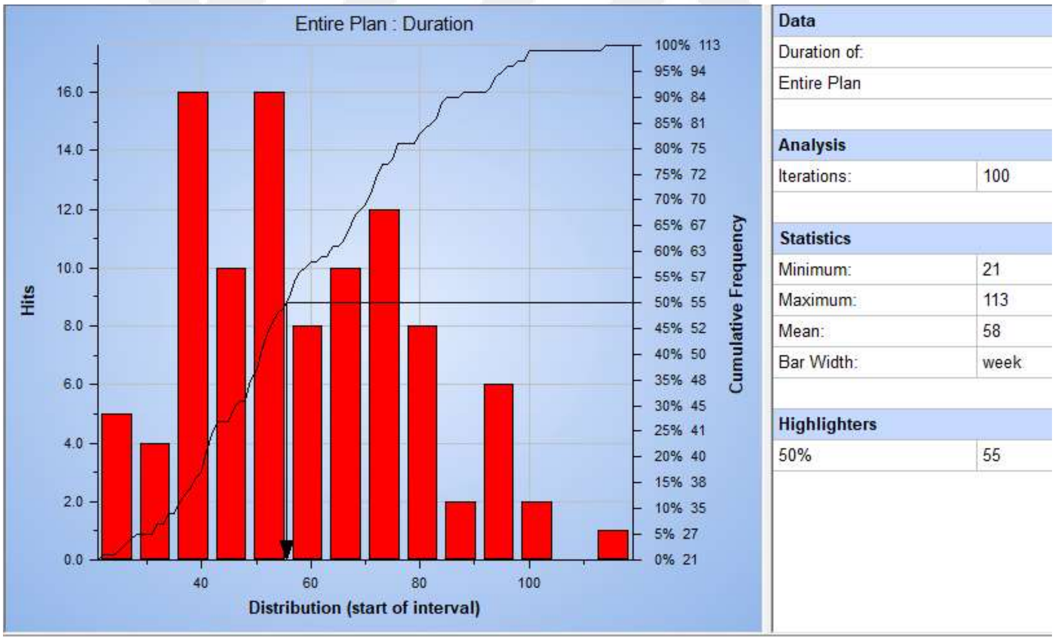
Data	
Duration of:	
Entire Plan	
Analysis	
Iterations:	100
Statistics	
Minimum:	50
Maximum:	386
Mean:	196
Bar Width:	week
Highlighters	
50%	178

Figure 4.8. The distribution graph of project 1



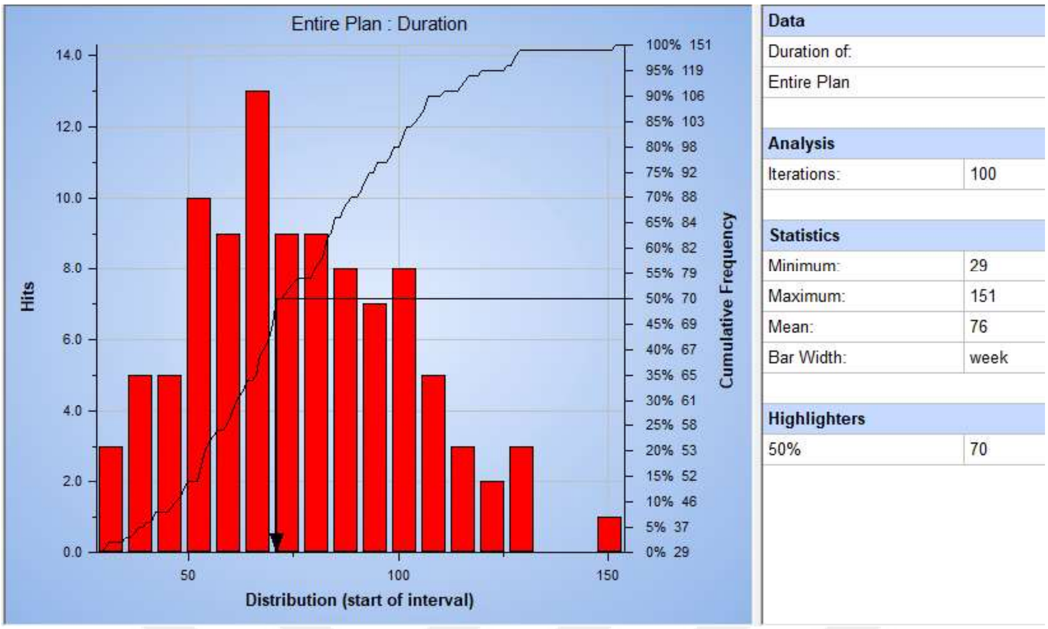
Data	
Duration of:	
Entire Plan	
Analysis	
Iterations:	100
Statistics	
Minimum:	44
Maximum:	150
Mean:	94
Bar Width:	week
Highlighters	
50%	89

Figure 4.9. The distribution graph of project 2



Data	
Duration of:	
Entire Plan	
Analysis	
Iterations:	100
Statistics	
Minimum:	21
Maximum:	113
Mean:	58
Bar Width:	week
Highlighters	
50%	55

Figure 4.10. The distribution graph of project 3



Data

Duration of:

Entire Plan

Analysis

Iterations:	100

Statistics

Minimum:	29
Maximum:	151
Mean:	76
Bar Width:	week

Highlighters

50%	70

Figure 4.11. The distribution graph of project 4

The results of Monte Carlo Simulation method can be also examined in Table 4.3 for all the projects.

Table 4.3. Differences between the actual durations and the range estimations of Monte Carlo Simulation

	Project 1	Project 2	Project 3	Project 4
Actual Duration (days)	89	69	23	33
%5 Probability Value (days)	77	56	27	37
%95 Probability Value (days)	335	138	94	119
%50 Probability Value (days)	178	89	55	70
Actual - %50 Probability Difference (days)	89	20	32	37

4.1.1.3. Comparison of the Results

The results of each method are provided together with the actual durations of each project for showing the performances of each method individually. The general comparison table, which involves the results of the bootstrap scheduling tool, PERT and Monte Carlo Simulation along with the actual durations, is provided in Table 4.4 below.

Table 4.4. General comparison table of all the methods and actual durations

Predicted Project	Actual Durations	Prediction Method	%5 Probability Value (days)	%95 Probability Value (days)	%50 Probability Value (days)	Actual - %50 Probability Difference (days)
1	89	Bootstrap Scheduling Method	69	153	90	1
		Monte Carlo Simulation	77	335	178	89
		PERT	51	230	140	51
2	69	Bootstrap Scheduling Method	51	64	58	11
		Monte Carlo Simulation	56	138	89	20
		PERT	53	112	82	13
3	23	Bootstrap Scheduling Method	21	28	24	1
		Monte Carlo Simulation	27	94	55	32
		PERT	19	65	42	19
4	33	Bootstrap Scheduling Method	26	35	31	2
		Monte Carlo Simulation	37	119	70	37
		PERT	28	87	57	24

As it can be clearly seen in Table 4.4, the proposed integrated regression-bootstrap sampling scheduling method outperforms the other two classical estimation methods, which are PERT and Monte Carlo Simulation. The 90% probability range of the bootstrap scheduling tool is narrower and hence more realistic when compared to other ones. Narrower range also shows that the variation is smaller in the estimation of the tool when compared to PERT and Monte Carlo Simulation. Moreover, the difference between the durations corresponding to the 50% probability levels of the tool and the actual durations of the predicted projects are smaller than the difference between the durations corresponding to the 50% probability levels of the existing tools and the actual durations of the predicted projects.

4.2. Second Case Study of Construction Project Scheduling using the Integrated Regression-Bootstrap Sampling Scheduling Method

All the procedure is the same for the tool, PERT and Monte Carlo Simulation as the first case study, which is explained in detail before. Therefore, the steps will not be explained for the methods, however only the results will be provided. In addition, the data will be mentioned initially, before continuing with the results of the bootstrap scheduling tool, since the data is different than the data of the first case study.

The data used in this case study is collected from a Turkish contractor and the information belong to a real completed building project in Turkey, which is built between the years 2012 and 2014. The data consist of the daily production records and six activities are available in the historical data. Their productivity values are calculated for each day, since the progresses are recorded on a daily basis. Those six activities are excavation, rebar, formwork, concrete pouring, walls and plastering and the number of their daily data points are 70, 137, 108, 51, 9 and 4, respectively. All the activities are linked as finish-to-start and the predecessors of each activity are given in Table 4.5 and none of the activities can start before its predecessor is completely finished. Moreover, the affecting factors of the productivities are determined according to the available information. Then they are dropped from the model until obtaining a parsimonious model by conducting regression analyses, as explained in previous chapters. The available and final determined affecting factors are given in Table 4.6, Table 4.7, Table 4.8, Table 4.9, Table 4.10 and Table 4.11 below for each activity, separately.

Table 4.5. Activities and their predecessors

Activities	Predecessors
Excavation	-
Rebar	Excavation
Formwork	Rebar
Concrete pouring	Formwork
Walls	Concrete pouring
Plastering	Walls

Table 4.6. Factors of excavation

Available Factors	Final Factors
Quantity	Quantity
Crew size	Crew size
Foreman ratio	Foreman ratio
Temperature	Working hours per day
Humidity	
Precipitation	
Working hours per day	
Number of operators (equipment)	

Table 4.7. Factors of rebar

Available Factors	Final Factors
Quantity	Quantity
Crew size	Crew size
Foreman ratio	Foreman ratio
Temperature	Qualified workman ratio
Humidity	
Precipitation	
Qualified workman ratio	

Table 4.8. Factors of formwork

Available Factors	**Final Factors**
Quantity	Quantity
Crew size	Crew size
Foreman ratio	Qualified workman ratio
Temperature	
Humidity	
Precipitation	
Qualified workman ratio	

Table 4.9. Factors of concrete pouring

Available Factors	**Final Factors**
Quantity	Column & Curtain wall ratio
Crew size	
Foreman ratio	
Temperature	
Humidity	
Precipitation	
Qualified workman ratio	
Column & Curtain wall ratio	
Slab ratio	
Foundation ratio	

Table 4.10. Factors of walls

Available Factors	**Final Factors**
Quantity	Quantity
Crew size	Crew size
Temperature	
Humidity	

Table 4.11. Factors of plastering

Available Factors	Final Factors
Quantity	Crew size
Crew size	
Temperature	
Humidity	
Qualified workman ratio	

The initial factors included for the excavation model are quantity, crew size, foreman ratio, weather temperature, humidity, precipitation, working hours per day and number of operators (equipment). After conducting the first regression analysis it comes out to be that number of operators does not affect the model significantly due its large P-value. So that, it is the first factor to be dropped from the model and the regression analysis is conducted again to check the R^2 value if the accuracy is significantly reduced or not. After checking the accuracy of the model and the P-values of the remaining factors, it is concluded that the absence of the number of the operators does not significantly affect the model. So that, it is dropped from the model. After dropping number of operators, precipitation, humidity and weather temperature are also dropped from the model, respectively using the same manner, one by one as explained previously. Finally, the model included quantity, crew size, foreman ratio and working hours per day as the factors for productivity.

The initial factors included for the rebar model are quantity, crew size, foreman ratio, weather temperature, humidity, precipitation and qualified workman ratio. After conducting the first regression analysis, it comes out to be that humidity does not affect the model significantly due its large P-value. So that, it is the first factor to be dropped from the model and the regression analysis is conducted again to check the R^2 value if the accuracy is significantly reduced or not. After checking the accuracy of the model and the P-values of the remaining factors, it is concluded that the absence of the humidity does not significantly affect the model. So that, it is dropped from the model. After dropping humidity, weather temperature and precipitation are also dropped from the model, respectively using the same manner, one by one as explained previously. Finally, the model included quantity, crew size, foreman ratio and qualified workman ratio as the factors for productivity.

The initial factors included for the formwork model are quantity, crew size, foreman ratio, weather temperature, humidity, precipitation and qualified workman ratio. After conducting the first regression analysis, it comes out to be that precipitation does not affect the model significantly due its large P-value. So that, it is the first factor to be dropped from the model and the regression analysis is conducted again to check the R^2 value if the accuracy is significantly reduced or not. After checking the accuracy of the model and the P-values of the remaining factors, it is concluded that the absence of the precipitation does not significantly affect the model. So that, it is dropped from the model. After dropping precipitation, foreman ratio, weather temperature and humidity are also dropped from the model, respectively using the same manner, one by one as explained previously. Finally, the model included quantity, crew size and qualified workman ratio as the factors for productivity.

The initial factors included for the concrete pouring model are quantity, crew size, foreman ratio, weather temperature, humidity, precipitation, qualified workman ratio, column & curtain wall ratio, slab ratio and foundation ratio. After conducting the first regression analysis, it comes out to be that crew size and quantity do not affect the model at all, since crew size is the same in all the daily data points and the linear correlation between the productivity and quantity is one due to the same crew size in all the data points. So that, quantity and crew size are the first factors to be dropped from the model and the regression analysis is conducted again to check the R^2 and P values. After checking the accuracy of the model and the P-values of the remaining factors, it comes out to be that qualified workman ratio does not affect the model significantly due its large P-value. So that, it is the third factor to be dropped from the model and the regression analysis is conducted again to check the R^2 value if the accuracy is significantly reduced or not. After checking the accuracy of the model and the P-values of the remaining factors, it is concluded that the absence of the qualified workman ratio does not significantly affect the model. So that, it is dropped from the model. After dropping qualified workman ratio, foundation ratio, weather temperature, slab ratio, foreman ratio, humidity and precipitation are also dropped from the model, respectively using the same manner, one by one as explained previously. Finally, the model included only column & curtain wall ratio as the factor for productivity.

The initial factors included for the walls model are quantity, crew size, weather temperature and humidity. After conducting the first regression analysis, it comes out to be that the linear correlation between the weather temperature and humidity is one. Therefore, humidity is dropped from the model initially and the regression analysis is conducted again to check the R^2 and P values. After checking the accuracy of the model and the P-values of the remaining factors, it comes out to be that weather temperature does not affect the model significantly due its large P-value. So that, it is the second factor to be dropped from the model and the regression analysis is conducted again to check the R^2 value if the accuracy is significantly reduced or not. After checking the accuracy of the model and the P-values of the remaining factors, it is concluded that the absence of the weather temperature does not significantly affect the model. So that, it is dropped from the model. After dropping weather temperature, the R^2 and P values are checked again and it is concluded that the parsimonious model is obtained. So that, the final model included quantity and crew size as the factors for the productivity.

The plastering activity has four daily data points. Therefore, only two factors can be included in the regression model. The records of quantity, crew size, weather temperature, humidity and qualified workman ratio are available in the database. However, weather temperature and humidity are directly eliminated, since they are also not contributed to the final models of the other activities. The regression analysis is conducted for the remaining factors, which are quantity, crew size and qualified workman ratio, including all the combinations. However, it is concluded that only the inclusion of the crew size provides a good fit as the final model.

Among all the factors given in the tables above, only precipitation is a binary variable. It equals to one if there is precipitation and zero otherwise. In addition, working hours per day factor is only included in the excavation activity, because this factor differs only in excavation activity. Besides that, considering the final models after elimination, quantity, foreman ratio and qualified workman ratio increase the productivity, whereas crew size, working hours per day and column & curtain wall ratio decrease the productivity, which are expected. However, in rebar activity, foreman ratio and qualified workman ratio also decrease the productivity, which is not logical. The reason behind this issue may be the inefficient performances of the foremen and qualified workmen.

The final data set then used to schedule the construction project. It is resampled 100 times using the non-parametric bootstrap method and then the productivity and duration of each activity are found using linear regression analysis for each bootstrap sample as it is already demonstrated in the previous case study. The method is exactly the same. However, all the daily data of each activity are used for analysis in this case study and the prediction is made for the five times and ten times of the average daily quantities of each activity, separately. The quantities of the activities are as shown in Table 4.12 and only the five times and ten times of the average daily quantities are used for the comparison of the proposed method and the traditional methods. Since the original data contain daily quantities and the project consists of six activities, it is expected that the total project duration should be approximately 30 days for the prediction using the five times of average quantities of each activity and 60 days for the prediction using the ten times of average quantities of each activity. The distribution graphs found for five times of the average quantities and ten times of the average quantities using the bootstrap scheduling method are presented in Figure 4.12 and Figure 4.13, respectively.

Table 4.12. Average daily quantities and the quantities used in the case study

Activities and Units	Average Daily Quantities	Five Times of Average Daily Quantities	Ten Times of Average Daily Quantities
Excavation (m^3)	659	3295	6590
Rebar (ton)	29	145	290
Formwork (m^2)	411	2055	4110
Concrete pouring (m^3)	188	940	1880
Walls (m^2)	44	220	440
Plastering (m^2)	89	445	890

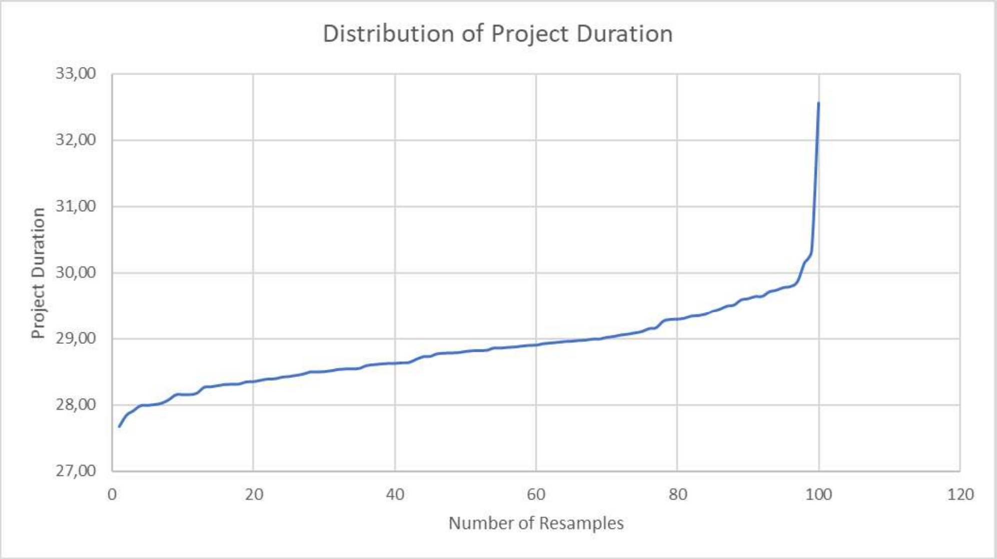


Figure 4.12. Distribution graph obtained using five times of the average daily quantities of each activity

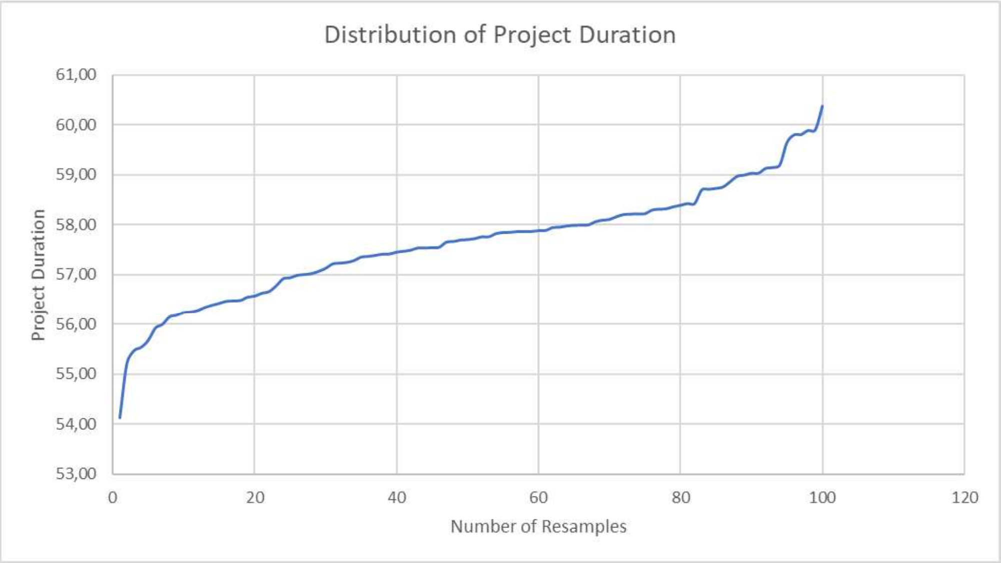


Figure 4.13. Distribution graph obtained using ten times of the average daily quantities of each activity

In addition, the range of 5% - 95% probabilities are provided in the Table 4.13 below together with the 50% probability value and actual durations, which is 30 days and 60 days.

Table 4.13. Differences between the actual durations and range estimations of the tool

	Five Times of Average Daily Quantities	**Ten Times of Average Daily Quantities**
Actual Duration (days)	30	60
%5 Probability Value (days)	28	56
%95 Probability Value (days)	30	60
%50 Probability Value (days)	29	58
Actual - %50 Probability Difference (days)	1	2

As it can be seen from the results, the developed tool using the proposed bootstrap scheduling method, performs accurate enough. It predicts the durations close to the actual durations, which is important for probabilistic risk analysis.

The results of the existing methods, such as PERT and Monte Carlo Simulation will also be presented and they will also be compared with the results of the developed tool.

4.2.1. Analysis with PERT

The smallest productivity value is used for the pessimistic estimation, the greatest productivity value is used for the optimistic estimation and the average value of the productivity is used for the most likely estimation, which is the same method as the previous case study. So that, pessimistic, optimistic and most likely estimations are obtained for each activity. After taking the productivity values as stated, the average values of total quantity, average crew size and average working hours per day of each activity are also used in order to convert the productivity values to the durations.

After computing the three-point estimates and taking the required parameters for converting the productivity to duration, the expected duration and the standard deviation of the project are calculated and then the duration values corresponding to 5%, 50% and 95% probability levels are computed using the Z-table as explained in detail previously. The activities are again linked as finish-to-start and none of the activities can start before its predecessor is completely finished. The results of PERT are given in Table 4.14 below for five times and ten times of the average daily quantities of each activity.

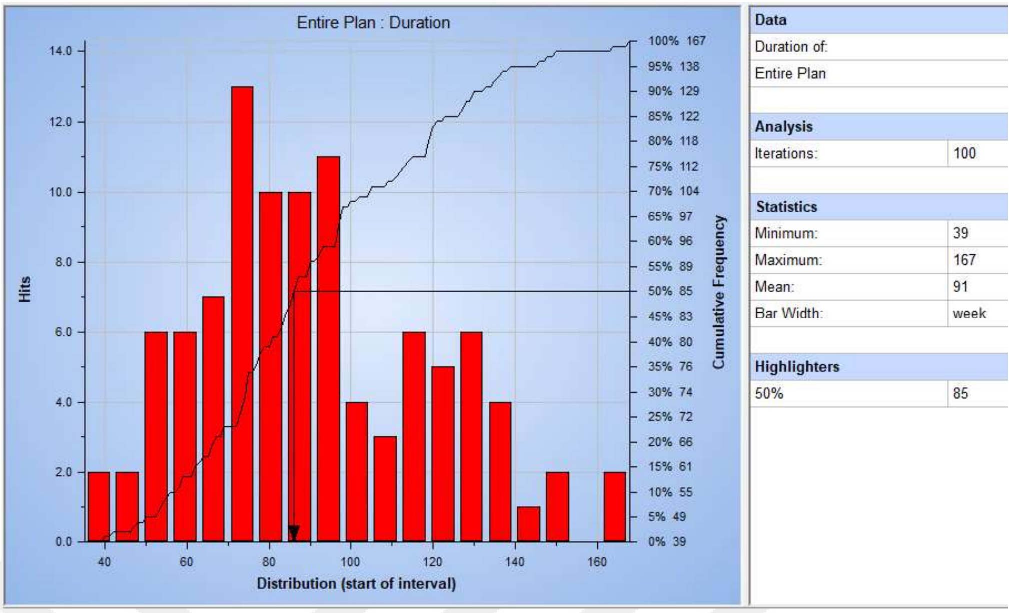
Table 4.14. Differences between the actual durations and range estimations of PERT

	Five Times of Average Daily Quantities	**Ten Times of Average Daily Quantities**
Actual Duration (days)	30	60
%5 Probability Value (days)	28	56
%95 Probability Value (days)	92	183
%50 Probability Value (days)	60	120
Actual - %50 Probability Difference (days)	30	60

In the next section, the results of the Monte Carlo Simulation method will be presented.

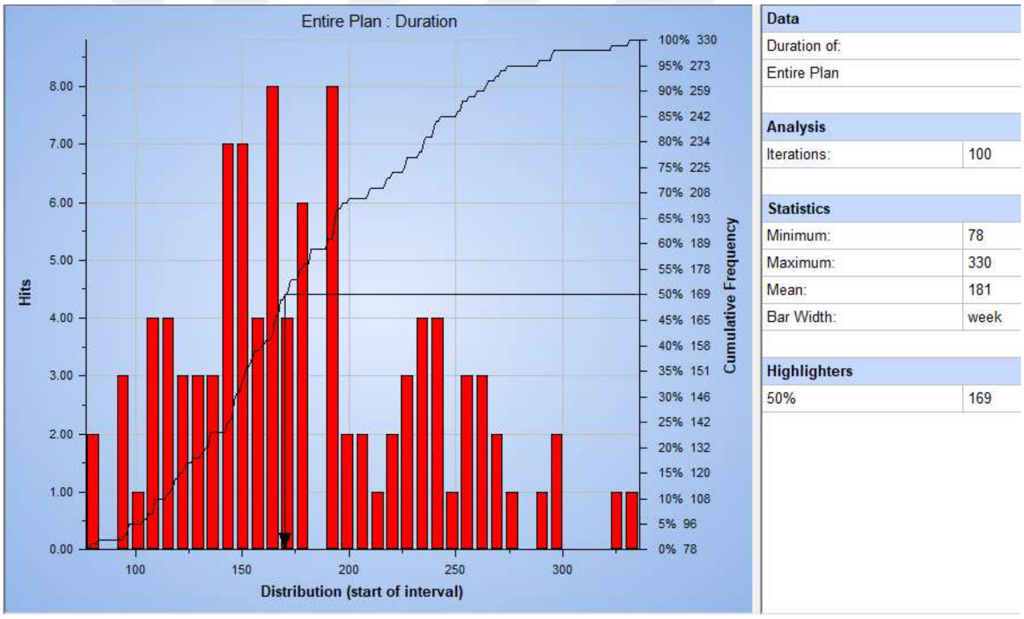
4.2.2. Analysis with Monte Carlo Simulation

Triangular distribution is used with 7-day calendar for each activity while conducting Monte Carlo Simulation method as it was also the case in the previous case study. Again, all the steps are the same as the previous one. Therefore, only the results will be provided. The distribution graphs for the five times and ten times of the average quantities are presented in Figure 4.14 and Figure 4.15, respectively. The results of Monte Carlo Simulation method are also presented in Table 4.15.



Data	
Duration of:	
Entire Plan	
Analysis	
Iterations:	100
Statistics	
Minimum:	39
Maximum:	167
Mean:	91
Bar Width:	week
Highlighters	
50%	85

Figure 4.14. Distribution graph for five times of the average quantities



Data	
Duration of:	
Entire Plan	
Analysis	
Iterations:	100
Statistics	
Minimum:	78
Maximum:	330
Mean:	181
Bar Width:	week
Highlighters	
50%	169

Figure 4.15. Distribution graph for ten times of the average quantities

Table 4.15. Differences between the actual durations and range estimations of Monte Carlo Simulation

	Five Times of Average Daily Quantities	Ten Times of Average Daily Quantities
Actual Duration (days)	30	60
%5 Probability Value (days)	49	96
%95 Probability Value (days)	138	273
%50 Probability Value (days)	85	169
Actual - %50 Probability Difference (days)	55	109

4.2.3. Comparison of the Results

The results of each method are provided together with the actual durations for showing the performances of each method individually. The general comparison table, which involves the results of the bootstrap scheduling method, PERT and Monte Carlo Simulation along with the actual durations, is provided in Table 4.16 below for a better comparison of the methods.

Table 4.16. General comparison table of all the methods and actual durations

Predicted Project	Actual Durations	Prediction Method	%5 Probability Value (days)	%95 Probability Value (days)	%50 Probability Value (days)	Actual - %50 Probability Difference (days)
Five times of the average daily quantities	30	Bootstrap Scheduling Method	28	30	29	1
		Monte Carlo Simulation	49	138	85	55
		PERT	28	92	60	30
Ten times of the average daily quantities	60	Bootstrap Scheduling Method	56	60	58	2
		Monte Carlo Simulation	96	273	169	109
		PERT	56	183	120	60

As it can be clearly seen from the Table 4.16, the proposed integrated regression-bootstrap sampling scheduling method provides more accurate results than the existing methods such as PERT and Monte Carlo Simulation. The ranges of Monte Carlo Simulation even do not contain the actual durations. Moreover, the estimated durations of PERT and Monte Carlo simulation corresponding to 50% probability are not even close to the actual durations. The results show that the usage of PERT or Monte Carlo Simulation for the duration prediction purposes may produce misleading results, at least for the practiced problems.

As the outcomes of the two case studies, the integrated regression-bootstrap sampling scheduling method outperformed the existing methods. The reasons may be explained as follows. Non-parametric bootstrap method eliminates the necessity of determining the distribution types of the activities such as normal, triangular distributions etc., which also eliminates the assumptions based on the distribution types. On the other hand, it is not an easy task to add up the activity durations of different distribution types. Although it can be achieved by using the simulation methods, the correlations between the activities get higher and these correlations should also be taken into account while estimating the project durations. However, non-parametric bootstrap method also resolves the correlation problem and therefore provides more realistic results as already proven. Beside these advantages of non-parametric bootstrap method over the existing methods, it can also be used to analyze the small samples (data sets), since it resamples the data sets as many times as requested by the user. Moreover, the tool also conducts regression analysis in addition to bootstrap resampling. Using the regression analysis, the relation between the productivity and the affecting factors are revealed and using the productivity, the variations in the activities are also taken into account for performing a more realistic approach.

The quantitative analyses of two different case studies are provided in this chapter with comparisons. The superiority of the proposed bootstrap scheduling method to the other methods is shown by using the actual projects. In the next chapter, all the information and remarks provided up to this point will be summarized and the thesis will be concluded.

CHAPTER 5

CONCLUSION

Since all the construction projects have a certain budget and deadline, project duration is a very crucial parameter for construction projects for schedule risk assessment, especially in tendering stages to foresee the duration ranges with the corresponding probabilities. Poor estimations of durations can yield lost opportunities or low profits if the tender is awarded, since the duration of a construction project directly affects the total cost. Moreover, it is also important to predict the duration range with its probability in order to evaluate the risk of time overrun while tendering. Because of these reasons, the estimations of project durations should be completed at the tendering stage, even if the project scope is not clear at all. Therefore, there is a need for a method to enable adequate duration range estimates at that early time of the construction projects. Based on this method, the contractor should be able to evaluate its decision for the tender, since the duration range of the project would be known with its probability and so that, the delay risk will also be known beforehand.

The existing methods such as CPM, PERT and Monte Carlo Simulation are not sufficient for fulfilling these needs of the contractors. CPM is a deterministic method, which provides only a single point estimation. On the other hand, PERT is a probabilistic method but it adopts three-point estimation (PERT-beta distribution) which may not be valid for all of the activities. Therefore, PERT may mislead the decision makers while evaluating the delay risk of the projects. Although Monte Carlo Simulation is a probabilistic and more enhanced method than PERT, it is also not a sufficient method again due to the assumptions of the distribution types of the activity durations. Even if the distributions of the activities are found using the statistical methods, adding up the different distribution types of the activities comes with high correlations between the activities. In typical Monte Carlo analysis, correlations are not considered, which causes unrealistic results. However, non-parametric bootstrap method eliminates the assumptions of the distribution types and also the bias occurred due to the distributions of the activities, since it does not require any assumptions or determination of the distribution types of the activities. Therefore, the main goal of this study is to overcome this drawback by developing an integrated method, which integrates the non-parametric bootstrap method with the regression analysis, in order to predict the duration ranges of the construction projects by using productivity values for probabilistic risk analysis. Productivity values are used to include the variations within the activities in the prediction model and regression analysis is used to consider the affecting factors of the productivity values, while the non-parametric bootstrap method eliminates the assumptions of the distributions in order to obtain a robust model for duration range estimation of the construction projects with the corresponding probabilities.

The bootstrap scheduling tool developed in this thesis, which uses the proposed integrated regression-bootstrap sampling scheduling method, is also capable of analyzing the small data sets, when the data are scarce. This is another advantage of using the bootstrap method, since it can resample the original data set as many times as the user or decision maker desires. This resampling advantage allows the users to obtain realistic results even for small data sets, when the available data is limited.

As the first case study, four real projects, each consisting of two activities were used to compare the results of the bootstrap scheduling method, PERT and Monte Carlo Simulation method along with the actual durations. 90% probability ranges (the range between the durations corresponding to 5% and 95% probability levels) and the durations corresponding to 50% probability levels of the mentioned three methods and the actual durations of the four projects were considered for the comparison. 100 resamples were generated with the bootstrap scheduling tool and 100 iterations were made with Monte Carlo Simulation, whereas PERT has no capability of iterating or resampling. After the computations, the results were provided in the same table with the actual durations for the sake of comparing them easily. The results have clearly shown that the proposed method outperformed PERT and Monte Carlo Simulation, since the estimated 90% range of the proposed method was narrower and more realistic than the estimated 90% ranges of PERT and Monte Carlo Simulation. It also means that the variation was smaller in comparison with the other two methods. Moreover, the durations corresponding to 50% probability levels obtained by the bootstrap scheduling tool were very close to the actual durations and additionally, they were also relatively closer to the actual durations when compared to the estimated durations corresponding to 50% probability levels obtained by PERT and Monte Carlo Simulation.

As the second case study, another real project consisting of six activities was used to compare the estimation accuracies of the proposed bootstrap scheduling method, PERT and Monte Carlo Simulation along with the actual durations. For the prediction of the duration, five times and ten times of the average daily quantities of each activity were used separately as the quantities of the activities of the predicted project. The same procedures as the first case study were applied and again the 90% probability ranges and 50% probability levels were compared. The results were provided in the same table with the actual durations for the sake of better comparison. As the results clearly show that the proposed integrated method outperformed the existing methods, since it predicted the durations more accurately. The 50% probability levels and the ranges estimated using PERT and Monte Carlo Simulation were not even close to the actual durations.

Even though the proposed method has major advantages compared to PERT and Monte Carlo Simulation, it has also some limitations, which may be eliminated in the further studies focusing on the probabilistic range estimation of the durations. First of all, all the activities can be only linked as finish to start relation and the lags between the activities cannot be taken into account, which may not be realistic for the construction projects. In addition to the limitation of the relations between the activities, the total float values of the activities may also be provided as the output in order to see the buffer times of each activity of a project. It may be helpful to see the total float values, while the decision makers evaluate the delay risks to avoid the delay penalties. Moreover, the tool is capable of conducting only the linear regression analysis. However, the prediction performance of a regression model increases if the relation between the dependent and independent variables are presented good enough. It means that the other relations such as quadratic relation between the dependent and independent variables could also be tested instead of using only the linear relations between the parameters for further developments. Additionally, it takes time for the tool to provide results and this time increases with the increasing numbers of data points, activities, bootstrap samples and predecessors.

REFERENCES

AbouRizk, S., Knowles, P., & Hermann, U. R. (2001). Estimating Labor Production Rates for Industrial Construction Activities. *Journal of Construction Engineering and Management*, *127*(6), 502–511.

Al-Zwainy, F. M. S., Abdulmajeed, M. H., & Aljumaily, H. S. M. (2013). Using Multivariable Linear Regression Technique for Modeling Productivity Construction in Iraq. *Open Journal of Civil Engineering*, *3*, 127–135. https://doi.org/10.4236/ojce.2013.33015

Arunmohan, A. M., & Lakshmi, M. (2018). Analysis of Modern Construction Projects Using Montecarlo Simulation Technique. *International Journal of Engineering & Technology*, *7*(2.19), 41–44.

Barraza, G. A. (2011). Probabilistic Estimation and Allocation of Project Time Contingency. *Journal of Construction Engineering and Management*, *137*(4), 259–265. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000280

Dallah, H. (2012). A Bootstrap Approach to Robust Regression. *International Journal of Applied Science and Technology*, *2*(9), 114–118.

Dorp, J. R. van, & Duffey, M. R. (1999). Statistical Dependence in Risk Analysis for Project Networks Using Monte Carlo Methods. *International Journal of Production Economics*, *58*, 17–29.

Elaiwi, A. H. (2018). Efficiency of Critical Path Method (CPM) and PERT Technique for Yacht Construction. *International Journal of Mechanical Engineering and Technology (IJMET)*, *9*(11), 48–54.

Ergin, A., Balas, C. E., & Keyder, E. (1995). A Network Planning Model for Offshore Structures. In *Proceedings of the Fifth (1995) International Offshore and Polar Engineering Conference* (Vol. 1, pp. 227–230). Hague: The International Society of Offshore and Polar Engineers.

Gardner, B. J., Gransberg, D. D., & Rueda, J. A. (2017). Stochastic Conceptual Cost Estimating of Highway Projects to Communicate Uncertainty Using Bootstrap Sampling. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, *3*(3). https://doi.org/10.1061/AJRUA6 .0000895

Hajdu, Miklós. (2013). Effects of the Application of Activity Calendars on the Distribution of Project Duration in PERT Networks. *Automation in Construction*, *35*, 397–404. https://doi.org/10.1016/j.autcon.2013.05.025

Hajdu, Miklos, & Bokor, O. (2014). The Effects of Different Activity Distributions on Project Duration in PERT Networks. In *Procedia - Social and Behavioral Sciences* (Vol. 119, pp. 766–775). Elsevier Ltd. https://doi.org/10.1016/j.sbspro.2014.03.086

Hajdu, Miklós, & Bokor, O. (2016). Sensitivity analysis in PERT networks: Does activity duration distribution matter? *Automation in Construction*, *65*, 1–8. https://doi.org/10.1016/j.autcon.2016.01.003

Hashemi, H., Mousavi, S. M., & Mojtahedi, S. M. H. (2011). Bootstrap Technique for Risk Analysis with Interval Numbers in Bridge Construction Projects. *Journal of Construction Engineering and Management*, *137*(8), 600–608. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000344

Hashemi, H., Mousavi, S. M., Tavakkoli-Moghaddam, R., & Gholipour, Y. (2013). Compromise Ranking Approach with Bootstrap Confidence Intervals for Risk Assessment in Port Management Projects. *Journal of Management in Engineering*, *29*(4), 334–344. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000167

Hegazy, T. (1999). Optimization of Construction Time-Cost Trade-off Analysis Using Genetic Algorithms. *Canadian Journal of Civil Engineering*, *26*, 685–697.

Isidore, L. J., & Back, W. E. (2002). Multiple Simulation Analysis for Probabilistic Cost and Schedule Integration. *Journal of Construction Engineering and Management*, *128*(3), 211–219. https://doi.org/10.1061/(ASCE)0733-9364(2002)128:3(211)

Karabulut, M. (2017). Application of Monte Carlo Simulation and PERT/CPM Techniques in Planning of Construction Projects: A Case Study. *Periodicals of Engineering and Natural Sciences*, *5*(3), 408–420. https://doi.org/10.21533/pen.v5i3.152

Kazaz, A., & Ulubeyli, S. (2007). Drivers of Productivity among Construction Workers: A Study in a Developing Country. *Building and Environment*, *42*, 2132–2140. https://doi.org/10.1016/j.buildenv.2006.04.020

Kelley, J. E., & Walker, M. R. (1959). Critical-Path Planning and Scheduling. In *PROCEEDINGS OF THE EASTERN JOINT COMPUTER CONFERENCE* (pp. 160–173). Boston, Massachusetts: IRE-AIEE-ACM.

Kong, Z., Zhang, J., Li, C., Zheng, X., & Guan, Q. (2015). Risk Assessment of Plan Schedule by Monte Carlo Simulation. In *International Conference on Information Technology and Management Innovation (ICITMI 2015)* (pp. 509–513). Atlantis Press.

Lee, D.-E. (2005). Probability of Project Completion Using Stochastic Project Scheduling Simulation. *Journal of Construction Engineering and Management*, *131*(3), 310–318. https://doi.org/10.1061/(ASCE)0733-9364(2005)131:3(310)

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. https://doi.org/10.1007/s12205-013-0147-x

Lee, D., & Arditi, D. (2006). Automated Statistical Analysis in Stochastic Project Scheduling Simulation. *Journal of Construction Engineering and Management*, *132*(3), 268–278.

Lee, H. C., Lee, E. B., & Alleman, D. (2018). Schedule Modeling to Estimate Typical Construction Durations and Areas of Risk for 1000 MW Ultra-Critical Coal-Fired Power Plants. *Energies*, *11*(10). https://doi.org/10.3390/en11102850

Li, X., Chow, K. H., Zhu, Y., & Lin, Y. (2016). Evaluating the Impacts of High-Temperature Outdoor Working Environments on Construction Labor Productivity in China: A Case Study of Rebar Workers. *Building and Environment*, *95*, 42–52. https://doi.org/10.1016/j.buildenv.2015.09.005

Liu, M. (2013). Program Evaluation and Review Technique (PERT) in Construction Risk Analysis. *Applied Mechanics and Materials*, *357–360*, 2334–2337. https://doi.org/10.4028/www.scientific.net/AMM.357-360.2334

Lu, M., & AbouRizk, S. M. (2000). Simplified CPM/PERT Simulation Model. *Journal of Construction Engineering and Management*, *126*(3), 219–226.

Nasirzadeh, F., & Nojedehi, P. (2013). Dynamic Modeling of Labor Productivity in Construction Projects. *International Journal of Project Management*, *31*, 903–911. https://doi.org/10.1016/j.ijproman.2012.11.003

Nguyen, L. D., Phan, D. H., & Tang, L. C. M. (2013). Simulating Construction Duration for Multistory Buildings with Controlling Activities. *Journal of Construction Engineering and Management*, *139*(8), 951–959. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000677

Öztaş, A., & Ökmen, Ö. (2005). Judgmental risk analysis process development in construction projects. *Building and Environment*, *40*, 1244–1254. https://doi.org/10.1016/j.buildenv.2004.10.013

Park, H.-S. (2006). Conceptual Framework of Construction Productivity Estimation. *KSCE Journal of Civil Engineering*, *10*(5), 311–317. Retrieved from http://link.springer.com/10.1007/BF02830084

Plebankiewicz, E., Juszczyk, M., & Malara, J. (2015). Estimation of Task Completion Times with The Use of The PERT Method on The Example of a Real Construction Project. *Archives of Civil Engineering*, *61*(3), 51–62. https://doi.org/10.1515/ace-2015-0024

Salas-Morera, L., Arauzo-Azofra, A., García-Hernández, L., Palomo-Romero, J. M., & Ayuso-Muñoz, J. L. (2018). New Approach to the Distribution of Project Completion Time in PERT Networks. *Journal of Construction Engineering and Managemet*, *144*(10). https://doi.org/10.1061/(ASCE)CO.1943-7862.0001552

Silvianita, Aprillia, N., Mulyadi, Y., Citrosiswoyo, W., & Suntoyo. (2018). Cost and Time Analysis of Graving Dock Project. In *MATEC Web of Conferences* (Vol. 177). EDP Sciences. https://doi.org/https://doi.org/10.1051/matecconf/201817701028

Sonmez, R. (2008). Parametric Range Estimating of Building Costs Using Regression Models and Bootstrap. *Journal of Construction Engineering and Management*, *134*(12), 1011–1016. https://doi.org/10.1061/(ASCE)0733-9364(2008)134:12(1011)

Sonmez, R. (2011). Range Estimation of Construction Costs Using Neural Networks with Bootstrap Prediction Intervals. *Expert Systems with Applications*, *38*, 9913–9917. https://doi.org/10.1016/j.eswa.2011.02.042

Sonmez, R., & Rowings, J. E. (1998). Construction Labor Productivity Modeling with Neural Networks. *Journal of Construction Engineering and Management*, *124*(6), 498–504.

Tsai, T.-I., & Li, D.-C. (2008). Utilize Bootstrap in Small Data Set Learning for Pilot Run Modeling of Manufacturing Systems. *Expert Systems with Applications*, *35*, 1293–1300. https://doi.org/10.1016/j.eswa.2007.08.043

Usukhbayar, R., & Choi, J. (2018). Determining the Impact of Key Climatic Factors on Labor Productivity in the Mongolian Construction Industry. *Journal of Asian Architecture and Building Engineering*, *17*(1), 55–62. https://doi.org/10.3130/jaabe.17.55

Valášková, K., Spuchľáková, E., & Adamko, P. (2015). Non-parametric Bootstrap Method in Risk Management. In *Procedia Economics and Finance* (Vol. 24, pp. 701–709). Kazan, Russia: Elsevier B.V. https://doi.org/10.1016/S2212-5671(15)00678-4