Applied Research and Innovation Branch

A Tool to Estimate Contract Time

Guillermo Nevett
Paul M. Goodrum
University of Colorado

Report No. CDOT-2019-07
July 2019

The contents of this report reflect the views of the author(s), who is(are) responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views of the Colorado Department of Transportation. This report does not constitute a standard, specification, or regulation.

Table of Contents

Table of Contents *2*

Technical Report Documentation Page *3*

Chapter 1 - Introduction *4*

Literature Review **6**
Background and Terminology 6
Typical Methods used for Duration Estimation 8
Artificial Neural Networks (ANN) 13

Table 1.2. Prior Construction Focused ANN Models **18**

Chapter 2 – Methodology *20*

The Multiple Linear Regression Approach **20**
Construction Quantities 20
Project Characteristics 20
Durations 22

The Multiple Linear Regression Mechanics **22**
Data Preformatting 22
Variable Grouping 23
Visual Normality Check 25
Multicollinearity Assessments 26
Automatic Variable Selection 27

Analyzing the Model **28**
Interpreting the Regression Coefficients for Continuous Variables 29
Understanding coefficients in MLR models 30

Overall model performance ..... **xxxiv**
Test for Significance of the Regression Coefficients xxxiv

Data Split ..... **xxxiv**

Model Validation ..... **xxxv**
Testing the Underlying Assumptions xxxv

The Artificial Neural Network Approach ..... **xxxvi**
Data for Model Development xxxvi
Model Training xxxix
Overfitting and How to Mitigate it xli
Model Testing xliii

Results and Discussion ..... *xliii*

Chapter 3 – The Estimating Contract Time Tool ..... *xlv*

Chapter 4 – References ..... *lvii*

Appendix A – Steps for Tool Server Installation ..... *lx*

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
CDOT-2019-07		

4. Title and Subtitle	5. Report Date
A Tool to Estimate Contract Time Construction Production Rates for Estimating Contract Time	July, 2019
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Guillermo Nevett Paul Goodrum	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Regents of University of Colorado 1800 Grant Street STE 600 Denver, CO 80203	11. Contract or Grant No. TPF – 5(260) Project 4 PO 411009667

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Colorado Department of Transportation - Research 4201 E. Arkansas Ave. Denver, CO 80222	Final, research and report-writing spanned from 8/1/2015 to 4/30/19
	14. Sponsoring Agency Code

15. Supplementary Notes
Prepared in cooperation with the Colorado Department of Transportation

16. Abstract

It is in the best interest of the traveling public for state transportation agency (STA) construction projects to be completed in the least amount of time possible, to minimize lane closures and construction impacts. In addition, the public takes note regarding activity on construction sites and on occasion, STAs receives complaints when it appears that road construction is not being aggressively pursued. To better estimate durations, this project conducted an exhaustive literature review that allowed the researchers to identify state of the art practices for contract time determination. This literature review oriented the research team to what they believe was the best approach to estimate contract times using historical data. At First, a MS Excel tool was the more desirable tool, but after further research, the team designed and created an Artificial Neural Network (ANN) based tool, that provided more accurate estimates, and is easier to use.

Implementation

The research produced a tool to estimate the duration of Transportation Construction Projects. The tool was created by studying historical data. Said data was then used to create a machine learning algorithm that estimates project durations. This algorithm was then complemented with a user interface (web application) to estimate durations. The website can be accessed through computers, tablets, and smartphones, where a conversion tool and user manual can be found.

17. Keywords	18. Distribution Statement
CDOT, Construction, Contract Time, Estimation, Project Duration, Neural Networks	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	63	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Chapter 1 - Introduction

The objective of this research was to develop a tool to effectively estimate contract time for future construction projects. This will improve the accuracy and consistency of estimating contract time, leading to shorter construction periods and reduced construction impacts to the traveling public. It will also improve the efficiency of creating estimates, thereby saving the designer's time. The implementation product of this research study is a contract time estimating tool that provides state transportation agencies the flexibility of estimating contract time across a broad range of project types. The tool was developed from a database containing state transportation agency historical project data. The tool includes parametric models that can be used to estimate contract time based on project quantities for specific project types based on both size and scope.

Estimating durations of highway construction is a challenge. State Transportation Agencies (STAs) often have to develop an estimate during project development, when many of the project parameters are in flux. Furthermore, STAs' development of an accurate duration requires knowledge of both construction means and methods as well as contractor strategy, which relies of keen insights into the resource constraints that drive contractors' decisions (Zhai et al. 2016). In order to help address this challenge, the Federal Highway Administration (FHWA) developed a Guide For Construction Time Determination Procedures, but even when using it, State Transportation Agencies (STAs) struggle to calculate accurate estimates of project durations (Zhai et al. 2016). These inaccurate estimates, which are often set in the contract documents, affect owners (STAs) and road users, i.e., taxpayers. Some of the consequences of inaccurate estimates are: (1) requiring the STA to allocate extra resources; (2) road users (taxpayers) spending more time driving; (3) longer commutes; (4) and unsafe roads (due to construction hazards) (Zhai et al. 2016).

In order to achieve accurate estimates, it is necessary to understand what drives the durations of projects and what are the actual production rates. Generally, contract times are determined by production rates (e.g. Critical Path Method or Bar Charts) or predicting the duration using cost (e.g .Estimated Cost Method). If the preferred method is based on production rates, the accuracy of the estimate is only as good as the accuracy of the production rates (Jiang and Wu 2007). Achieving accurate production rates and estimates require experienced engineers.

Alternatively, *estimated cost method* (also referred to as regression method) models can be used quickly and do not require such expertise (FHWA 2002).

Zhai et al. (2016) conducted a study to gather information about the different methods and how they are used among STAs. In this study, they determined that only 29 of the 50 STAs have their time determination procedures available online and are subject to inaccuracy. This inaccuracy was explained with mean absolute percent errors (MAPE) (Equation 1.1). Having such inaccurate models is yet another incentive to determine what are the primary drivers of a project's duration, in order to be able to come up with more realistic estimates.

$$MAPE = \left|\frac{Predicted\ duration - Observed\ duration}{Observed\ duration}\right| * 100 \qquad [1.1, \text{APE}]$$

In order to understand why estimates are inaccurate, it is necessary to explain *which* factors have a relationship with project durations, which is a focus of the described research. To determine which are these factors, this research will explore parametric and nonparametric methods for duration estimation. The team chose Multiple Linear Regression (MLR) as the parametric method and Artificial Neural Networks (ANNs), was the chosen to be the studied nonparametric method.

Another issue influenced by production rates and duration estimates is the ability to accurately measure productivity over time. Several difficulties of measuring productivity have been studied over the years. One difficulty highlighted by Goodrum et al. (2002) is the challenge of accurately measuring construction inflation in order to develop accurate measures of real industry output. Challenges noted by others include the heterogenous nature of the construction industry, lack of consistent data standards, and lack of consensus about the techniques for measuring different inputs and outputs of construction productivity (Building Futures Council (2006)). Such difficulties lead to differences in the way productivity is measured. Vereen et al. (2016) explained how some of these differences occur. They measured labor productivity using the same metric but different data sources. A total of four different combinations of the data sources were used by Vereen et al. (2016), and unfortunately none of data sources produced similar results. They concluded that, depending on the input and output data sources, productivity has been increasing or decreasing. The studies by Goodrum et al. (2002), the Building Futures Council (2006), and Vereen et al. (2016) might help explain why there are two

main schools of thought when it comes to construction productivity trends. First some, like Teicholz (2013), suggest that productivity in construction has been declining. On the other hand, there are economists like Sveikauskas et al. (2016), that suggest that productivity has been increasing since 2006.

Highway construction is a multi-billion dollar industry, and the expenditure trend is expected to grow (US Census Bureau Construction Expenditures 2018). Considering its effect on both the overall construction industry and impact on national and regional economic growth, a better understanding of the factors that influence project durations and how to better estimate durations too is warranted. As a result, this report addresses three main issues:

(1) Determine which factors are most influential on a highway project's durations;

(2) Determine if nonparametric approaches to determine such durations, can produce more accurate time estimates compared to traditional parametric approaches, such as Multiple Linear Regression (MLR) models; and

(3) Create a metric that allows to explain the industry-level productivity trends in highway construction over the last 14 years.

Literature Review

Background and Terminology

Part of this report explores the differences between parametric and non-parametric models. *Parametric* modeling uses data that follow certain rules – or parameters – to estimate the value of one dependent variable. In order to estimate the dependent variable, one or several independent variables are used. Both, the dependent and the independent variables, have to follow the parameters established for the kind of relationship used in the modeling (Sheskin 2003). In the case of parametric time estimation, cost has been the most popular independent variable used to predict the duration (dependent variable) of a project (Herbsman and Ellis 1995). On the other hand, *Non-parametric* models do not rely on an assumed distribution or other type of parameter but instead is determined from data. The term *non-parametric* is not meant to imply that such models completely lack parameters but that the number and nature of the parameters are flexible and not fixed in advance. Furthermore, Conover (1980) states that a statistical method can be considered nonparametric if the distribution of the data is unspecified. This means that nonparametric methods can ignore assumptions that are used to analyze data with parametric methods.

Jiang and Wu (2004) conducted research to analyze the factors affecting production rates as well as factors affecting durations in highway construction projects. After determining these factors, they determined the relationships (e.g. linear, exponential, and logarithmic) that best fitted their data, between a project's cost estimate and its duration. These relationships allowed them to create models that help predict project durations during the different stages of a project. One model was created per project type, namely, resurfacing, bridge replacement, among others. This method provided satisfactory accuracy for time estimation on the 95% confidence interval. The models' mean accuracies, within this confidence interval, ranged from ± 0.2 σ (New Road Construction) to ± 0.9 σ (Bridge Replacement (County Road)). Taylor et al. (2013) developed a similar method but included more estimators in the regression equations. Instead of developing an individual model per type of project, they used the type of project as a prediction factor and also included bid quantities as independent variables. In this research project, they came to the conclusion that a linear regression equation is satisfactorily accurate (See Table 1.1) for projects above $1M, suggesting that for smaller projects, unit rate-based estimators were more accurate. The median absolute percent error in Table 1 refers to the median value of the computed APEs (Equation 1.1). The median percent difference is the median value of the percent differences, computed according to Equation 1.2.

$$\Delta\% = \frac{Predicted\ duration - Observed\ duration}{Observed\ duration} * 100 \quad \text{[1.2, Percent Difference]}$$

Table 1.1. Taylor et al. (2013) Accuracy Results

Project Type	Sub-group	Absolute Percent Error		Percent Difference	
		Mean	Median	Mean	Median
Limited Access	All Projects	53.31%	27.33%	28.51%	-1.21%
	Only >$1,000,000	70.36%	28.89%	48.73%	8.13%
	Only >$3,000,000	32.87%	21.53%	13.21%	3.96%
Open access	All Projects	189.59%	75.89%	150.35%	45.66%
	Only >$1,000,000	61.26%	34.98%	26.34%	1.23%
	Only >2,000,000	60.82%	23.36%	38.76%	2.92%
New route	All Projects	206.09%	69.78%	177.56%	36.96%
	Only >$1,000,000	72.31%	54.69%	28.02%	10.70%
	Western KY	66.02%	43.78%	36.46%	2.81%
	Central KY	287.25%	60.75%	270.73%	48.83%
	Eastern KY	148.98%	61.60%	114.21%	34.03%
	Western KY Only >$1,000,000	91.12%	33.37%	67.32%	9.12%
	Eastern KY Only >$1,000,000	77.23%	45.35%	55.53%	33.45%
Bridge Rehab	All Projects	102.00%	48.73%	73.82%	17.16%
	Western KY	106.43%	66.90%	81.74%	66.90%
	Only >$1,000,000	60.06%	77.20%	-50.44%	-77.26%
Bridge Replacement	All Projects	57.77%	35.77%	32.27%	0.47%
	Only >$1,000,000	27.36%	17.03%	9.98%	5.67%

Woldesenbet (2010), demonstrated that location and traffic condition are big influencers of productivity in highway construction. This is due to proximity to metropolitan areas, type of terrain, and how they affect delivery of materials to the jobsite, which why this research is including project locations as a factor that might help explain the project durations.

Typical Methods used for Duration Estimation

According to the FHWA (2002), there are several methods used to estimate the duration of highway construction projects. However, the FHWA recommends one of three methods: 1)

Critical Path Method, 2) Bar Charts Method, and 3) Estimated Cost Method. The first two methods rely highly on the accuracy of the production rates and require skilled engineers to produce reliable estimates due to required expertise in construction methods in order to replicate realistic construction schedules. On the contrary, the estimated cost method relies on the relationship between cost and duration to produce the estimates and does not require as much skill and experience to create estimates. That being said, the engineers' judgement is still key on determining whether the estimates produced are reasonable.

Critical Path Method (CPM)

The Federal Highway Administration (2002) describes the CPM as an analysis of the relationships amongst activities. CPMs emphasize the relationships of activities that must be completed in order to start a succeeding activity. The CPM shows such relationships in a diagrammatical way. These diagrams also contain information about the time required to complete each task and the float that each activity has (total float and free float). Total float is the time that an activity can be delayed without delaying the entire project and free float is the time an activity can be delayed without affecting its successors. In the critical path method, information is shown about which activities will cause a change in the project's completion day if delayed.

The FHWA (2002) defines the five steps required to develop a CPM: (1) project breakdown by activities; (2) defining the relationships between activities, specifically, which activities need to be completed (preceding) before the beginning of another activity (succeeding) or whether activities can be done simultaneously; (3) develop a graphic representation of the relationships defined in the previous step; (4) by using production rates, each activity's duration is estimated and shown in the diagram, along with float, early start, and early finish; and (5) with the use of the CPM, the ideal amount of work days required to complete the project can be estimated.

CPM is popular among DOTs because of a number of advantages. Some of those advantages include: the ability to perform analysis of delays and how to lessen them, activity breakdown visualization, and, as stated in the name, track the critical path of a project. Basically, CPMs help calculate contract time while showing the hierarchy of operations (Herbsman and Ellis 1995; Khallaf et al. 2016). On the downside, in order to achieve accurate estimates, CPM schedules require experienced labor and reliable production rates (FHWA 2002).

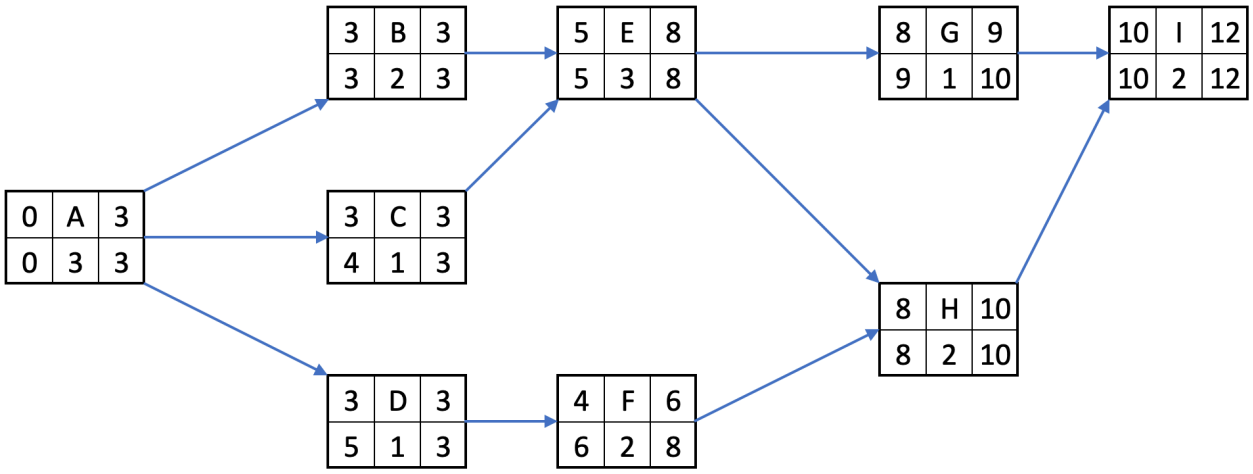


Figure 1.1. CPM

Louisiana DOT is one STA that uses CPM to estimate project durations. Louisiana DOT uses production rate-based time determination tool (Lotus), developed by (McCrary et al. 1995). They developed a set of three templates to estimate project durations. The templates, Lotus 1-2-3, have their own sets of production rates. Quantities of each activity are input into the templates. These quantities are then related to production rates to compute the project durations. These templates also have logic-based relations that, along with the production rates, are then used to estimate project durations (Werkmeister et al. 2000). The durations of these activities can be used in bar charts to create project schedules.

Bar Charts Method

Bar Charts, also known as Gantt Charts, display information related to duration using horizontal lines. Each line represents the duration of an activity and the duration is shown in dates, usually on the top of the chart. Like CPM, before constructing a bar chart, work breakdown is conducted in order to determine the list of activities to be depicted in the chart. Unlike the CPM, Bar Charts do not show relationships between activities. However, they do display the overall duration of a project (Mubarak 2015).

When DOTs select this as their duration estimation method, it is because of its advantages, which include facilitating the tracking a project's actual duration and comparing it to the planned one, ease of understanding, and powerful visual impact (FHWA 2002; Herbsman and Ellis 1995). On the other hand, but they do not show the relationships between different

phases of the project, which is a major disadvantage and why they are not recommended for large complex construction projects (FHWA 2002). Another downside for bar charts is that, like with CPM, bar charts rely on the quality of the production rates in order to produce trustworthy estimates

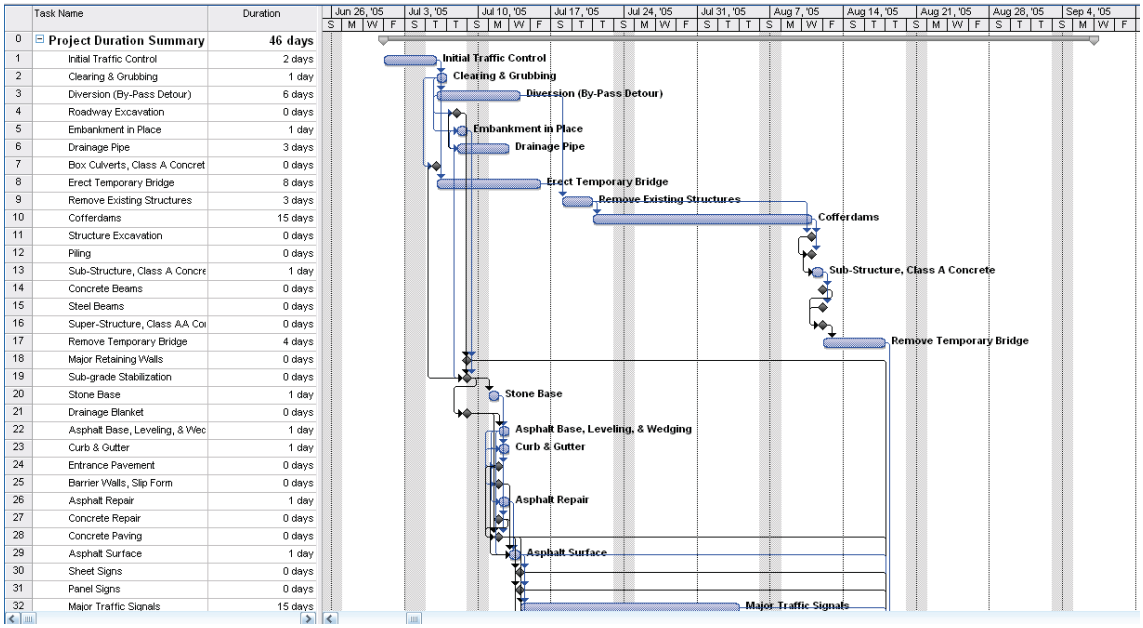


Figure 1.2. Bar charts example (Taylor et al. 2017)

One STA that uses bar charts to create their final schedules is Texas DOT (TxDOT), which incidentally is one of the pioneers in developing custom contract time determination systems (CTDS). TxDOT used a CTDS in which production rates and activity quantities are used to estimate durations. The production rates are developed by engineers with expertise to create better estimates. The production rates are selected by the user and the quantities are introduced in the software. These durations can then be adjusted with correction factors that depend on geography, traffic, and other characteristics pertinent to each project. Finally, the TxDOT Contract Time Determination system creates a output schedule that produces Gantt charts to present the durations.

Estimated Cost Method

Per FHWA (2002), the estimated cost method relates dollar value to duration, based on historical data. This method utilizes different charts to depict cost versus duration for projects

with different locations, traffic volumes, and scope. Following the estimated cost method, Zhai et al. (2016) created MLR models that included cost and other variables, like bid quantities. In their research, Zhai et al. (2016) justify the usage of MLR with the following arguments: (1) bid quantities are highly correlated with duration; (2) highway construction projects are very repetitive; (3) practices are similar across the United States; and (4) parametric modeling utilizes historical data for prediction. In the present research, we try to expand what Zhai et al. (2016) did with their MLR models by adding additional variables and creating a single model for all project types. The new independent variables being studied include: Location, Project Type, Project Condition (New Vs. Old), Terrain Type, and Average Annual Daily Traffic (AADT), as part of the predicting factors. Studying the relationships that all these variables have on project durations is going to be studied in this research. This study will help address the statement by FHWA (2002) about the estimated cost method: "Many items affecting the completion of a project are not taken into consideration when applying this method. Any special features that are unique to a specific project cannot easily be accounted for when using this very simplistic procedure."

The history of the Estimated Cost Method can be summarized and described by reviewing several researches conducted since the 1960s. Fulkerson (1961) conducted an early effort by creating a Simple Linear Regression model to estimate duration of projects (dependent variable) by using cost as an independent variable. From that point on, several variations of the relationship were studied. For example, Falk and Horowitz (1972) studied concave (non-linear) relationships between cost and duration. Additionally, Jiang and Wu (2007) created several regression equations to estimate duration using cost. Each project type (e.g. bridge rehabilitation, resurfacing, etc.) was represented in a different equation. Furthermore, Irfan et al. (2011) created one different exponential per project type (e.g. bridge, resurfacing, or maintenance) to predict duration using cost and contract type. Alternatively, Zhai et al. (2016) created various multiple linear regression models to estimate the duration of highway transportation projects. The different models were developed for different combinations of 'project type', 'project size', and 'accessibility'. These models used not just cost but also specific construction quantities that were observed to be significant predictors of required project duration.

A good example for the estimated cost method is the tool developed by the Kentucky Transportation Cabinet (KTC). KTC uses a custom time determination tool (KY-CTDS). The

tool was first developed in 2000 by (Werkmeister et al. 2000) and later updated by Taylor et al. (2013). In the last update, the final tool created was a regression-based parametric model. This tool lets the user decide which approach to use, based on project size (budget and duration) and type. For larger projects (≥$1,000,000), the tool guides the user into selecting one of the 5 regression models, based on project type (i.e. *Limited Access, Bridge Rehabilitation, New Route, Open Access, and Bridge Replacement).* Each of the regression model has its specific coefficients that allow the user to calculate the duration. These durations are then translated to working days, which help develop the project schedule. For smaller projects (<$1,000,000), a production rate approach was designed. In that approach, the user selects the production rates and inputs them, along with quantities, into a worksheet. The output of range of durations is then used to develop the project schedules.

Combined Methods

Some STAs use a combination of time estimation methods to develop their schedules. A good example is Indiana DOT (INDOT) time estimation tool. Jiang and Wu (2004) developed a tool for INDOT that incorporates *Regression Method, Mean Production Rate Method,* and *Adjustment for Contract Time*. The method chosen to estimate the duration and the adjustment, depends on a what is known about each project. If the production rates cannot be identified, the tool uses the *Regression Method*, which is comprised by 15 different Univariate Regression equations, in which the independent variable is the project's cost and the dependent variable is the duration (in work days). The 15 equations developed were to satisfy the 15 project types (e.g. Asphalt Resurface, Bridge Painting, Bridge Rehabilitation (Deck Replacement), Bridge Rehabilitation (Superstructure Repair), among others) identified in their research. The type of regression (linear, exponential, or logarithmic) varies depending on the type project. This variation is given to the nature of the data. After imputing the cost, these equations then compute the preliminary duration of the projects. On the other hand, if the production rates can be identified, the tool uses *Mean Production Rate Method*. In this method, the installed construction quantities are associated with each of their production rates to develop a preliminary estimate, similar to the *Small Projects* in Ky-CTDS and the Lotus model used by Louisiana DOT.

Artificial Neural Networks (ANN)

Artificial Neural Networks are machine-learning algorithms. Contrary to regression modeling, machine-learning algorithms are nonparametric and specifically nonlinear by nature,

because they do not require to satisfy any assumptions (e.g. normality, linearity, and independence for MLR) in order to analyze the data contained within them. These networks are structured in a similar manner to the neural networks within the human brain. This similarity helps us understand the functioning of ANNs, because both consist of independent units (neurons) that combine to form a larger, more powerful unit. Minimally, ANNs consist of three layers of neurons, including an initial input layer (equivalent to independent variables), followed by hidden layer(s) of interconnected neurons, and finally an output layer (equivalent to dependent variables) (Figure 1.3). The neurons of each layer are connected to all the neurons of the preceding and succeeding layers and each connection has a unique weight. Such weights are optimized during the training process (Klerfors and Huston 1998). To optimize these weights, ANNs use a function called backpropagation (Frandina et al. 2013), which adjusts the individual weights after comparing the predicted and the observed values. This backpropagation, often referred to as training the ANN, happens through several iterations that are specified within the ANN's code. One drawback of backpropagation is the challenge of overfitting the model, which is expanded upon in later chapters of this report.

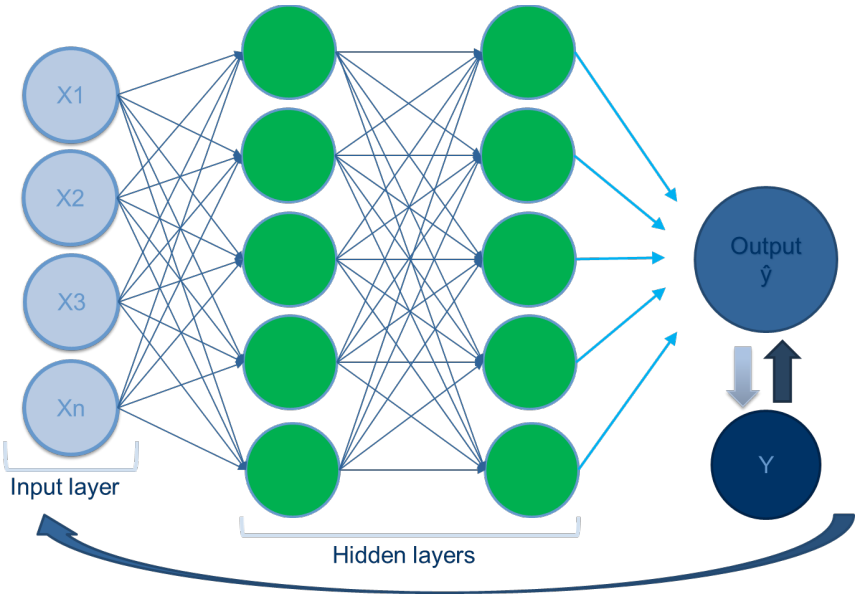


Figure 1.3. Artificial Neural Networks

ANNs have been used to create different estimating models in areas that are not limited to construction. Starting in the late 1990's, Boussabaine and Elhag (1997) compared traditional cost and time estimation methods with Neurofuzzy Models - a combination of ANNs and Fuzzy

Logic - to determine which performed better. They found Neurofuzzy models had stronger performance at predicting cost and duration of vertical construction projects. To create these predictions, they used characteristics (input variables) that described each project. Such characteristics included: height, size, location, market behavior, area, and whether or not excavation was required for the foundation systems. In that same year, Smith and Mason (1997) concluded that ANNs performed better at estimating than regression models. One of the strengths they found was the robustness to assessments required for parametric modeling (e.g. linear or quadratic regressions). The data they predicted was the cost (output variable) of pressure vessels for chemical production. Their input variables were: height, diameter, and wall thickness. Their findings replicated those of Boussabaine and Elhag (1997), concluding that ANNs perform better than commonly used methods, particularly when the data does not meet the assessments required for parametric models. In an effort to explore the use of machine learning to estimate the development effort for software engineering, Finnie et al. (1997) compared the advantages and disadvantages among MLR, ANNs, and Case-Based Reasoning (CBR), another type of machine learning. They concluded that ANNs and CBR outperformed MLR models, but they have a disadvantage. Even when they perform better, ANNs are harder to interpret, meaning that it is harder to explain the contribution of each independent (input) variable in the explanation of the dependent (output) variable. Later on, Hegazy and Ayed (1998) compared optimization methods used within ANNs. They compared the performance of *backpropagation* and *simple optimization* in adjusting weights. The networks they developed were used to predict the cost of highway construction projects (output variable). Their input variables were project type, geographic elements, and project scope.

A couple years later, Kim et al. (2004) explored the same comparison used by Finnie et al. (1997). They compared how ANNs, CBR, and MLR performed at estimating cost (output variable) of multi-story residential construction (considered commercial construction in the US) in Korea. Their input variables were year, area, stories, total units, duration, roof type, foundation system, basement usage, and finishing grades. With this study, they came to the conclusion that ANNs performed better at predicting cost than the other two (CBR and MLR). However, ANNs were still challenging to interpret and they also concluded that CBR performs better over time. This is due to the fact that CBR appends the cases studied to its database, whereas ANNs have to be manually updated. In that same year, Günaydın and Doğan (2004)

studied the capabilities of ANNs when used for cost estimation. They also predicted the cost of multi-story residential construction projects, achieving 93%, using several project characteristics as input variables. Wilmot and Mei (2005) used ANNs to estimate cost indices for several pay items in highway construction. They used market conditions to predict cost indices. With this study, they were able to predict cost indices estimates – not different to the observed ones – at a 95% confidence level. Pewdum et al. (2009) also compared ANNs vs. a conventional estimation method for highway construction projects. In their study, they found that ANNs produce a better cost and duration estimates than the ones obtained by the Earned Value. Some other efforts have been made at estimating cost of different areas of construction in the transportation industry. One study concluded that ANNs performed better at estimating the cost of Road Tunnel Construction than MLR. This study created the ANNs models in a two-step fashion. First, they created models to estimate the values of what would be the inputs of the subsequent model in order to estimate the final cost of projects. (Petroutsatou et al. 2011). A summary of this literature review can be found in Table 1.2.

Table 1.2. Prior Construction Focused ANN Models

AUTHOR AND YEAR	COUNTRY AND JOURNAL	INDUSTRY	RESULTS	DEPENDENT / OUTPUT VARIABLES	INDEPENDENT / INPUT VARIABLES
BOUSSABAINE AND ELHAG, 1997	England - RICS	Commercial Construction	Neurofuzzy models are more accurate than traditional methods for cost and time estimation on commercial buildings.	Cost and Duration	Height, site, foundation, market, area, location, excavation required (y/n)
SMITH AND MASON, 1997	USA – The Engineering Economist	Engineering/ Manufacturing	Neural networks are a better choice than Regressions when the data does not enable the use of a commonly used model, such as linear.	Pressure Vessel Cost	Height, Diameter, and Wall Thickness
FINNIE ET AL., 1997	Australia – Journal of Systems & Software	Software Engineering	Artificial Neural Networks (ANNs) and Case-Based Reasoning (CBR) perform better than Multiple Linear Regression (MLR) at providing software effort estimation.	Estimated Development Effort	System size, programing environment, and general software characteristics
HEGAZY AND AYED, 1998	Canada - JCEM	Highway Construction	Simple Optimization performed better than Backpropagation and Generic Algorithms for cost estimation in highway construction.	Cost	Project type, scope, year, season, location, duration, size, capacity, water body (y/n), and soil condition
KIM ET AL., 2004	Korea – Building and Environment	Residential Construction	ANNs and CBR perform better than MLR for cost estimation in residential buildings. ANN is more accurate but CBR performs better in long term use.	Cost	Year, area, stories, total units, duration, roof type, foundation system, basement usage, and finishing grades
GÜNAYDIN AND DOĞAN, 2004	Turkey – International Journal of Project Management	Multi-Story Residential Construction	Models utilizing ANNs achieved a 93% accuracy for cost estimation of structural systems of buildings.	Cost	Area, ratio of typical floor area to total area, ratio of ground floor area to total area, stories, console direction of the building, foundation system, floor type, and building's core location
WILMOT AND MEI, 2005	USA – JCEM	Highway Construction	An ANN model produced cost indices estimates not significantly different from the observed ones at a 95% confidence level.	Cost indices for different pay items	Price of labor, price of material, price of equipment, pay item quantity, contract duration, contract location, quarter in which contract was let, annual bid volume, bid volume variance, number of plan changes, and changes in standards or specifications
PEWDUM ET AL., 2009	Thailand – Engineering, Construction, and Architectural Management	Highway Construction	ANNs provide better cost and duration forecasting than Earned Value Method.	Cost and Duration	Traffic volume, topography, weather conditions, evaluating date, contract duration (for cost estimation), percent of as planned completion, and percent of actual completion.
PETROUTSATOU ET AL., 2011	Greece - JCEM	Road Tunnel Construction	ANNs provide better cost estimates than MLR for road tunnel construction projects. ANNs were developed in two steps.	Step 1: Steel sets, shotcrete, rockbolts, concrete, steel Step 2: Cost	Step 1: Geology, geological strength index, strain of geological environment, depth of overburden, excavated area of the mined section Step 2: Output variables of step 1

Although previous studies in construction have used ANN, the use of ANN to study the relationships between duration (dependent variable) and construction quantities, cost, and project characteristics (independent variables) is a novel concept. Studying these relationships using ANNs is the primary objective of this research.

Chapter 2 – Methodology

The research examined two mathematical approaches to estimate contract time using various project quantity and project characteristics from CDOT data, including Multiple Linear Regression and Artificial Neural Networks. The process and processes for each approach are described below. The reader will note that there is overlap in the sections describing each approach. This was intentional to allow each description to be viewed independently of each other.

The Multiple Linear Regression Approach

Two different data sources are included in this study. The first one, is a compilation of projects executed by CDOT between 2004 and 2016. The second one is a compilation of projects executed by GDOT during the same period. The data gathered from CDOT was provided by the agency in a Microsoft Access file. Such file contained information about each project in, linked to the individual project IDs. These project IDs allowed the researchers to link the variables (characteristics and quantities) to each individual project. These variables were then used in the analysis of this research. A summary of this data can be found in Table 2.1.

Table 2.1. Data Source

State	***n***	***Criterion Measure***	***Independent Variables***	***Min (2003) US$***	***Max (2003) US$***	***Mean (2003) US$***
Colorado	1500	Charge Days	23	33,284	60,035,291	2,783,012

Construction Quantities

The construction quantities refer to the amount required to be installed, per bid item for the completion of a project. For Colorado, construction quantities were extracted from bid tabulations, *after* the completion of project, i.e. installed quantities.

Project Characteristics

The project characteristics can be separated into two different types, continuous and categorical variables.

The continuous variables are:

- Cost, measured in 2003 USD, converted using the National Highway Construction Cost Index (NHCCI) and

- Annual Average Daily Traffic (AADT), which describes the number of axels that use the road in which the project is being executed.

The categorical variables, represented in the model as dummy variables, are:

- Project Type, a description of the category of work to be executed (e.g. bridge rehabilitation, resurfacing, and road widening,
- Terrain Type, which describes topography in which the project is being executed.
- Project Condition, whether the project is a new project or a revamp of an existing project, and
- Project Size with three levels, small (between $0 and $1,000,000), medium (over $1,000,000 and under $10,000,000], and large ($10,000,000 and over)

Dummy variables refer to a categorical variable with *n* levels represented by *n-1* binary variables. For example, project size has three levels (s1, s2, and s3) that are represented by s2 and s3. In this case, if a project is size=1, the values for s2 and s3 would be 0. The interpretation of the coefficients for s2 and s3 would be relative to projects of size = 1. More about this interpretation is after the model interpretation.

The variable *Project Type* refers to the type of project executed on each contract (as presented by CDOT). This variable is presented as a dummy variable, in the same way as project size. In this case, the variable is a factor variable with 23 levels. Levels 1 through 22 are presented as a binary variable. In this case, when all levels (1 though 22) equal 0, project type = 23. A list of all project types is presented in Table 2.2.

Table 2.2. Project Types dummy variables

Project Type	Variable	Project Type	Variable
Resurfacing	Type1	New Construction	Type13
Bridge Restoration/Rehabilitation	Type2	Rest Area	Type14
Bridge Replacement	Type3	Noise Walls	Type15
Restoration/Rehabilitation	Type4	Landscaping	Type16
Safety	Type5	Miscellaneous	Type17
Hazardous Locations	Type6	Enhancement	Type18
Rail/Highway Separation	Type7	Planning	Type19
Transit System Management (TSM)	Type8	Major Surface Treatment	Type20
Traffic Signals	Type9	Minor Surface Treatment	Type21
Minor Widening	Type10	Routine Maintenance	Type22

Major Widening	Type11	Other	Type23
Reconstruction	Type12		

Note: these Project Types are assigned by CDOT and are not modified by the author.

Durations

The duration for the projects that comprise the data is expressed in number of *charge days*. *Charge days* refer to the number of days in which work is actually performed in a project. *Charge days* account for weekends not worked, weather related stoppage days, accident related stoppage days or how many days per week were worked during a project. This measure represents an advantage when compared to other states that only collect data in terms of *Calendar* Days. States that don't measure *Charge* Days need to use factors to convert *calendar days* into *charge* days, in order to have durations in contractual language (Werkmeister et al. 2000). Zhai et al. (2016) created a tool that generated estimates based on *calendar days.* They offered an option to adjust this for *work days,* but these are based purely on estimates. These estimates are due to uncertainties about unforeseen stoppages. Other considerations for project durations is that neither of the databases include the type of shifts were worked on any project. They also did not include how long the work days were, in terms of hours worked per day. To account for the difference in durations, this research is going to use a similar approach. In order to come up with a *calendar day*-to-*work day* coefficient the average number of days worked during construction season will be used and assigned to projects according to the months in which they were executed.

The Multiple Linear Regression Mechanics

The specific approach used to analyze the data using Multiple Linear Regression is described below in detail.

Data Preformatting

In order to incorporate all the data in a single model, every project was formatted to fit a template created by the team (Figure 2.4). This involved combining all projects into a single file, where each row represents a project and each column represents a variable. This process also included filling blanks with zeroes. For example, a project in which concrete was not used, should have "0" as the value for such variable. Also, since not all the projects were executed in the same year, all projects' costs were transformed to 2003 USD to have consistency across projects. To achieve this transformation, the team used the National Highway Construction Cost

Index (NHCCI). This index tracks the most important factors affecting construction costs in transportation projects. As described by Shahandashti and Ashuri (2015), it is an output index, which means that FHWA measures items contained in construction cost as charged by the contractors, including overhead, material, labor, equipment, and profit.

CID	Rock Exca	Much Exc	Structural	Unclassifi	Embankm	Structural	Concrete	Asphalt P	Aggregati	Asphalt R	Concrete	Concrete	Pavemen
C15919	7815	0	0	0	0	0	0	34764.19	14251.244	0	0	0	4(
C17637	0	0	0	4453	0	0	0	20139.32	13652.324	0	0	0	1
C19941	0	0	270	0	0	43.8	0	30312.77	113.4	0	219.6	219.6	
C13216	0	0	0	0	0	0	0	0	0	0	0	0	
C15223	0	0	0	0	0	0	0	0	0	0	0	0	
C17442	0	0	65	359	0	338	0	87.64	317.352	0	5.8	0	
C19065	0	0	0	0	0	0	0	0	0	0	0	0	
C20236	0	0	0	0	0	0	0	0	0	0	0	0	
C14612	0	0	0	0	0	0	0	0					
C15917A	0	0	0	0	0								

Figure 2.4 Sample of Data Format

Variable Grouping

Since multiple regression was used, the research was observant of the ratio of specimens to the number of independent variables (k). A common practice is to use Fisher's generalization of $n / k > 10$ (Duin 1995). With the purpose of achieving an acceptable ratio of specimens to independent variable, the team proceeded to group variables into variables with similar physical characteristics. For example, PVC piping is one single variable instead of having one per diameter. By doing so, the number of independent variables was reduced from over a thousand to just 23. Of those 23 variables, 17 belong to the category "Construction Quantities" and the other 6 belong to project characteristics (Table 2.3)

Table 2.3. List of Independent Variables

Construction Quantities			**Project Characteristics**
Perforated Pipe	Muck Excavation	Embankment	Cost (2003 USD)
PVC Pipe	Rock Excavation	Asphalt	Size
Concrete Pipe	Concrete	Unclassified Excavation	Type
Class D Concrete	Concrete Pavement	Structural Excavation	New Project
Pavement Marking	Structural Backfill	Asphalt Reclamation	Terrain Type
Aggregate	Sewage		Annual Average Daily Traffic (AADT)

For example, Table 2.4 shows an example of the grouping of sub-variables required for *Concrete Pipe.* In this case, 70 variables were grouped into one variable, *Concrete Pipe*. In this Table, some of the names seem incomplete, due to the data source. This condition made the grouping of

the variables a cumbersome process and inflated the number of sub-variables per independent variable.

Table 2.4. List of Sub-variables Present in the Independent Variable *Concrete Pipe*

15 Inch Reinforced Concrete Pipe	18 Inch Reinforced Concrete Pipe (Jacked)	53x34 Inch Reinforced Concrete Pipe Elliptical	90 Inch Reinforced Concrete Pipe (Complete In Place)
18 Inch Reinforced Concrete Pipe	60x38 Inch Reinforced Concrete Pipe Elliptical	48 Inch Reinforced Concrete Pipe	48 Inch Reinforced Concrete Pipe Special
24 Inch Reinforced Concrete Pipe	66 Inch Reinforced Concrete Pipe	30x19 Inch Reinforced Concrete Pipe Elliptical	78 Inch Reinforced Concrete Pipe Special
30 Inch Reinforced Concrete Pipe	36 Inch Reinforced Concrete Pipe Special	84 Inch Reinforced Concrete Pipe	48 Inch Reinforced Concrete Pipe (Jacked)
36 Inch Reinforced Concrete Pipe	23x14 Inch Reinforced Concrete Pipe Elliptical	45x29 Inch Reinforced Concrete Pipe Elliptical	54 Inch Reinforced Concrete Pipe (Jacked)
42 Inch Reinforced Concrete Pipe	36 Inch Reinforced Concrete Pipe Special (Install Only)	91x58 Inch Reinforced Concrete Pipe Elliptical	24 Inch Reinforced Concrete Pipe Special
54 Inch Reinforced Concrete Pipe	60 Inch Reinforced Concrete Pipe Special (Install Only)	106x68 Inch Reinforced Concrete Pipe Elliptical	83X53 Inch Reinforced Concrete Pipe Elliptical
42 Inch Reinforced Concrete Pipe (Jacked)	24 Inch Reinforced Concrete Pipe (Jacked)	54 Inch Reinforced Concrete Pipe (Special) (Install Only)	68x43 Inch Reinforced Concrete Pipe Elliptical
18 Inch Reinforced Concrete Pipe (Complete In Place)	30 Inch Reinforced Concrete Pipe (Jacked)	66 Inch Reinforced Concrete Pipe (Special) (Install Only)	18 Inch Reinforced Concrete Pipe (Complete In Place)(Instal
24 Inch Reinforced Concrete Pipe (Complete In Place)	27 Inch Reinforced Concrete Pipe (Complete In Place)	12 Inch Reinforced Concrete Pipe (Complete In Place)	24 Inch Reinforced Concrete Pipe (Complete In Place)(Instal
30 Inch Reinforced Concrete Pipe (Complete In Place)	18 Inch Reinforced Concrete Pipe Special	42 Inch Reinforced Concrete Pipe (Complete In Place)	30x19 Inch Reinforced Concrete Pipe Elliptical (Complete In
76x48 Inch Reinforced Concrete Pipe Elliptical	72 Inch Reinforced Concrete Pipe	38x24 Inch Reinforced Concrete Pipe Elliptical	38x24 Inch Reinforced Concrete Pipe Elliptical (Complete In
23x14 Inch Reinforced Concrete Pipe (Complete In Place)	68x43 Inch Reinforcement Concrete Pipe Elliptical	60x38 Inch Reinforced Concrete Pipe Elliptical (CIP)	21 Inch Reinforced Concrete Pipe
36 Inch Reinforced Concrete Pipe (Complete In Place)	12 Inch Reinforced Concrete Pipe	36 Inch Reinforced Concrete Pipe (Jacked)	45x29 Inch Reinforced Concrete Pipe Elliptical (Complete In
48 Inch Reinforced Concrete Pipe (Complete In Place)	21 Inch Reinforced Concrete Pipe (Complete In Place)	54 Inch Reinforced Concrete Pipe (Complete In Place)	76x48 Inch Reinforced Concrete Pipe Elliptical (Complete In
60 Inch Reinforced Concrete Pipe	78 Inch Reinforced Concrete Pipe (Complete In Place)	66 Inch Reinforced Concrete Pipe (Complete In Place)	53x34 Inch Reinforced Concrete Pipe Elliptical (Complete In

60 Inch Reinforced Concrete Pipe (Complete In Place)	34 x 22 Inch Reinforced Concrete Pipe Elliptical (Complete	72 Inch Reinforced Concrete Pipe (Complete In Place)	23x14 Inch Reinforced Concrete Pipe Elliptical (Complete In

Visual Normality Check

Prior to conducting any advanced variable transformations, each variable underwent a graphical normality assessment. This process consists of plotting histograms for each variable and, if needed, standard transformations are executed. Figure 2.1 shows an example of the visual normality assessment and log transformation for the variable *Charge Days*.

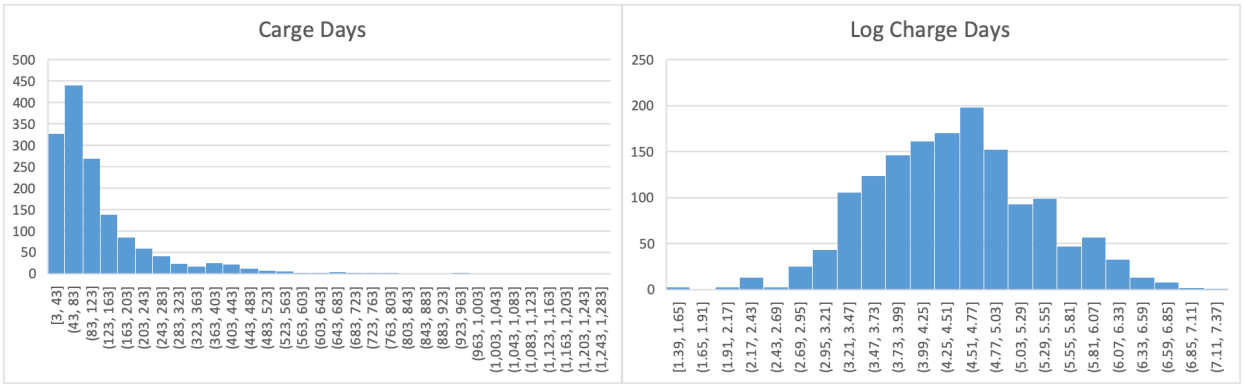


Figure 2.1 Sample of data format

After conducting this visual normality check, all continuous variables were transformed using a logarithmic transformation. Equation 2.1 shows a sample of the transformations performed after this visual normality check. Since *Ln(0)* is undefined, a value of 1 was added to all the independent variables before proceeding to the transformations. This conversion has to be reverted before interpreting any results.

$$\text{Transformed } \boldsymbol{concrete} = Ln(\boldsymbol{concrete} + 1) \quad \text{[2.1, Sample log transformation]}$$

Transforming all variables using a log transformation has one major benefit. When interpreting the coefficients, the concept of elasticity can be applied. This concept refers to how the response (dependent) variable changes compared to the predictor (independent) variable. This change is represented with percent changes. For example, Equation 2.2 shows a log-log transformation equation used in the analyses. In this case, with each 1% increment in *Y, X* would

experience β_1 % change (Duin 1995). More about individual interpretations is detailed in the individual coefficient interpretations.

$$Ln(Y) = \beta_0 + \beta_1 \cdot Ln(X) \quad \text{[2.2, log-log transformation]}$$

Multicollinearity Assessments

After conducting a visual normality assessment, the Variance Inflation Factor (VIF) was measured, which is an estimated of the degree of collinearity for each variable (Craney and Surles (2002)). While there is not consensus on an acceptable VIF value before multicollinearity should become a concern, a value of 4 is recognized as being a conservative value; any VIF value above 4 indicates the presence of multicollinearity (Kabacoff 2015). Per Table 2.5, none of the variables' VIF exceed 4, therefore the data used for the model remained unchanged.

Table 2.5. VIF Values for Variable Selection.

Variable	***VIF***
Sewage	1.37
Concrete	3.80
Asphalt	3.31
Perforated Pipe	1.24
PVC Pipe	1.39
Concrete Pipe	1.82
Class D Concrete	3.85
Pavement Marking	1.91
Muck Excavation	1.34
Rock Excavation	1.04
Concrete Pavement	1.18
Structural Backfill	2.53
Embankment	1.88
Unclassified Excavation	1.90
Structural Excavation	1.87
COST (2003 USD)	2.47
Asphalt Reclamation	2.85
Aggregate Base	2.23
AADT	1.09

Automatic Variable Selection

For this step of model development, two-way stepwise selection was used. This is an automated variable selection process that combines forward and backward stepwise selection. The process compares forward and backward stepwise and selects the best possible model. The forward method starts by creating a model that has no predictors and includes predictors, one by one, until a stopping criterion is reached. It finally produces one model that only includes those variables present in the best model analyzed. Alternatively, the backward stepwise process begins with a model that includes all the predictors and subtract them one by one, similarly to forward elimination, but in the opposite direction, using the same stopping criterion (Kabacoff 2015). The criterion used in this process is a metric called Akaike Information Criterion (AIC), which is used to measure the quality of the analyzed models. With AIC, a smaller value means a better model considering the use of as few independent variables as possible (James et al. 2013; Kabacoff 2015). Table 2.6 shows a summary of the AIC values for each iteration, until the optimum model was achieved. By using two-way stepwise regression, the number of variables present in the model decreased. These variables are shown in Table 2.7.

Table 2.6. Summary of AIC values produced by the two-way stepwise regression

Start:	AIC=-2110.9
Step 1	AIC=-2110.9
Step 2	AIC=-2112.87
Step 3	AIC=-2114.82
Step 4	AIC=-2116.58
Step 5	AIC=-2118.31
Step 6	AIC=-2119.83
Step 7	AIC=-2121.35
Step 8	AIC=-2122.8
Step 9	AIC=-2123.74
Step 10	AIC=-2123.86

Table 2.7. List of independent variables after stepwise regression

Construction Quantities	**Project Characteristics**
Sewer	Cost (2003 USD)
PVC Pipe	Size
Pavement Marking	Type
Structural Backfill	AADT

Asphalt	
Structural Excavation	

Analyzing the Model

There are several steps that comprise analyzing the model, which include: Estimating the regression coefficients; Interpreting the regression coefficients; Testing for significance of the model (overall); Calculating the coefficient of multiple determination (R^2 and adjusted R^2); and Calculating the standard error of the estimate or residual standard error (RSS). The final model based on two-step stepwise regression is summarized in Table 2.8). The stepwise technique described previously, includes all variables that make the model better. In other words, this automatic variable selection technique can include variables that are not significant, but still help improve the overall model. An example of this would be the presence of *Sewer* and *PVC Pipes* in Table 2.8.

Table 2.8. Output Table from the statistical software

Variable	***Coefficients***	***Std. Error***	***t value***	***p value***
(Intercept)	-1.440242	0.355283	-4.054	5.34E-05
Sewer	-0.012434	0.007016	-1.772	0.07657
PVC Pipes	0.014707	0.007762	1.895	0.05833
Pavement Marking	-0.035238	0.007239	-4.868	1.27E-06
Structural Backfill	0.027228	0.010444	2.607	0.00924
Asphalt	-0.021042	0.0073	-2.882	0.00401
Structural Excavation	0.015237	0.007781	1.958	0.0504
Cost (2003 USD)	0.505969	0.027163	18.627	<2.00E-16
AADT	0.016733	0.01011	1.655	0.09814
Size between $1M and $10M	0.109954	0.0604	1.82	0.06892
Size over $10M	0.217396	0.092127	2.36	0.01843
Type_Resurfacing	-0.069376	0.23399	-0.296	0.7669
Type_Minor Widening	-0.203911	0.240318	-0.849	0.39631
Type_Major Widening	-0.174805	0.24808	-0.705	0.48117
Type_Reconstruction	-0.146697	0.56241	-0.261	0.79426
Type_New Construction	-0.078505	0.564364	-0.139	0.88939
Type_Rest Area	-0.263502	0.236444	-1.114	0.2653
Type_Noise Walls	-0.096163	0.285851	-0.336	0.73662
Type_Landscaping	-0.350158	0.247082	-1.417	0.15667
Type_Miscellaneous	-0.63488	0.235148	-2.7	0.00703
Type_Enhancement	-0.252819	0.2936	-0.861	0.38934
Type_Planning	-0.227456	0.233746	-0.973	0.33069
Type_Bridge Restoration/Rehab.	-0.3172	0.240872	-1.317	0.18811
Type_Major Surface Treatment	-0.355098	0.282957	-1.255	0.20972
Type_Minor Surface Treatment	0.126309	0.249627	0.506	0.61295
Type_Routine Maintenance	-0.098788	0.266376	-0.371	0.7108

Type_Bridge Replacement	-0.503307	0.343892	-1.464	0.14356
Type_Restoration/Rehabilitation	-0.162659	0.257054	-0.633	0.52699
Type_Safety	-0.297805	0.378922	-0.786	0.43205
Type_Hazardous Locations	-0.505773	0.254252	-1.989	0.04688
Type_Rail/Highway Separation	-0.319532	0.256892	-1.244	0.21378
Trans. System Management	-0.918822	0.24458	-3.757	0.00018
Type_Traffic Signals	-0.187167	0.251583	-0.744	0.45704

Interpreting the Regression Coefficients for Continuous Variables

Interpreting regression coefficients is a highly important step, because it helps explain how different construction quantities and project characteristics impact the duration of a project. Since there were transformations conducted prior to creating the final model, the interpretation of the coefficients can be misleading. Since all the continuous variables were transformed using a logarithmic transformation, the interpretation of their coefficients can be done using the concept that economists know as *elasticity.* Elasticity is used to interpret the coefficients of variables when the dependent and independent variables are transformed using a logarithmic transformation. With elasticity, the coefficients represent the percent change of an independent variable with a 1% change in the dependent variable (Fox 2015). The explanation of the continuous variables present in the final model is similar for most of them. However, an individual interpretation of each coefficient is presented next.

Sewer

Sewer represents the amount of linear feet installed of *sewage* pipes. Since the only transformation needed for this variable was the logarithmic transformation, the coefficient *-0.012434* can be interpreted as follows: having all other variables held constant, each -.012434% change in *Sewer* relates to a 1% increase in project durations.

PVC Pipes

PVC Pipes refers to the installed quantities of PVC pipes, expressed in linear feet. Again, since the only transformation needed for this variable was the logarithmic transformation, the coefficient .014707 means that a .014707% increment in the installed feet of PVC Pipes are associated with a 1% increase in project durations.

Pavement marking

Pavement marking refers to the installed quantities of pavement marking for the projects, expressed in linear feet. Since the only transformation needed for this variable was the logarithmic transformation, the coefficient -0.035238 can be interpreted as follows:

having all other variables held constant, for each .035238% decrease in Pavement marking, the duration increases by 1%. This is due to the concept of elasticity.

Structural Backfill

Structural Backfill is a representation of to the number of Cubic Yards of backfill-related items present in the projects. In that way, the coefficient .027228 explains that with a .027228% change on quantity of cubic yards an associated 1% increment is experienced in a project's duration.

Structural Excavation

Structural Excavation refers to the number of Cubic Yards excavated from structures in a highway transportation project. In that way, the coefficient .015237 explains that with a .015237% change on quantity of cubic yards an associated 1% increment is experienced in a project's duration.

Asphalt

Asphalt refers to the number of tons of asphalt pavement used in a highway transportation project. In that way, the coefficient -.021042 explains that with a -.021042% change on quantity of tons an associated 1% increment is experienced in a project's duration.

Cost

Cost refers to the amount of 2003 USD needed to execute a project. The interpretation of this coefficient is the same as the previous variables. In this case, a .505969% increment in the cost of a project represents a 1% increase in the project's duration

AADT

AADT refers to the annual average daily traffic that uses the road associated with a particular project on a daily basis. This metric is actually the number of axels use a road per day. In that way, the coefficient .016733 explains that with a .016733% change on AADT an associated 1% increment is experienced in a project's duration.

Understanding coefficients in MLR models

In order to help understand the interpretation of *partial correlation coefficients*, the author developed a Table with several MLR models. Table 2.9 shows how the coefficients, R^2s, Adjusted R^2s, F-statistics, and p-values change for each model. The models shown below include the final model created by the stepwise regression (Model I). Also shown are the six models that

include cost and exclude variables one at a time (Models II through VIII) to demonstrate each variable's effect on the statistical strength on the regression model and how coefficients change with each variable removed. Model IX is included to show the predictive power of cost and the categorical variables. Finally, Model X is included to show how the coefficient of asphalt changes from negative (with all variables present) to positive.

Table 2.9. Comparison of coefficients with different variables present

	I	II	III	IV	V	VI	VII	VIII	IX	X
Variable	Coeff.	Coeff.	Coeff.	Coeff.	Coeff.	Coeff.	Coeff.	Coeff.	Coeff.	Coeff.
(Intercept)	-1.44*	-1.40*	-1.50*	-1.34*	-1.48*	-1.27*	-1.24*	-1.12*	3.255	3.81*
Sewer	-0.01	-	-	-	-	-	-	-	-0.0	-
PVC Pipe	0.014	0.01	-	-	-	-	-	-	0.03*	-
Pvmnt Mkg	-0.03*	-0.03*	-0.03*	-	-	-	-	-	-0.01*	-
Str. Backfill	0.027*	0.02*	0.02*	0.02*	-	-	-	-	0.06*	-
Asphalt	-0.02*	-0.02*	-0.02*	-0.03*	-0.03*	-	-	-	-0.002	0.001
Str. Exc.	0.015	0.01	0.01	0.01*	0.02*	0.02*	-	-	0.00	-
AADT	0.016	0.01	0.01	0.02*	0.02*	0.02*	0.02*	-	0.04*	-
Cost	0.505*	0.50*	0.51*	0.49*	0.51*	0.48*	0.49*	0.49*	-	-
size2	0.109	0.11	0.10	0.09	0.07	0.06	0.06	0.05	0.69*	0.69*
size3	0.217*	0.21*	0.20*	0.19*	0.17	0.14	0.14	0.11	1.49*	1.55*
type1	-0.06	-0.08	-0.08	-0.08	-0.07	-0.05	-0.08	-0.06	0.064	0.11
type10	-0.20	-0.23	-0.21	-0.22	-0.25	-0.19	-0.27	-0.24	-0.17	-0.30
type11	-0.17	-0.20	-0.18	-0.22	-0.22	-0.20	-0.27	-0.24	0.037	0.05
type12	-0.14	-0.17	-0.16	-0.19	-0.23	-0.24	-0.34	-0.28	-0.11	-0.26
type13	-0.07	-0.09	-0.09	-0.41	-0.44	-0.35	-0.45	-0.42	-0.12	-0.47
type14	-0.26	-0.29	-0.28	-0.33	-0.34	-0.33	-0.39	-0.37	0.050	0.00
type15	-0.09	-0.12	-0.10	-0.14	-0.15	-0.15	-0.20	-0.17	-0.05	-0.09
type16	-0.35	-0.38	-0.37	-0.40	-0.44	-0.41	-0.50*	-0.49*	-0.25	-0.44
type17	-0.63*	-0.65*	-0.64*	-0.74*	-0.80*	-0.87*	-0.98*	-0.95*	-0.53*	-0.86*
type18	-0.25	-0.27	-0.27	-0.28	-0.31	-0.23	-0.32	-0.31	-0.15	-0.32
type19	-0.22	-0.25	-0.25	-0.28	-0.31	-0.30	-0.40	-0.37	-0.19	-0.37
type2	-0.31	-0.34	-0.34	-0.36	-0.39	-0.36	-0.44	-0.42	-0.24	-0.41
type20	-0.35	-0.38	-0.37	-0.44	-0.48	-0.49	-0.59*	-0.61*	-0.30	-0.62
type21	0.126	0.09	0.10	0.05	0.03	0.05	-0.0	-0.01	0.130	-0.0
type22	-0.09	-0.12	-0.12	-0.12	-0.14	-0.10	-0.20	-0.16	-0.07	-0.22
type3	-0.50	-0.53	-0.52	-0.51	-0.53	-0.56	-0.60	-0.56	-0.70	-0.75
type4	-0.16	-0.19	-0.17	-0.18	-0.19	-0.16	-0.26	-0.25	-0.08	-0.23
type5	-0.29	-0.33	-0.30	-0.32	-0.35	-0.29	-0.40	-0.36	-0.20	-0.28
type6	-0.50*	-0.53*	-0.53*	-0.64*	-0.70*	-0.75*	-0.85*	-0.83*	-0.30	-0.64*
type7	-0.31	-0.35	-0.32	-0.37	-0.35	-0.36	-0.41	-0.39	-0.03	0.083
type8	-0.91*	-0.93*	-0.93*	-1.08*	-1.12*	-1.13*	-1.24*	-1.23*	-0.83*	-1.17*
type9	-0.18	-0.22	-0.21	-0.28	-0.31	-0.36	-0.45	-0.43	-0.17	-0.35
RSE	0.5097	0.5101	0.5107	0.5153	0.5164	0.5208	0.5224	0.5222	0.5734	0.5943
R^2	0.6643	0.6635	0.6625	0.6561	0.6544	0.6482	0.6458	0.6444	0.5748	0.5395
Adjusted R^2	0.656	0.6555	0.6547	0.6484	0.6469	0.6409	0.6387	0.6377	0.5646	0.5308
F-stat.	80.45	82.81	85.26	85.77	88.24	89.11	91.65	95.97	56.77	62.05
p-val.	2E-16	2E-16	2E-16	2E-16	2E-16	2E-16	2E-16	2E-16	2E-16	2E-16

Note: (*) denotes significance at $p < .05$

From the Table above, there are several observations worth highlighting. First the model that uses only cost as an independent variable explains about 50% of the variability. On top of that, as variables are included or excluded from the model, the *partial correlation coefficients* change. These changes can be so drastic that coefficients that vary may from positive to negative. Finally, the model that includes more variables (from the Table above), has the greatest amount of explained variability. With the previous observations in mind, it is easier to understand how the *partial correlation coefficients* can be interpreted. The *partial correlation coefficients* are used to explain the influence that one independent variable has over the dependent variable in the presence of the other variables, so the amount of explained variability changes when the number of independent variables changes. The statistical explanation for the change of the coefficients is called *Omitted Variable Bias*. This term explains the way coefficients change in a regression model when excluding – or including - other independent variables in the model (Clarke 2005). A hypothetical practical explanation of negative coefficients could be linked to how the productivity of a certain activity increases when the quantity increases. This would reduce the time required per unit installed. This concept is known as *economies of scale* (Baumers et al. 2016). These two explanations help understand the changes in coefficients between models.

1.1.1.1 Interpreting the Regression Coefficients for Categorical Variables

The categorical variables in this model produce auto-generated dummy variables. This means, that a project can only fulfil one of the dummy variables per category, i.e. a project that is Type2 cannot be Type3, Type4, or any other TypeX, nor a project that is Size2 can be Size1. That being said, the interpretation of this coefficients is different from variables that are continuous. Since these variables were not transformed and they are dummy variables, the general interpretation of these coefficients could be explained as the percent difference shown in Equation 2.3.

$$\Delta Y = (e^{\beta_j} - 1) \quad \text{[2.3, percent difference]},$$

Where:

ΔY: *Percent change in duration*

β_j : *J eth coefficient*

For example, if we replace the coefficient .217396 of Size3 in Equation 5, ($\Delta Y = (e^{.217396} - 1) \approx .24$), we will conclude that a project with Size3 will take 24% longer to execute than a project Size1 (base dummy variable).

Overall model performance

The overall performance of the model is summarized in (Table 2.10). First, the *overall model* is statistically significant. This is shown by the p-value of the F statistic. These low values infer that at least one of the predictor variables has a linear relationship with the response variable (charge days) and that the model is statistically significant. Second, the *coefficient of multiple determination* (R^2) tells us that we can explain 66.4% of the variability in the duration of projects with the variables present in the model. This value drops to 65.6% even after adjusting for the number of independent variables and computing the adjusted R^2. Lastly, the *standard error of the estimate* or *Residual Standard Error* (RSS) of the model is 0.5097 on 1470 degrees of freedom.

Table 2.10. Overall model performance

Statistics of the overall model

F-statistic: 80.45 on 32 and 1301 DF, p-value: < 2.2e-16
Multiple R-squared: 0.6643, Adjusted R-squared: 0.656
Residual standard error: 0.5097 on 1301 degrees of freedom

Test for Significance of the Regression Coefficients

As we can see from Table 11, not all the continuous variables present in the final model are significant, satisfying a level of $\alpha = .1$, but they still improve the model's accuracy levels. For the factor variables (Size and Type) we can also find significance, because at least at one level of each has a p-value < 0.05 (Type6, Type8, Type17, and Size3). Therefore, we can conclude that all the *Type* and *Size* are significant.

Data Split

Prior to the creation of the model, the dataset was split into training and testing subsets containing 90% and 10% of the data, respectively. This was done in order to test the model with

data that was not used in the creation of the model. By doing so, the researcher can provide an unbiased evaluation of the model accuracy.

Model Validation

After creating the model with the training subset, the test subset was used to validate the model's accuracy. To do this, the final model was used to predict the duration of the cases in the testing subset. After predicting the durations, Median Absolute Percent Error (MdAPE) and the Mean Absolute Percent Error (MAPE) Errors were computed. The MdAPE is simple the median of all APEs (Equation 1) and the MAPE is the mean of the APEs. The results obtained for MdAPE and MAPE from the test sample are shown in Table 2.11.

Table 2.11. MdAPE and MAPE for MLR model

	Train Data (CO)		***Test Data (CO)***	
	MdAPE	**MAPE**	**MdAPE**	***MAPE***
MLR	*43.70%*	*44.50%*	*43.18%*	*45.20%*

Testing the Underlying Assumptions

Linearity and Homoschedasticity

An MLR model must fulfil the assumptions of linearity and homoschedasticity (equal variance), which can be assessed by looking at the residual plots of the fitted values (Figure 2.2). Here, we can attest that the fitted values are linear and have equal variance, so the data used fits a linear model.

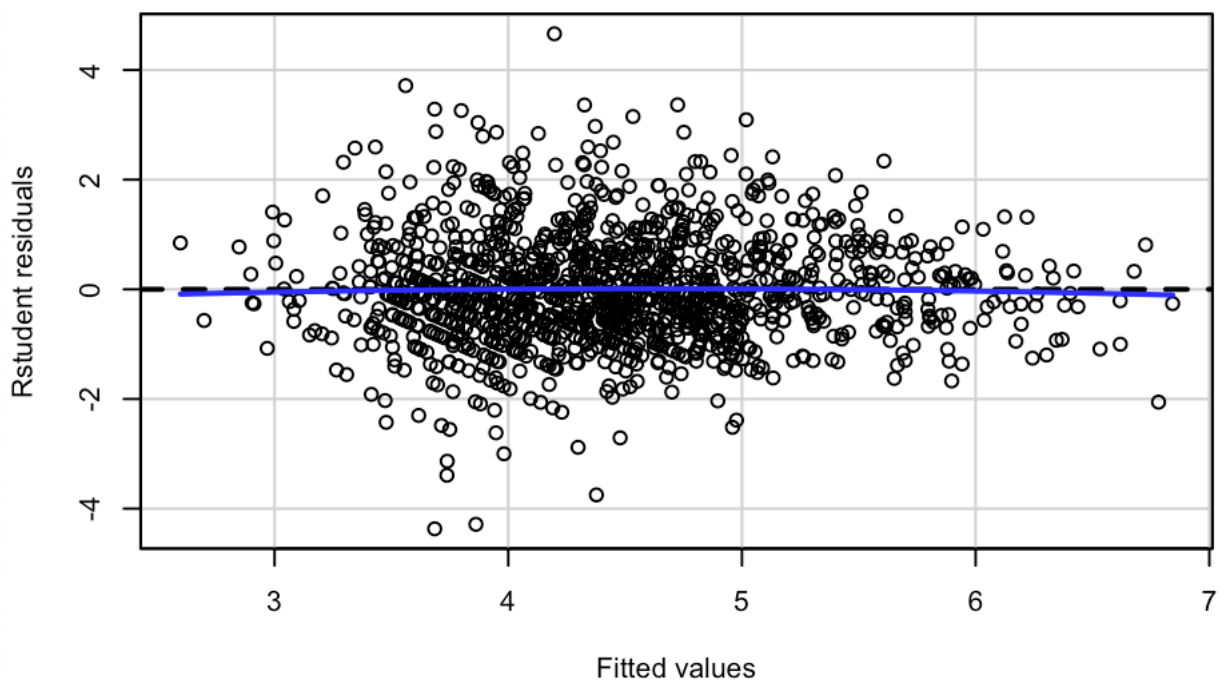


Figure 2.2 Scatterplot of fitted values

Normality

For Linear Regression models, data has to fit a normal distribution. However, the central limit theorem states that when sample sizes are large, the data tends to fit a normal distribution. According to Kwak and Kim, once the sample sizes reaches $n = 30$, it can be assumed that the studentized distribution tends towards a normal distribution. With this in mind, it is safe to assume that, given the sample size used of this model (n>1300), the residuals approximate a normal distribution.

The Artificial Neural Network Approach

Data for Model Development

The data used to develop the model was obtained from Colorado's Department of Transportation (CDOT). The data involved in the creation of the model consisted of historical projects finished for CDOT from 2004 up to and including 2016. After cleaning and organizing the data, over 1,500 projects were considered *usable* for the model. In order for a project to be considered usable, its data recorded had to include observed duration (work days), which is the dependent or output variable and overall project cost. Once the projects were considered usable, more variables were added to the model as input variables, including: Contract (bid) items, Project

Type, Project Size, Terrain Type, Annual Average Daily Traffic, and Project Condition (new or restoration).

In order to reduce the number of variables, the author proceeded to group them into variables with similar physical characteristics. For example, concrete piping is one single variable instead of having one independent variable per pipe diameter. By doing so, the number of independent variables was reduced from over a thousand to just 23, categorized into "Construction Quantities" (17 variables) "Project Characteristics" (6 variables) (Table 2.12)

Table 2.12. List of independent variables

Construction Quantities			**Project Characteristics**
Perforated Pipe	Muck Excavation	Embankment	Cost (2003 USD)
PVC Pipe	Rock Excavation	Asphalt	Size
Concrete Pipe	Concrete	Unclassified Excavation	Type
Class D Concrete	Concrete Pavement	Structural Excavation	New Project
Pavement Marking	Structural Backfill	Asphalt Reclamation	Terrain Type
Aggregate	Sewage		Annual Average Daily Traffic (AADT)

The data collected for this research only includes CDOT data, but it does not necessarily include all the factors that can have a significant effect in project durations. To better explain this, an influence diagram was developed (Figure 3.x). Since this paper is trying to explore all possible explanations of project durations, all the variables available in the data collected were included in the different ANN models developed.

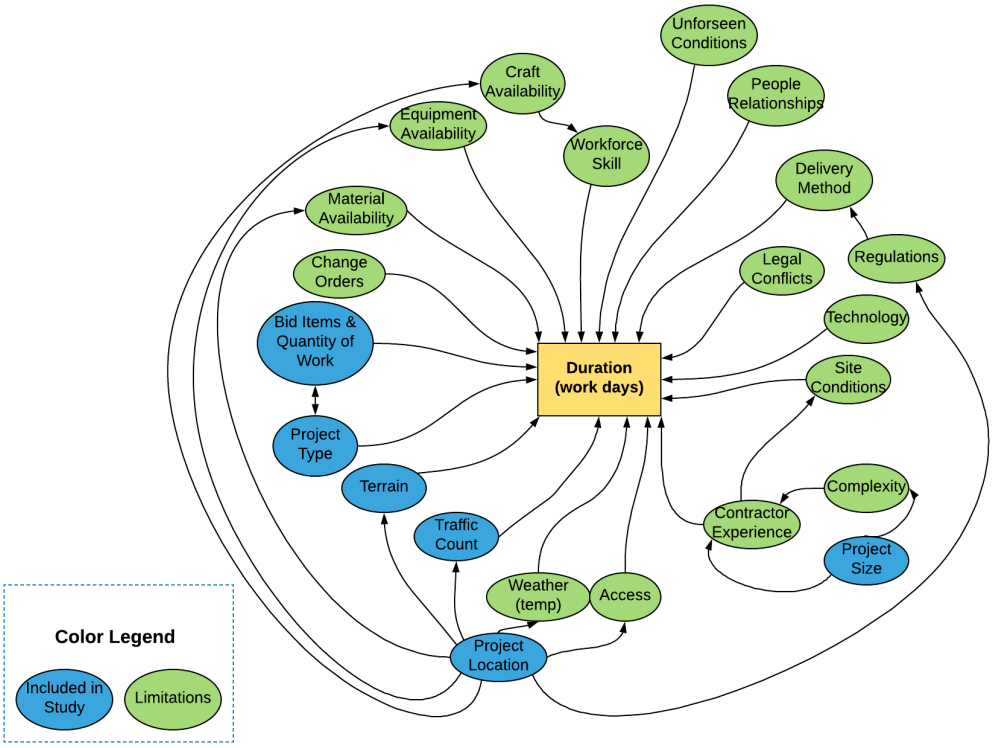


Figure 2.3. Influence Diagram

Project Type

Project Type refers to the category assigned by CDOT to each particular project. CDOT uses 22 different project types. This variable is a categorical variable, and the CDOT categories are shown in Table 2.13.

Table 2.13. List of Project Types

Bridge replacement	Major widening	Planning	Safety
Bridge restore/rehab	Minor surface treatment	Reconstruction	Unassigned
Enhancement	Minor widening	Rest area	Traffic signals
Hazardous locations	Miscellaneous	Restoration/rehab	Resurfacing
Landscaping	New construction	Trans systems management (TSM)	
Major surface treatment	Noise walls	Planning	

*These Project Types are assigned by CDOT and are not modified by the research.

Project Size

Project size refers to the overall cost of the project, expressed in terms of 2003 USD. In order to achieve this, the cost of every project was converted using the National Highway Construction

Cost Index (NHCCI), which measures the historical fluctuations of highway construction costs ("NHCCI / Description - Policy | Federal Highway Administration" n.d.).
Project size is represented as a continuous variable as well as a categorical variable, in which projects are assigned a value – 1, 2, or 3 – depending whether they are under $1M, between $1M and $10M, or over $10M.

Terrain Type

Terrain type refers to the characteristics of the topography in which the project was built. There are 5 categories present in the data, Level (1), Mountainous (2), Plain (3), Rolling (4), and Unassigned (5).

Annual Average Daily Traffic (AADT)

Annual average daily traffic is a metric of the total number of axels that drive through a specific road during a year divided by 2 to represent *cars* and the divided by 365 to calculate *daily traffic.* AADT is also a continuous variable. It is a weighted average of the AADT for the whole project, using mile markers for the project and its respective AADT, the weighted AADT was computed and assigned to each individual project.

Project Condition

Project condition is a categorical variable with two possible outcomes, whether the project is new (0) or is a restoration of an existing project (1).

Model Training

Model training consists on a series of automated iterations conducted by the ANN algorithm (Figure 3.3). This process is done until a set parameter (Median Average Percent Error) reaches its optimum value, i.e. until it stops decreasing. For the training of the model, only 90% of the existing data, randomly selected, was used. This was done to have an *untouched* set with which the model was tested after each iteration. In ANNs, the training is conducted automatically, and it is an equivalent process of the manual processes followed in MLR (e.g. variable selections, autocorrelation assessment, transformations, and unusual observations). Training is the iterative process (Fig. 2.4) in which the algorithm adjusts the weights of the variables after comparing observed versus predicted values if the input variable. This is the process in which the model *learns* and improves the predictability. In order to train the model, the data has to be *fed*. Feeding the data refers to running it through the network to start the training process.

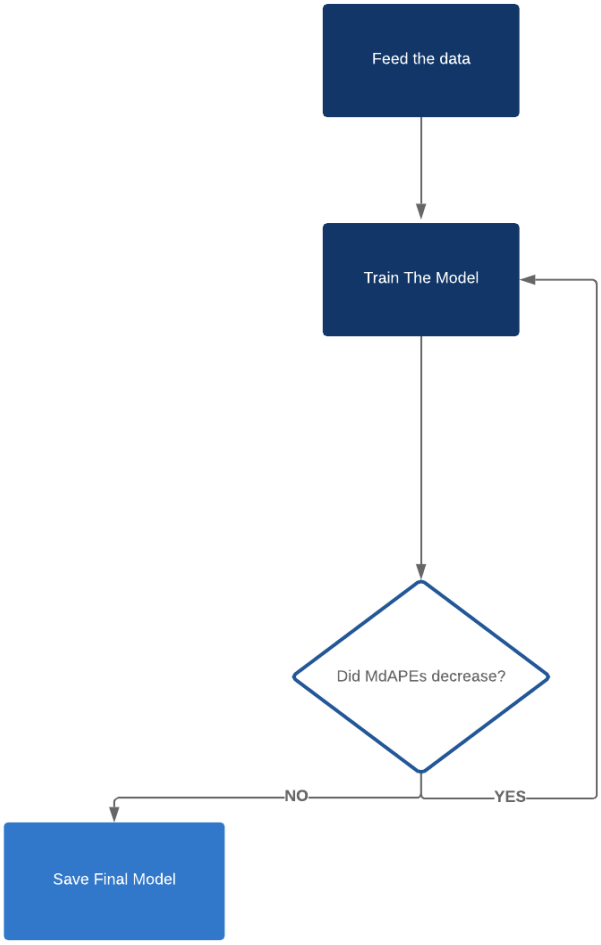


Figure 2.4 Model Training Process

Two metrics were used to measure the accuracy of the model, including Median Absolute Percent Error (MdAPE) and Mean Absolute Percent Error (MAPE), where the absolute percent error is calculated as follows (Equation 2.4). MdAPE and MAPE are computed after each epoch for both, the training and the testing datasets.

$$Absolute\ Percent\ Error = \frac{|Observed\ Duration - Predicted\ Duration|}{Observed\ Duration} * 100 \qquad [2.4, \text{APE}]$$

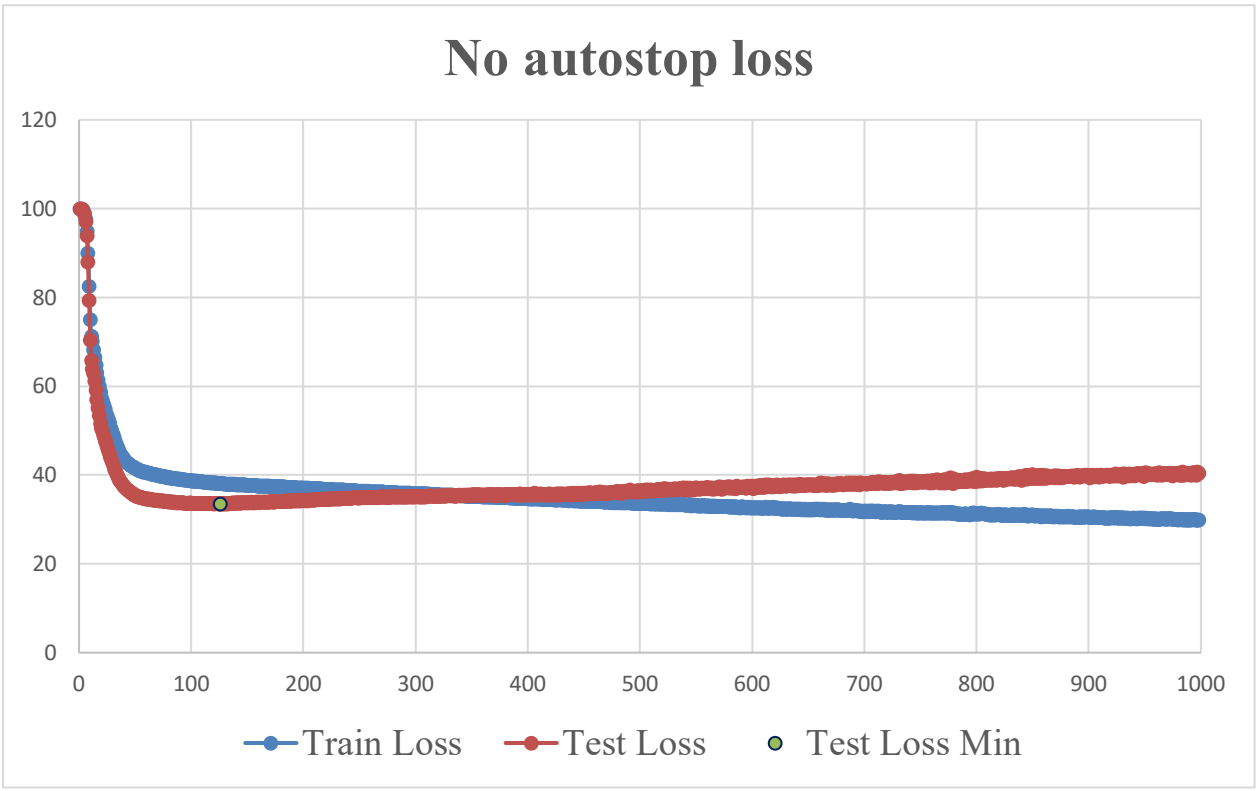
Overfitting and How to Mitigate it

A major issue encountered when training ANNs is what is known as overfitting, which Piotrowski and Napiorkowski (2013) define as the deterioration of the generalizability of the model. Model overfitting results in the decrease in power for predicting new cases. This happens when the model is trained with too many epochs, which causes the model to be extremely accurate at predicting the data in the *train* dataset, but very inaccurate for *test* data or any exterior data.

In order to mitigate overfitting in this research, an *auto stop* function was added to the script. The goal of this function is to stop training the model based on specified parameters. In this case, the parameter was a ratio (Equation 2.5) of the MAPEs from the *train* and *test* (Figure 5) datasets. After every epoch, the MAPE is calculated for each of the datasets and the training stops when the ratio reaches the optimal point, or when it starts decreasing again (see Figure 2.5).

$$Ratio = \frac{MAPE_{Train}}{MAPE_{Test}} \qquad [2.5, \text{Ratio}]$$

-



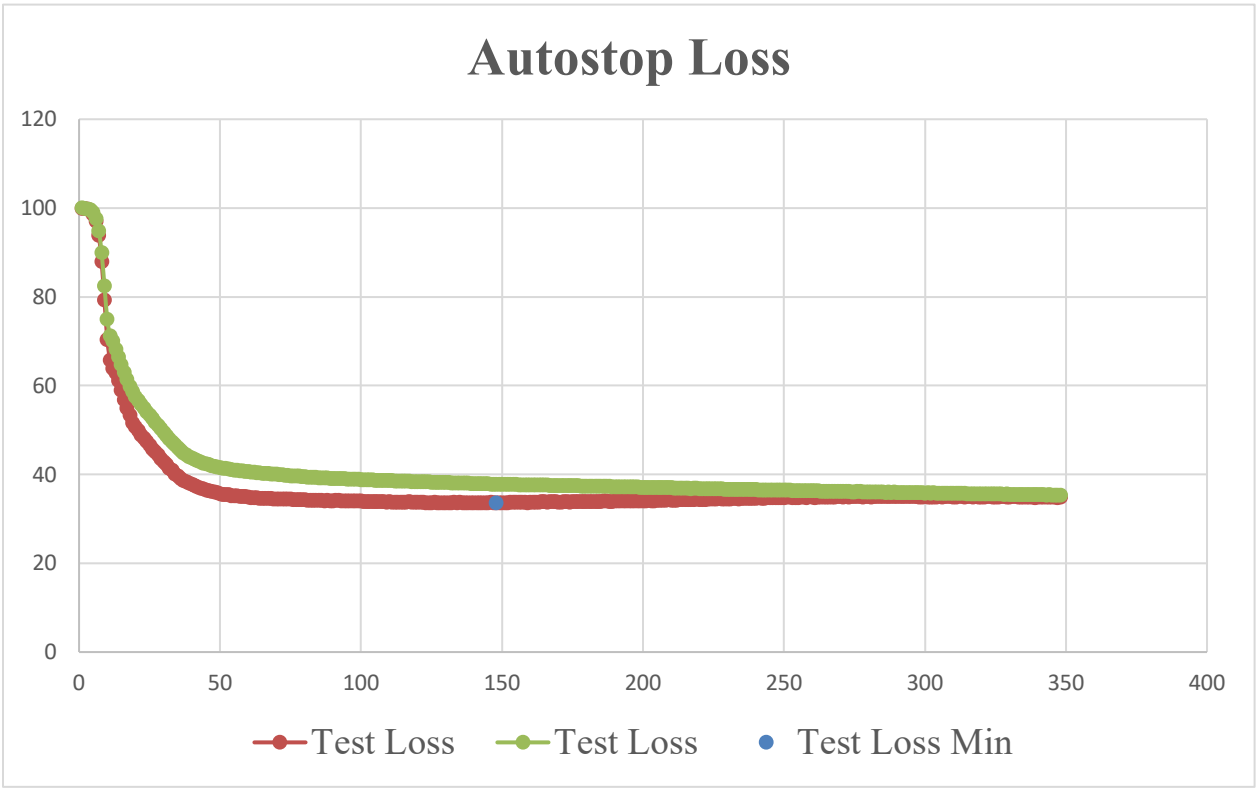


Figure 2.5. Illustration of auto stopping functions

Model Testing

The testing of the model is a process that happens simultaneously to the training by using the 10% of the data randomly left out from the training data set. The model predicts the duration of every test project and compares it to the observed duration. After every epoch, each absolute percent error is computed and the MAPE for that epoch is calculated. This process is also iterative and is performed until the optimal model is achieved.

Results and Discussion

Thus far, the ANN models developed outperform any of the parametric models developed by the author or those found in literature (Table 2.14). One explanation for this performance, might be that most models are based on MLR and, at least with the data used for this research, none of the variables present are normally distributed, which is one of the main requirements to develop a Linear Regression. Of course, transformations are a valid way to overcome this limitation, but the ANNs are robust to this – and other assumptions – allowing a broader interpretation of the available data, thanks to the nonparametric nature of ANNs. Such nature could help find what are the driving factors for highway construction projects' durations that might not even be considered important to the date.

The Kentucky Contract Time Determination System (KY-CTDS) was used as a reference point for a number of reasons. First, it is one of the most recently documented Contract Time Determination Systems for an STA. Second, the MLR approach used by the KY-CTDS relates it to the Estimated Cost Method suggested by the FHWA. Third, KY-CTDS is highly accurate when compared to those methods existing in the literature (Zhai et al. 2016). Lastly, the author created a preliminary MLR model to estimate durations before creating the ANN models used for this paper. This last point is of great importance, because it would help compare previous efforts with those developed by the author.

Table 2.14. Comparison of MLR, ANN, and KY-CTDS

	Train Data (CO)		*Test Data (CO)*	
	MdAPE	**MAPE**	**MdAPE**	***MAPE***
MLR	43.70%	44.50%	43.18%	*45.20%*
ANN	32.78 %	35.11%	25.56%	*32.26%*
*KY-CTDS**			*34.28%*	*61.25%*

* (Zhai et al. 2016)

Summary

The research found that greater accuracy of estimating contract time was obtained using ANN versus MLR. In addition, ANN was found to be more adoptable to future data in the sense that ANN can be more easily trained and therefore revise its estimates when new data is eventually added to the database. In the case of MLR, the analyses would have to be redone if new data was ever added to CDOT's database. Based on these two findings, the decision was made in conjunction with CDOT's oversight time to develop the Estimating Contract Time Tool with ANN.

Chapter 3 – The Estimating Contract Time Tool

Introduction

This tool was designed to estimate Project Durations (charge days) for projects in early stages. The Tool was developed by researchers at the University of Colorado Boulder, with Dr. Paul Goodrum as a Principal Investigator (PI) and Guillermo Nevett as a Research Assistant (RA). The tool was developed using Machine Learning, specifically an Artificial Neural Network (ANN), to estimate project durations. Artificial Neural Networks use historical data to train a model that can then be used to predict a variable. In this case, several project characteristics and construction quantities were used to predict the duration of transportation construction projects.

The data used for the creation of this tool was provided by CDOT to CU's researchers. Over 1,500 projects (from 2004 to 2017) were analyzed in the creation of this tool. The researchers extracted the most relevant variables and grouped them by similarity of characteristics (physical and daily outputs) to reduce the number of variables required to create predictions.

How it works:

ANNs are machine learning algorithms that learn by predicting data outputs and comparing them to historic, actually observed values. This is done to adjust the weights of predictors within the model, in order to incrementally improve the accuracy of the predictions. In this case, the ANN uses historic project data (Project Quantities and Characteristics) to predicted project durations. These predicted durations are then compared to the actual (observed) durations, the ANN computes an error for each duration estimation, and then weights are adjusted within the ANN to minimize overall prediction error. After adjusting the weights, the ANN again estimates the durations of the same projects, computes predicted duration errors, and attempts to adjust weights to minimize the error again. This process is conducted hundreds of times until the lowest possible prediction error is achieved, resulting in the best possible model.

Why ANNs?

ANNs were used for several reasons:

- They are more accurate than anything observed in literature (See Figure 1.1);
- Their ability to adapt to changes that improve productivity (e.g. new technologies, new methodologies, new materials, etc.);

- The final interface is easy to use; and
- Their ability to produce quick estimates.

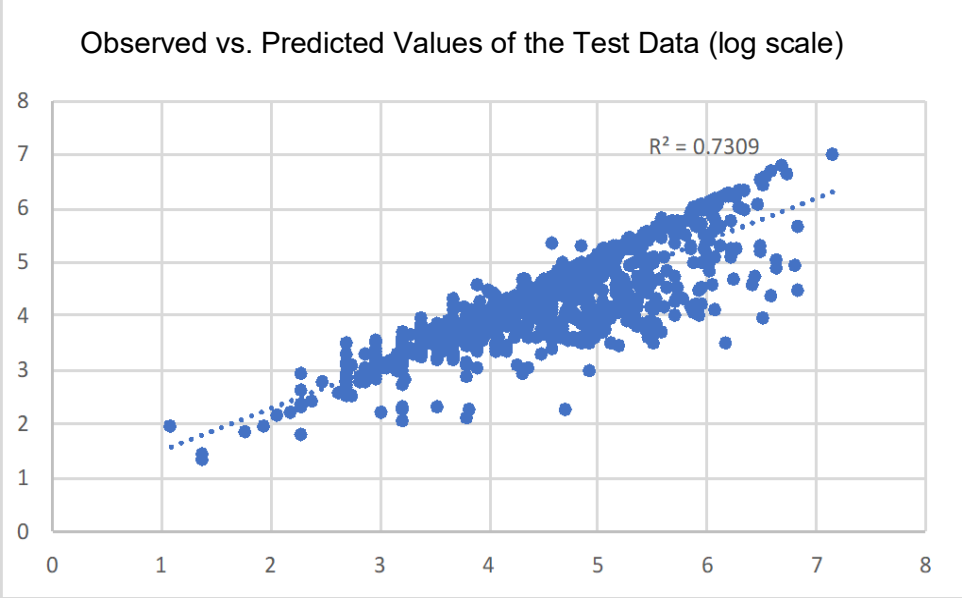


Figure 3.1 Graphical representation of model accuracy

As a quick overview of the tool, the process to develop a contract time estimate includes the following steps.

- Access tool through link on Project Management website https://pmo.codot.gov/contracttimes/.
- Input project characteristics.
- Input item quantities.
- Hit "Compute", tool then tells you a working day range.
- The amount under "Expected" should be entered in the Project Management tool "MS Project Preconstruction Schedule" https://www.codot.gov/business/project-management/business/project-management/scoping under the task "Construct the Project".
- The difference between the "Expected" and the "Expected +30%" should be entered under the task "Schedule Contingency".

More specific details about the process to estimate contract time using the tool is descrbed in the next section.

Interface of the Tool (Figure 3.2)

The tool has two categories of variables: Project Characteristics and Bid Quantities

Project Characteristics (all six variables required)

This category refers to conditions relevant to the project locations and budget. It includes:

Project size: Sub-grouping category used to describe the cost of the project.

Project type: Type of construction that best describes the project.

Terrain type: Describing the topography of the location of the project.

Condition of the project: This refers to whether the project is new or is a reconstruction or renovation.

Engineers estimate (EE): The numerical value of the cost of the project. It has to be in 2003 constant dollar value. To convert current dollar value to constant dollar value, use the quarterly National Highway Construction Cost Index (NHCCI) 2.0 (https://www.fhwa.dot.gov/policy/otps/nhcci/pt1.pdf).

$$EE\ in\ constant\ dollar = \frac{EE\ in\ current\ dollar}{NHCCI\ for\ the\ current\ Quarter}$$

For example: Engineering Estimate in December 2018: $30,000,000; NHCCI for December 2018 is 1.8727 (Choose the nearest quarter).

$$EE\ in\ constant\ dollar = \frac{\$30{,}000{,}000}{1.8727} = \$16{,}019{,}651$$

Annual Average Daily Traffic (AADT): this value is the weighted average of the traffic metrics of the project location.

Bid Quantities (only those that apply)

This category refers to the quantities of installed materials or executed work. Using the conversion tool (section 2.B), input all that apply. If a project doesn't have one of the variables shown in the interface, the value must be **zero.**

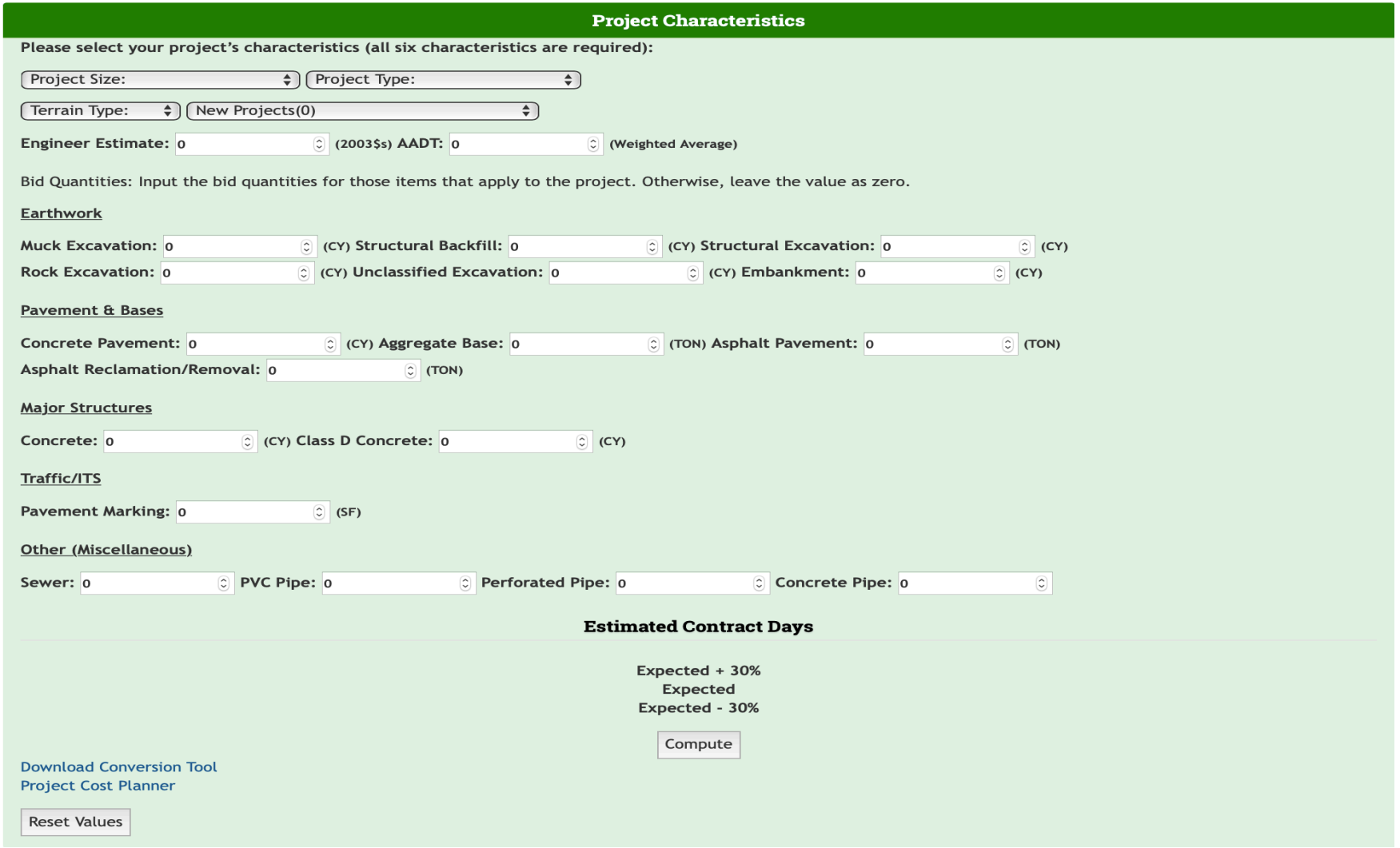


Figure 3.2. Contract Time Tool – Interface

Table 3.1. Details of the bid items used as independent variables.

Name	Description	Unit
Sewer[1]	Sum of all corrugated and culvert pipe quantities	LF
Perforated pipe	Sum of all perforated pipe quantities	LF
PVC pipe	Sum of all PVC pipe quantities	LF
Concrete pipe	Sum of all concrete pipe quantities	LF
Class D concrete	Sum of all Class D concrete quantities	CY
Pavement Marking	Sum of all pavement marking quantities	CY
Muck Excavation	Sum of all muck excavation quantities	CY
Rock excavation	Sum of all rock excavation quantities	CY
Concrete	Sum of all concrete quantities (excludes piping and pavement)	CY
Concrete pavement	Sum of all concrete pavement quantities	Ton
Structural backfill	Sum of all structural backfill quantities	CY
Embankment	Sum of all embankment quantities	CY
Asphalt pavement	Sum of all asphalt pavement quantities	Ton
Unclassified excavation	Sum of all unclassified excavation quantities	CY
Structural excavation	Sum of all structural excavation quantities	CY
Aggregate base	Sum of all aggregate base quantities	Tons

[1] Piping items exclude connections, manholes, and anything not measurable in LF.

Conversion Tool

A complementary conversion tool (Figure 3.3) was designed to facilitate the use of the Contract Time Tool. This tool is needed because of the amount of different bid items that make up one single variable for the model. It is also required because different regions use different units for some project items. In order to run the conversion tool, the user has to allow editing and enable macros in MS Excel. The program will prompt the user to do so.

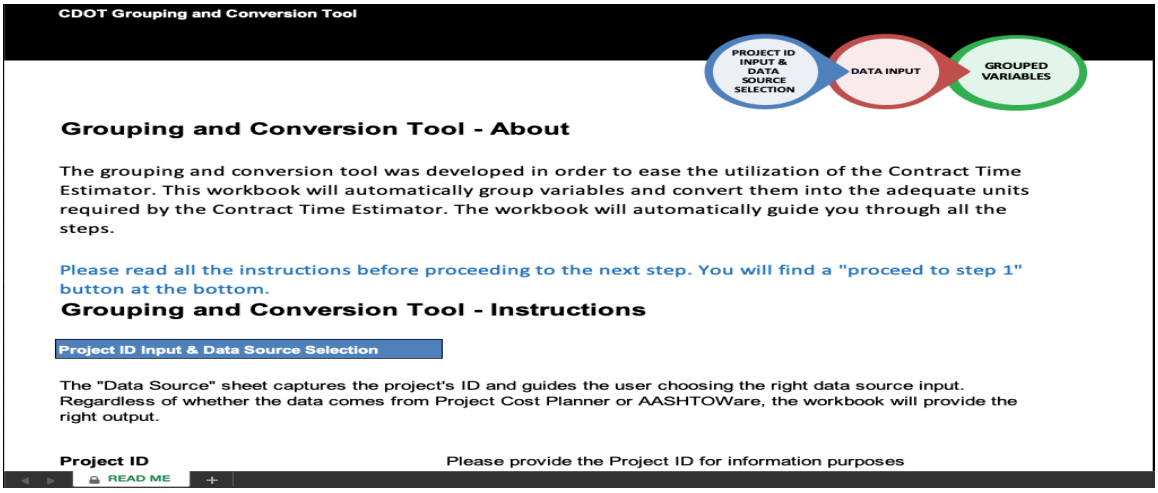


Figure 3.3 Conversion Tool

Since the data source can be from AASHTOWare or Project Cost Planner Tool, the conversion tool has one specific tab for each of the data sources.

Project Cost Planner

For the Project Cost Planner data, the tool will require data to be copied from the tab ***Model_RegionEstimate*** (Figure 3.4) and pasted (using paste special > values) into the Conversion Tool (Figure 3.5). The conversion tool has buttons to guide the users into selecting the appropriate data source.

EARTHWORK	Unit	Qty
203-00060 - Embankment Material (Complete In Place)	CY	80,000
203-00100 - Muck Excavation	CY	15,000
203-00066 - Embankment Material (Complete In Place) (R40)	CY	0
206-00100 - Structure Backfill (Class 1)	CY	0
203-00066 - Embankment Material (Complete In Place) (R40)	CY	0
203-00066 - Embankment Material (Complete In Place) (R40)	CY	0
206-00100 - Structure Backfill (Class 1)	CY	0
203-00066 - Embankment Material (Complete In Place) (R40)	CY	0
206-00360 - Mechanical Reinforcement of Soil	CY	0
203-00066 - Embankment Material (Complete In Place) (R40)	CY	0

A-01 Cost:

PAVEMENT & BASES	Unit	Qty
304-06000 - Aggregate Base Course (Class 6)	TON	100,000
403-00720 - Hot Mix Asphalt (Patching) (Asphalt)	TON	500
403-34721 - Hot Mix Asphalt (Grading SX) (75) (PG 58-28)	TON	50,000
310-00610 - Full Depth Reclamation of Hot Mix Asphalt Pavement (8- 12")	SY	72,500
411-10255 - Emulsified Asphalt (Slow-Setting)	GAL	12,500
403-00721 - Hot Mix Asphalt (Patching) (Asphalt)	SY	0
403-09210 - Stone Matrix Asphalt	TON	0
310-00608 - Full Depth Reclamation of Hot Mix Asphalt Pavement (0- 8")	SY	0
310-00610 - Full Depth Reclamation of Hot Mix Asphalt Pavement (8- 12")	SY	0
403-34721 - Hot Mix Asphalt (Grading SX) (75) (PG 58-28)	TON	0

Figure 3.4 Sample Data from Major Construction Items (Section A in Model_Region Estimate Tab of the Project Cost Planner Tool).

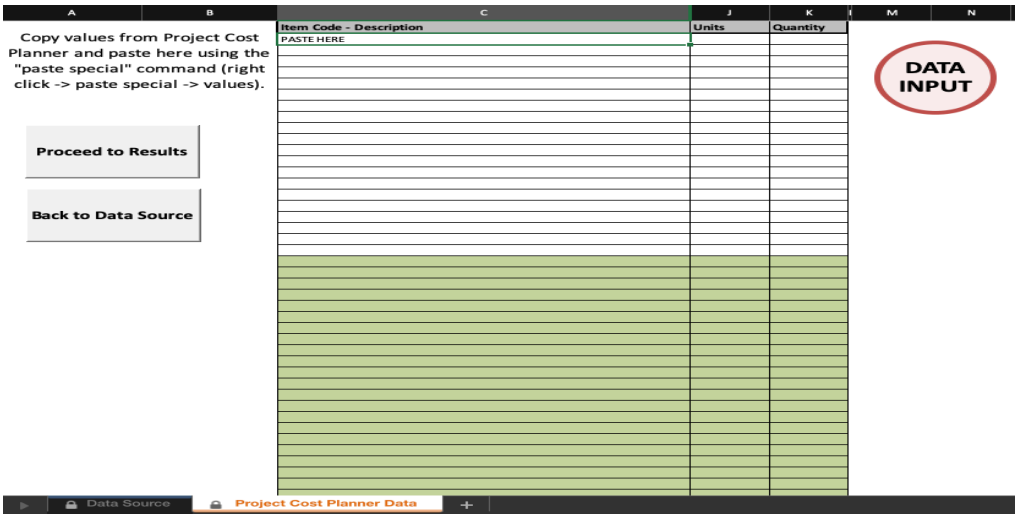


Figure 3.5 Conversion Tool Paste Area – PCP Tool

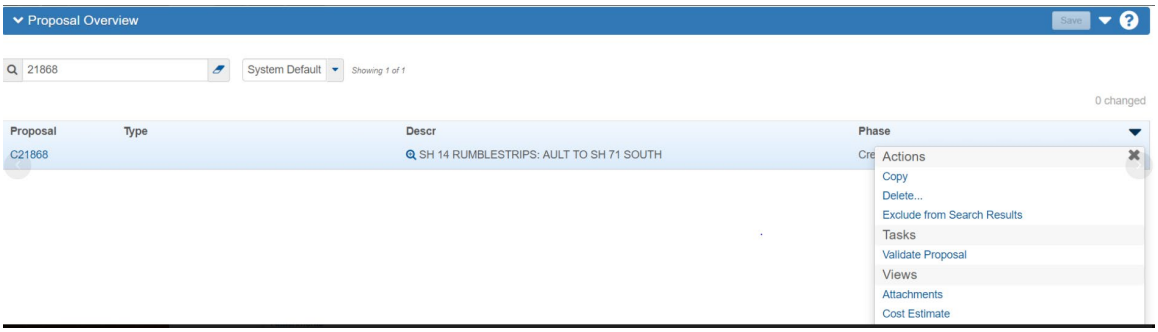
AASHTOWare

For the AASHTOWare data, copy the project's data from the excel spreadsheet (Figure 3.6) and paste using the paste special command into the designated cells (Figure 3.7).

Here are steps to print out (generate) a proposal price schedule (aka estimate) in AASHTOWare in excel. *You may need to contact EEMA for help in generating the report.*

- First go to Proposal Overview:

- Then use the action pull down arrow:



to select CDOT Proposal Price Schedule XLS:

	A	B	C	D
1	**PAYITEM**	**DESCR**	**QTY**	**UNIT**
2	614-80391	Rumble Strip (Grind)(Asphalt)	259683	LF
3	620-00020	Sanitary Facility	1	EACH
4	626-00000	Mobilization	1	L S
5	626-01104	Public Info Services (Tier IV)	1	L S
6	627-00008	Mod Epoxy Pvmt Mkg	1000	GAL
7	630-00000	Flagging	390	HOUR
8	630-00007	Traf Ctrl Inspection	4	DAY
9	630-00012	Traf Ctrl Mgmt	13	DAY
10	630-80341	Const Traf Sign (A)	16	EACH
11	630-80342	Const Traf Sign (B)	28	EACH
12	630-80355	Port Mesg Panel	2	EACH
13	630-80360	Drum Channel Dev	35	EACH
14	630-80520	Mobile Pvmt Mkg Zone	13	DAY

Figure 3.6: Sample of excel spreadsheet generated from AASHTOware.

Figure 3.7 Conversion Tool Paste Area - AASHTOWare

Running the Tool

Once the data is pasted and the user proceeds to results, the values obtained (Figure 3.8) have to be copied into the Contract Time Tool's respective fields (Figure 3.9) below. Once this is done, the user clicks "COMPUTE" and the Tool gives the estimated durations with the ±30% values.

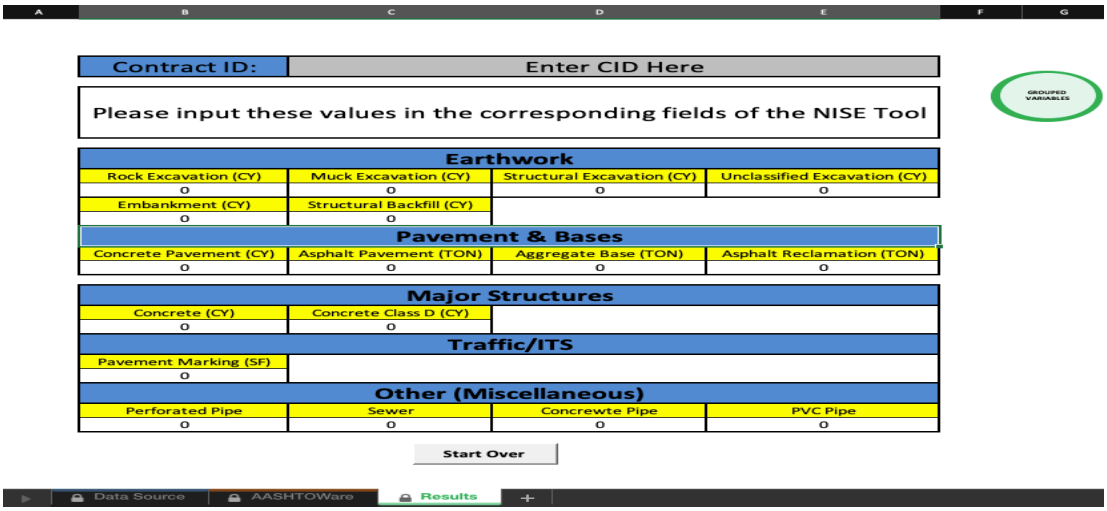


Figure 3.8. Grouped Values

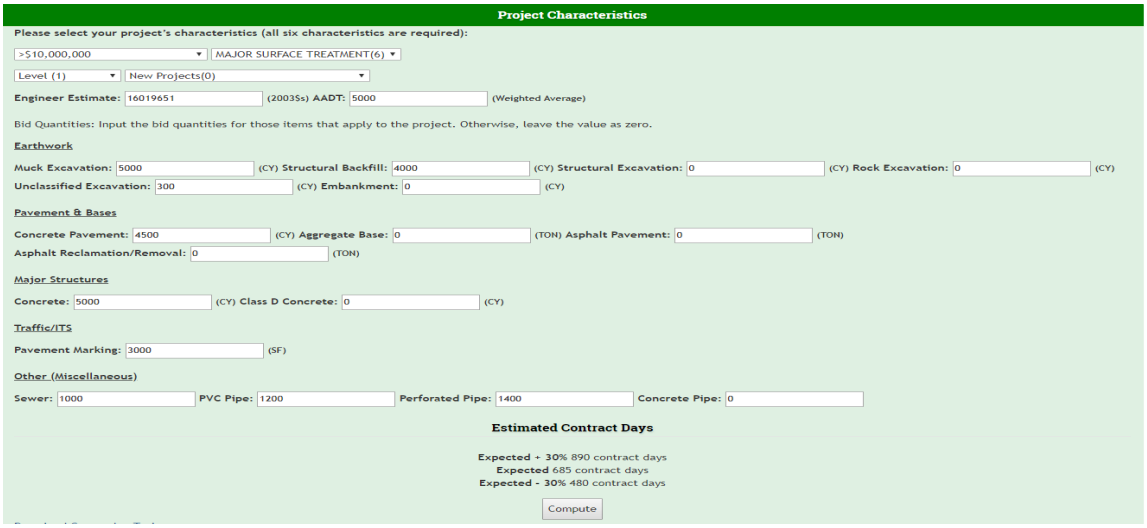


Figure 3.9: Contract Time Tool – Interface

Results and Interpretation

The results of the tool are an estimated working day duration and one value that is 30% above the working day estimate and another one that is 30% below the working day estimate. Since the average percent error of the tool is 25%, these values provide an interval in which the duration of a project should fall. These results, however, are intended to be used in the early stages of the project, so the Project Mangers' evaluation is required to determine whether the estimates are reasonable. This tool is not intended to be used to develop contract time or replace the Form 859 and Critical Path Method schedule.

There may also be other variables the Project Manager or Engineer should consider when determining contract time, such as procurement of items with a long lead time, special events, or

a seasonal project shutdown. There may be additional considerations like these that are not strictly related to actual construction time, and the tool will not automatically accommodate those as one would reflected in the Construction End Date. The user has to take all the above factors into account when estimating the Construction End Date or construction duration (in months). This tool is meant to help validate, or serve as a starting point for a more detailed Microsoft Project schedule as required in the form 859.

If it is necessary, converting a Working Day estimate to a Calendar Day estimate is done by multiplying the Estimated Contract Days by 1.4 to add weekends. This is based on the assumption that work time is five days per week. Once the weekends are added the next step is to account for the additional no-work periods such as holidays and a reasonable number of weather days.

$$\boldsymbol{Calendar\ Days} = (\textbf{Estimated Contract Days} \times \frac{\boldsymbol{7\ calendat\ days}}{\boldsymbol{5\ work\ days}}) + \boldsymbol{No\ work\ periods}$$

For example, the Contract Time Tool provides a result of 250 contract days, with an **estimated Construction start date of April 1, 2019.**

$$\boldsymbol{Calendar\ Days} = \mathbf{250}\ \boldsymbol{Days} \times \mathbf{1.4} = \mathbf{350\ days}$$

March 16, 2020 is 350 Calendar days after April 1, 2019 and at this point, the contract time does not account for no-work periods.

There are eight holidays between April 1, 2019 and March 16, 2020. Add 12 days to account for the holidays. This is based on this estimate of non-work days per holiday: Memorial Day (1.5), Independence Day (1.5), Labor Day (1.5), Thanksgiving (2.5), Christmas Day (1.5), New Year's Day (1.5), Martin Luther King Jr. (1) and President's Day (1). A half-day was added to some holidays per 108.08, because six of these holidays restrict work after noon on the day before the holiday.

Next assume seven days per month for weather. This is an estimate for this example and will vary based on project location and time of year. The project spans 12 months, therefore add 84 days for weather.

$$\boldsymbol{Calendar\ Days}\ (\boldsymbol{including\ no\ work\ periods}) = \mathbf{350}\ \boldsymbol{Days} + \mathbf{12}\ \boldsymbol{Day} + \mathbf{84}\ \boldsymbol{Days} = \mathbf{446}\ \boldsymbol{days}$$

In summary, the tool provided an expected Estimated Contract Days of **446 days** and based on weekends, holidays and other no-work.

As part of the implementation, the research team assisted CDOT personnel in installing the tool on the agency's servers. The general overview of the installation steps are included in Appendix A.

Chapter 4 – References

Baumers, M., Dickens, P., Tuck, C., and Hague, R. (2016). "The cost of additive manufacturing: machine productivity, economies of scale and technology-push." *Technological Forecasting and Social Change*, 102, 193–201.

Boussabaine, A. H., and Elhag, T. M. S. (1997). "A neurofuzzy model for predicting cost and duration of construction projects." *RICS Research (9 p.). The Royal Institution of Chartered Surveyors*.

Building Futures Council. (2006). "Measuring productivity and evaluating innovation in the US construction industry." *Arlington, VA*, 1–13.

Clarke, K. A. (2005). "The phantom menace: Omitted variable bias in econometric research." *Conflict management and peace science*, 22(4), 341–352.

Conover, W. J. (1980). *Practical nonparametric statistics*. Wiley, New York.

Craney, T. A., and Surles, J. G. (2002). "Model-dependent variance inflation factor cutoff values." *Quality Engineering*, 14(3), 391–403.

Duin, R. P. (1995). "Small sample size generalization." *Proceedings of the Scandinavian Conference on Image Analysis*, PROCEEDINGS PUBLISHED BY VARIOUS PUBLISHERS, 957–964.

Falk, J. E., and Horowitz, J. L. (1972). "Critical path problems with concave cost-time curves." *Management Science*, 19(4-part-1), 446–455.

FHWA. (2002). "FHWA Guide for Construction Contract Time Determination Procedures - Contract Administration - Construction - Federal Highway Administration." <https://www.fhwa.dot.gov/construction/contracts/t508015.cfm> (Sep. 21, 2017).

Finnie, G. R., Wittig, G. E., and Desharnais, J.-M. (1997). "A comparison of software effort estimation techniques: using function points with neural networks, case-based reasoning and regression models." *Journal of systems and software*, 39(3), 281–289.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

Frandina, S., Gori, M., Lippi, M., Maggini, M., and Melacci, S. (2013). "Variational foundations of online backpropagation." *International Conference on Artificial Neural Networks*, Springer, 82–89.

Fulkerson, D. R. (1961). "A network flow computation for project cost curves." *Management science*, 7(2), 167–178.

Goodrum, P. M., Haas, C. T., and Glover, R. W. (2002). "The divergence in aggregate and activity estimates of US construction productivity." *Construction Management & Economics*, 20(5), 415–423.

Günaydın, M. H., and Doğan, Z. S. (2004). "A neural network approach for early cost estimation of structural systems of buildings." *International Journal of Project Management*, 22(7), 595–602.

Hegazy, T., and Ayed, A. (1998). "Neural Network Model for Parametric Cost Estimation of Highway Projects." *Journal of Construction Engineering and Management*, 124(3), 210–218.

Herbsman, Z. J., and Ellis, R. (1995). *Determination of contract time for highway construction projects*.

"How can I interpret log transformed variables in terms of percent change in linear regression? | SAS FAQ." (n.d.). <https://stats.idre.ucla.edu/sas/faq/how-can-i-interpret-log-transformed-variables-in-terms-of-percent-change-in-linear-regression/> (Apr. 11, 2019).

Irfan, M., Khurshid, M. B., Anastasopoulos, P., Labi, S., and Moavenzadeh, F. (2011). "Planning-stage estimation of highway project duration on the basis of anticipated project cost, project type, and contract type." *International Journal of Project Management*, 29(1), 78–92.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Jiang, Y., and Wu, H. (2004). "Determination of INDOT highway construction production rates and estimation of contract times."

Jiang, Y., and Wu, H. (2007). "A method for highway agency to estimate highway construction durations and set contract times." *International Journal of Construction Education and Research*, 3(3), 199–216.

Kabacoff, R. (2015). *R in Action: Data Analysis and Graphics with R*. Manning Publications Co., Greenwich, CT, USA.

Khallaf, R., Yoon, S., Hastak, M., and Nantung, T. (2016). "Simplified Construction Scheduling for Field Personnel."

Kim, G.-H., An, S.-H., and Kang, K.-I. (2004). "Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning." *Building and Environment*, 39(10), 1235–1242.

Klerfors, D., and Huston, T. L. (1998). "Artificial neural networks." *St. Louis University, St. Louis, Mo*.

McCrary, S. W., Corley, M. R., Leslie, D. A., and Aparajithan, S. (1995). *Evaluation of contract time estimation and contracting procedures for Louisiana Department of 649 Transportation and Development construction projects*.

Mubarak, S. A. (2015). *Construction Project Scheduling and Control (3rd Edition)*. John Wiley & Sons, Incorporated, Somerset, UNITED STATES.

"NHCCI / Description - Policy | Federal Highway Administration." (n.d.). <https://www.fhwa.dot.gov/policy/otps/nhcci/desc.cfm> (Oct. 29, 2018).

Petroutsatou, K., Georgopoulos, E., Lambropoulos, S., and Pantouvakis, J. P. (2011). "Early cost estimating of road tunnel construction using neural networks." *Journal of construction engineering and management*, 138(6), 679–687.

Pewdum, W., Rujirayanyong, T., and Sooksatra, V. (2009). "Forecasting final budget and duration of highway construction projects." *Engineering, Construction and Architectural Management*, 16(6), 544–557.

Piotrowski, A. P., and Napiorkowski, J. J. (2013). "A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling." *Journal of Hydrology*, 476, 97–111.

Shahandashti, S. M., and Ashuri, B. (2015). "Highway construction cost forecasting using vector error correction models." *Journal of Management in Engineering*, 32(2), 04015040.

Sheskin, D. J. (2003). *Handbook of parametric and nonparametric statistical procedures*. crc Press.

Smith, A. E., and Mason, A. K. (1997). "Cost Estimation Predictive Modeling: Regression Versus Neural Network." *The Engineering Economist*, 42(2), 137–161.

Sveikauskas, L., Rowe, S., Mildenberger, J., Price, J., and Young, A. (2016). "Productivity growth in construction." *Journal of Construction Engineering and Management*, 142(10), 04016045.

Taylor, T. R., Goodrum, P. M., Brockman, M., Bishop, B., Shan, Y., Sturgill, R. E., and Hout, K. (2013). "Updating the Kentucky Contract Time Determination System."

Taylor, T. R., Sturgill Jr, R. E., and Li, Y. (2017). *Practices for Establishing Contract Completion Dates for Highway Projects*.

Teicholz, P. (2013). "Labor-productivity declines in the construction industry: Causes and remedies (another look)." *AECbytes Viewpoint*, 67.

US Census Bureau Construction Expenditures. (2018). "US Census Bureau Construction Spending Survey." <https://www.census.gov/construction/c30/historical_data.html> (Sep. 20, 2018).

Vereen, S. C., Rasdorf, W., and Hummer, J. E. (2016). "Development and comparative analysis of construction industry labor productivity metrics." *Journal of Construction Engineering and Management*, 142(7), 04016020.

Werkmeister, R., Luscher, B., and Hancher, D. (2000). "Kentucky contract time determination system." *Transportation Research Record: Journal of the Transportation Research Board*, (1712), 185–195.

Wilmot, C. G., and Mei, B. (2005). "Neural Network Modeling of Highway Construction Costs." *Journal of Construction Engineering and Management*, 131(7), 765–771.

Woldesenbet, A. K. (2010). "Estmation Models for Production Rates of Highway Construction Activites." Oklahoma State University.

Zhai, D., Shan, Y., Sturgill, R. E., Taylor, T. R. B., and Goodrum, P. M. (2016). "Using Parametric Modeling to Estimate Highway Construction Contract Time." *Transportation Research Record: Journal of the Transportation Research Board*, 2573, 1–9.

Appendix A – Steps for Tool Server Installation

Part I: Ensure that app is able to run locally using flask.

1. Run as administrator.

2. Ensure that python 3.6.3 is installed, along with all the following libraries: numpy, pandas, keras, sklearn, flask and wtforms and that the version of Windows is Windows Server 2016 or earlier (Windows 10 as the earliest).
    a. Make sure to add Python to Path.
    b. Make Theano the backend framework for keras instead of TensorFlow. Go to the file "~/.keras/keras.json" and replace "tensorflow" with "theano".

3. Download miniconda for windows: https://docs.conda.io/en/latest/miniconda.html. Use the miniconda command line to run all the upcoming commands in this tutorial. (This allows Windows to use Theano which makes the epoch processing faster, and the command prompt for "conda" should appear by typing "anaconda" on search bar)
    a. Once you download miniconda, ensure to run the following commands inside its designated terminal:
        i. conda install m2w64-toolchain
        ii. conda install -c anaconda libpython

4. Download the folder "NiseCdot". This contains all of the files you will need for this application.

    a. Run the "newtrain.py" file using the "python newtrain.py" command on your terminal or the terminal of any editor you may be using. This will load the model and create a sample prediction of a dataset used for proofing and debugging.
    b. Run the "middle_man.py" file using the "python3 middle_man.py" command to ensure all is compiling correctly.

c. Finally, to run the application locally using flask run the following two commands: “export FLASK_APP=nisecdot.py”, “flask run”.
    i. It should default to the following port: http://127.0.0.1:5000/

5. In case the data changes and you need to upload a different excel file, as long as there isn’t any structural change to the columns or file (there is still the same number of columns and each input stays on its named column), nothing is necessary besides replacing the excel file.
    a. In case you do change or add a column, you will need to mark the range of columns you are going to use as inputs through “dataset.iloc”. The “iloc” command is what determines what data is getting retrieved from the file you inputted. In the “newtrain.py” file it is specified as columns 1 – 45 (“dataset.iloc[:, 1:45]”) currently. You can see this on line 17. You will also have to change the number of columns you have in line 18, currently 0 – 45. The same goes for line 53 in the same file because you need to specify the range for the testing data.
    b. You will also have to change line 22 on “nisecdot.py” and switch the two instances of the number “44” to one minus whatever number of inputs you may have. Currently the count is “45”, therefore the correct number is “44”.
    c. The “methods.py” file also needs to be edited. You need to add or remove whatever input you changed. If you added one, the format for adding a new one is “NAME = FloatFIeld(validators = [validators.InputRequired])” because this is what provides all of the different inputs that will be brought in though the website on order to make a prediction.
    d. The “middle_man.py” file would also have to be changed in a way similar to the “newtrain.py” file, on line 16 the “dataset.iloc” needs to be adjusted accordingly.
    e. The “nisecdot.py file would also have to be updated because adding or removing an input will change the list that is passed into the “compute” function called from “middle_man.py”. Similar to all the inputs from the request form that are in lines 19 – 92 you will have to add the added input with the same format (“nameUsed = form[‘Name From “methods.py” File’].data”). After adding this you will also

have to reference and add it to the “totalList” field in line 94 according to the order of the columns in the excel file that contains the training data. **Please note that the order of the values inside the list:“totalList” are in the exact same order as they appear in the columns from the excel file:“Final Data CO.xslx”. This order must be maintained in order for the prediction to work properly.**

f. Lastly you will have to edit “main_page2.html” and have to delete or add the new input so it is visible in the user interface. The format to add the input is:

<label>
<action=”” method=”POST”>
Name of input
<input id=”input ID” name=”Name of Input how you will reference it in other files” type=”number” class=”small” value=0>
<small>Abbreviation for quantity or Acronym</small>
</label>

More examples are between lines 213 and 436.

An addition to the “onsubmit” field on line 212 also has to be added with “saveValue(InputName)” next to all the other for the value to be saved. And finally an addition between lines 468 and 490 with the format: “document.getElementById(“InputName”).value = getSavedValue(“InputName”).

If you removed an input just make sure to remove all references to that input in this file and all other files.

6. Finally, once the application is successfully running locally and it will be deployed to a server, make sure to change the base url in “main_page2.html” because it is currently pointing to https://www.cdot.gov/programs-projects/landing-page. This is because the html structure of that webpage was used to render nisecdot. The last edit after this would be to change the “href” in line 463, currently http://127.0.0.1:5000/data, to whatever

location the downloadable file “NISE Grouping and Conversion Tool.xlsm” is moved to or whatever base url gets used for the web application.

Part II: Deploy the app using IIS.

1. Run every app as Administrator.
2. Run “pip install wfastcgi”.
3. Run “wfastcgi-enable”.
    a. It will produce output like this:
        i. “… Applied configutation changes to section “system.webserver/fastcgi” for “MACHINE/WEBROOT/APPHOST” at configuration commit “c:…\python36\python.exe|c:…\python36\lib\site-packages\wfastcgi.py” can now be used as a FastCGI script processor.
4. Go to “Control Panel”, then pick “Programs”, then “Turn Windows features on or off”. Check “Internet Information Services” and under that open “Application Development Features”. Check the “CGI” box, then press okay.
5. Run IIS
6. Select “Sites” under “Connections”
7. Click “Add Website” on right of console under “Actions”.
8. Fill in necessary site info: Site Name, Directory Containing the website content, IP address and port (I used 5000).
9. The physical path you specified should have an exact copy of the NiseCdot folder.
    a. Please make sure to change the “PYTHONPATH” inside “web.config” file to what path the “nisecdot” folder is.
    b. “ScriptProcessor” inside “web.config” needs to also be changed to wherever the “python.exe” and “wfastcgi.py”. These are inside the “Pyhton36” folder that you downloaded with python 3.6.3. They are also found when you do “wfastcgi-enable” from instruction 3.a.i.
10. Make sure that “nisecdot” and that “Python36” have permissions of “read” and “write” for IIS_IUSRS and IUSR.