Resilient Structures and Sustainable Construction
Edited by Pellicer, E., Adam, J. M., Yepes, V., Singh, A., and Yazdani, S.
Copyright © 2017 ISEC Press
ISBN: 978-0-9960437-4-8

USING STATISTICAL MODELS BASED ON HISTORICAL PROJECT DATA TO ESTIMATE DURATIONS FOR TRANSPORTATION PROJECTS

GUILLERMO NEVETT, DOUGLAS ALLEMAN, and PAUL GOODRUM

Dept of Civil, Environmental and Architectural Engineering, University of Colorado Boulder, Boulder, USA

Estimating the duration of highway transportation projects accurately has always been a challenge for State Transportation Agencies (STAs) in the United States (US) due to the differences in projects' location, scope, nature, size, and goals. Inaccurate estimates of project durations by STAs can lead to a lack of commitment from contractors, not needing to invest their maximum effort to fulfill a project's scope. STAs have used a variety of methods to estimate and set contract time, mainly relying on historical production rates in tandem with activity precedent logic programs. The Kentucky Transportation Cabinet has recently attempted to develop a statistical model which estimates contract times by using parametric modeling and historic contracting times with success. This paper attempts to build upon their findings, regressing historical data from Montana Department of Transportation projects using bid tabulations, engineers' estimates, location, and type of project, among other factors as inputs. The model presented is still under development and is expected to increase in accuracy as it reaches its final state.

Keywords: Highway, Regression analysis, Parametric modeling, Duration estimation.

1 INTRODUCTION

Project managers and project planners are continually expected to provide estimates of schedule length for upcoming projects. Anyone who has tried to construct and justify project duration estimates knows that it is almost an art rather than a science to achieve reasonable levels of accuracy and consistency. Project duration is influenced by factors such as change orders, worker unrest, material shortages, delayed payments, drawing changes, poor supervision, etc. (Kraiem 1987, Majid and McCaffer 1998, Kalibe *et al.* 2009). Yet it is in the planning stages that agencies are developing scheduling estimates, prior to any of these problems occurring. This, along with the differences in projects' location, scope, nature, size, and goals, has caused schedule estimating to be a big challenge for State Transportation Agencies (STAs). Accurate project duration estimation is of extreme importance as inaccurate project duration estimates can negatively impact contractor commitment, bid proposals pricing, public relations, safety implementation, project costs, and even life-cycle cost analyses (Williams 2006, Ifran *et al.* 2011).

To fulfill the need for accurate time estimation, this research's objective it to present a new method of time estimation based on historical project data, improving the accuracy of estimates on duration of transportation road projects. The following paper describes the process followed by the authors to build a stepwise regression preliminary time estimation model based on bid quantities from projects provided by Montana's Department of Transportation (DOT) (MDT). The data used consisted on bid tabulations, engineers' estimates, and type of project, among other factors. The process produced very encouraging results, though the authors feel that the accuracy of the model could be improved with additional data from MDT and perhaps merging the MDT with data from other state transportation agencies with similar scope of work and terrain (i.e. Colorado DOT), which could be supported through the existing Transportation Construction Management Pooled Fund Study, which includes MDT.

The current STA state-of-practice of time estimating modelling is developed from federal policies and guidelines. The Federal Highway Administration (FHWA) Guide for Construction Contract Time Determination Procedures to specify the essential steps in determining contract time include:

1) Establishing production rates.

2) Adjusting production rates to applicable project.

3) Understanding potential impacting factors and characteristics.

4) Calculating project duration using production rates with a progress schedule (FHWA 2002).

Taylor *et al.* (2012) performed a review of 29 states' finding STA's following FHWA estimating guidelines to a varying degree. The STA's estimating procedures were found to have the following characteristics: 28% were based solely on production rates, 17% used a predetermined job logic without production rates, and 49% had an integrated scheduling system combining job-logic templates with production rates. Taylor *et al.* (2012) tested two of these models used commonly by STAs on completed Kentucky highway projects finding an absolute percent error of 55% at best and 167% at worse. Furthermore, a survey of STA users of these two systems found that the majority had little confidence in the accuracy of the systems (Taylor *et al.* 2012). As such, existing STA project time estimating techniques may not be reliable for estimating contract time and require improvement. One such avenue of potential improvement is the use of parametric modelling in conjunction with historical project data.

Scheduling models have evolved throughout literature from a simple linear relationship between the cost of a project and the duration (Fulkerson 1961) to non-linear relationships (Falk and Horowitz 1972, Foldes and Soumis 1993), with more recent literature focusing on discrete formulations (Skutella 1998, Zheng *et al.* 2004) and discontinuous time-cost functions (Moussourakis and Haksever 2004, Yang 2005). Though previous models based on these theories have found a range of success, parametric modelling has been proposed as a more apt model for STA project duration estimation.

Zhai *et al.* (2016) justify parametric modelling for STA project duration estimation for the following reasons:

1) Use of correlation (construction durations are highly correlated with bid quantities).

2) Reliance on data similarities (highway construction projects are linear in nature, highly repetitive, and construction means and methods are relatively similar across the US).

3) Ability to be developed based on historical databases.

Existing literature corroborates Zhai *et al.*'s justifications with several successful uses to estimate highway construction duration (Boussabaine and Elhag 1997, Jiang and Wu 2007, Liu *et al.* 2011, Irfan *et al.* 2011). One of the most recent and greatest successes has been the use of multivariate regression analysis to estimate Kentucky highway transportation project durations.

Taylor *et al.* (2012) proposed a multi-variate regression model based on historic project schedule performance to improve the accuracy of the existing Kentucky Transportation Cabinet's system. Using 66 completed Kentucky projects as inputs, the develop model was found to increase the estimating accuracy by 61% to 87% in comparison to two existing estimating models used by STAs with a goodness of fit ranging from R^2 0.581 to 0.967. Zhai *et al.* (2016) used 2,503 Kentucky highway projects completed from 2002 to 2011 to further improve the model developed by Taylor *et al.* (2012). Their proposed multi-variate regression model increased the number of inputs from eight to 23. Using 20% of the 2,503 projects for validation, they developed a model to estimate project durations with a goodness of fit with a R^2 of 0.891.

Based on these successes, the following research chose to use a multivariate regression analysis to develop a model to estimate project duration for MDT projects. This paper contributes to the existing body of knowledge by testing a multivariate regression analysis's applicability to states other than Kentucky.

2 DATA COLLECTION AND METHOD OF ANALYSIS

This research uses a multivariate linear regression analysis to develop the MDT models. The linear regression model can be described as follows, Eq. (1) :

$$\Upsilon = \beta_0 + \beta_1 X_1 + \cdots + \beta_1 X_n + \epsilon \tag{1}$$

$$\epsilon \sim NID(0, {\sigma^2}_\epsilon)$$

$$\beta_i \ (i = 1 \ to \ n)$$

where, Υ = dependent variable, (Charge Days). β = regression coefficient (see first model). X = observed value of independent variables. ε = random error term or noise (accounts for all other factors that affect Y than X). NID = normally and independently distributed.

The regression model assumes the following are true about the error term (ε):

1) population mean of ε is zero.

2) εI have the same variance ${\sigma^2}_\epsilon$ for all values of X.

3) ε is normally and independently distributed.

Concerning the data used for this paper, it was provided by MDT and consisted on 43 bid items of 259 sample projects over a period of seven years (2009-2015). The data used for each project includes: critical project dates (award, notice to proceed, completion, etc.), charged days, EE, and bid item quantities. The dataset had 20 natural project type categories, as provided by MDT and shown in Table 1.

Stepwise regression was used to optimize this model as stepwise regression can minimize impacts of multi-collinearity, considers more relevant models during the process of obtaining optimal models, and limits the number of independent variables to those that have the highest statistical impact (Zhai 2016).

Since not all the projects were awarded or built in the same year, the first step in analyzing this data was to account for inflation. To do so, all the EE were transformed to 2015 USD using the National Highway Construction Cost Index (NHCCI) (FHWA 2015).

Given that the ratio of factors (bid items) to number of cases (projects) is relatively high and the low frequency of most of the bid items, the team created new factors by grouping similar bid items. The criteria used to group such items were characteristics (e.g., two types of asphalt pavement with the same thickness) and within these characteristics, similarities in daily rates per

RSMeans®. Grouping these items reduced the factor count from 43 to 20, as seen below in Table 2. This factor reduction increased the strengths of our findings and analysis by increasing the adjusted R2. Similarly to the R^2 value, the adjusted R^2 is the amount of variance in the dependent variable explained by the predictors, but contrary to the R^2 value, the adjusted R^2 takes into consideration whether the ratio of predictors to sample size is too large (Kabacoff 2015).

Table 1. Project types.

Project Type	Frequency	Project Type	Frequency
Overlays	87	Signals	3
Reconstruction, Grading	60	Miscellaneous	2
Bridge construction, rehab, and removal	28	Portland cement/concrete pavement	2
Safety	27	Bike and pedestrian	1
Slides or slope stabilization	14	Crack seal	1
Seal and cover	13	Fencing	1
Rehab (minor grade and overlay)	6	Micro-surfacing	1
Guardrail	4	Scour Projects	1
Drainage	3	Sidewalk	1
Environmental and Wetland	3	Signing	1

Table 2. Grouped bid items frequency in rank order.

Bid Item	Frequency	Project Type	Frequency
Crushed Aggregate Course	177	Commercial Asphalt Mix 3/4	20
Plant Mix 3/4	96	Plant Mix 9 mm	19
Excavation (unclassified)	92	Commercial Mix 3/8	6
General Asphalt Commercial Mix	82	Plant Mix 1/2	4
Special Borrow Neat Line	67	Plant Mix 3/8	4
Embankment in Place	57	Commercial Mix mm	3
Steel	41	Concrete Class Structure	2
Excavation Borrow	32	Concrete Class Deck	1
Concrete Class DD Bridge	31	Concrete Class SD Repair	1
Concrete General	31		

3 FINDINGS: MODEL DEVELOPMENT

As stated, the statistical method used in this model is stepwise regression. In such regression, the software (SPSS Statistics) runs an iterative process by removing variables that are not significant and it stops once it finds the best model keeping the factors included in that best model. Since the coefficient of determination (R2) does not guarantee predictive accuracy of the model, only 80% of the data was analyzed to create the models to be able to validate it with the remaining 20%. The validation consists on comparing the predicted value of the dependent variable (Charge Days) with following formula, Eq. (2) (Zhai *et al.* 2016):

$$Percent\ Error = \frac{|Predicted\ Value - Observed\ Value|}{Observed\ Value} \times 100 \tag{2}$$

When calculating Percent Errors, to account for extreme outliers, the median is the central tendency metric used to compare different model iterations. The model's iterations can be seen and explained below. There are several model characteristic identifiers that are used.

The first model developed by the team was a general model, all the projects in the sample were used (80% model and 20%) validation, producing a model with the following characteristics, Eq. (3):

$$Y = 44.532 + 9.253E - 6 * X_1 + 0.008 * X_2 + 0.001X_3 + 5.421E - 5 * X_5 + 0.002 * X_5 + \varepsilon \quad (3)$$

Goodness of fit F = 121.354, significance = 0.000 Adjusted R^2 = 0.746, Mean Percent Error = 44.59%, Median Percent Error = 29.54%.

After running the first model and analyzing the descriptive statistics of the most important predictor, EE, the data was split into three subgroups. The first group (second model) was for projects with budget $1,000,000 and under; the second (third model) one for projects between $1,000,001 and $3,000,000; and the third (fourth model) for projects $3,000,001 and above.

A summary of these analyses is shown in table 3:

Table 3. Summary of models' analyses.

Model	Goodness of Fit (F)	Significance	Adjusted R^2	Mean PE	Median PE	Sample Size
Second model	8.344	0.000	0.237	58.75%	24.75%	91
Third model	14.592	0.000	0.48	22.42%	19.35%	76
Fourth model	26.641	0.000	0.649	42.28%	19.59%	93

As it can be seen in the results of all the models, the first model iteration or general model has a larger Adjusted R^2, but the mean and the median percent error are higher than those values in all the models (except for the mean error for the second model). Another observation worth mentioning is that as the projects' size increase, R^2 also increases while the median percent error decreases, so the model becomes better at predicting the dependent variable, charge days. Lastly, it's also worth mentioning that the significant factors were different for each model even though all factors were included in each regression. This means that different project sizes require different models to use as estimating tools and shouldn't be treated as equals.

4 DISCUSSION AND CONCLUSION

Previous literature has found that current DOT contract estimating systems' accuracy is likely not suitable for efficient STA construction practices (Taylor *et al.* 2012). Inefficient estimating can negatively impact contractor commitment, bid proposals pricing, public relations, safety implementation, project costs, and even life-cycle cost analyses (Williams 2006, Ifran *et al.* 2011). The duration estimating model presented exhibits a goodness of fit with an R^2 of 0.647 which is within the range of existing models presented for Kentucky highway projects (Taylor *et al.* 2012). The model developed is an improvement to models currently in use with a percent error of 19.59%. This depicts an increase in accuracy and would be an advantageous tool for MDT. This model shows promising results on how statistical models can be used by DOTs to easily estimate durations of projects using bid quantities, during the planning phase or after a change order is experienced.

Though this model shows promise, several steps are required prior to successful dissemination to all STAs. Although the model is accurate, more projects and data are available to strengthen the model. The team is already working towards developing more accurate models by including more factors and different modeling techniques, including K-Fold cross validation and Artificial Neural Networks. Future research includes using the presented approach to DOTs throughout the US, determining its universal applicability. Other research includes converting the developed models to a user-friendly Microsoft Excel-based estimating tool and testing the accuracy and applicability of said tool at the STA project level.

Future research goals include testing the model's applicability to all STAs and developing a user-friendly model to be used by STA schedulers. For this to occur, the research team needs to both strengthen the existing model and test it on multiple DOTs. The team is already working towards developing more accurate models by including more factors, such as other bid items, delivery methods, projects' locations, among other factors.

References

Boussabaine, A. H. and Elhag, T. M. S., *A Neurofuzzy Model for Predicting Cost and Duration of Construction Projects*, RICS Research, Royal Institution of Charted Surveyors, 1997, 9 pp.

FHWA, Federal Highway Administration, Guide for Construction Contract Time Determination Procedures, 2002. Retrieved from http://www.virginiadot.org/business/resources/const/CTDR_Guidelines_FHWA.pdf on 19 Feb 2017.

FHWA, Federal Highway Administration, Construction Cost Trends for Highways | NHCCI - Policy | Federal Highway Administration, 2015. Retrieved from https://www.fhwa.dot.gov/policyinformation/nhcci/pt1.cfm on Jan. 10, 2017.

Falk, J. and Horowitz J. L., Critical Path Problems with Concave Cost-Time Curve, *Management Science*, 1972.

Foldes, S. and Soumis, F., Pert and Crashing Revisited: Mathematical Generalizations, *Eur. J. Oper. Res.*, 1993.

Fulkerson, *A Network Flow Computation for Project Cost Curves*, Manage. Sci., 7 pp. 167-178, 1961.

Irfan, M., Khurshid, M. B., Anastasopoulos, P., Labi, S., and Moavenzadeh, F., *Planning-stage Estimation of Highway Project Duration on the Basis of Anticipated Project Cost, Project Type, and Contract Type*. J. Constr. Manage., 21 (1), pp. 78-92, 2011.

Jiang, Y. and Wu, H., A Method for Highway Agency to Estimate Highway Construction Duration and Set Contract Times, *International Journal of Construction Education and Research*, Vol. 3, No. 3, pp. 199–216., 2007.

Kraiem, Concurrent Delays in Construction Projects, *J. Constr. Manage.*, 113 (4), pp. 591–601, 1987.

Majid, M., McCaffer, R., Factors of Non-Exusable Delays that Influence Contractor's Performance, *J. Manage. Eng*, pp.42-49, 1998.

Liu, W., *Duration Estimation Method for Highway Construction Work*, Proc., 2011 International Conference on Management and Service Science (MASS), Wuhan, China, Aug. 12–14, 2011.

Moussourakis, J. and Haksever, C., Flexible Model for Time–Cost Trade-off Problem, *J. Constr. Manage.*, 130 (3), pp. 307–314, 2004.

Skutella, M., Approximation Algorithms for the Discrete Time–Cost Trade-off Problem, *Math. Oper. Res.*, 23 (4), pp. 909–929, 1998.

Taylor, T., Brockman, M., Zhai, D., Goodrum, P., and Sturgill, R., Accuracy of Selected Tools for Estimating Contract Time on Highway Construction Projects, *Construction Research Congress*, West Lafayette, Ind., May 21–23, 2012.

Williams, R., *The Framework of a Multi-Level Database of Highway Construction Performance Times*, MS thesis. Virginia Polytechnic Institute and State University, Blacksburg, 2006. Retrieved from https://vtechworks.lib.vt.edu/bitstream/handle/10919/32148/RCW_Thesis.pdf?sequence=1&isAllowed=y on January 14, 2017.

Yang I. T., *Chance-Constrained Time–Cost Tradeoff Analysis Considering Funding Variability* J. Constr. Manage., 131 (9), pp. 1002–1012, 2005.

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

Zheng D.X.M., Ng S.T., and Kumaraswamy M. M., Applying a Genetic Algorithm-Based Multiobjective Approach for Time–Cost Optimization, *J. Constr. Manage.*, 130 (2), pp. 168–176, 2004.