INTEGRATING MODEL TREE AND MODIFIED STEPWISE REGRESSION IN CONCRETE SLUMP PREDICTION AND STEEL FABRICATION ESTIMATING

Journal:	<i>Canadian Journal of Civil Engineering</i>
Manuscript ID	cjce-2020-0753.R2
Manuscript Type:	Article
Date Submitted by the Author:	09-Jun-2021
Complete List of Authors:	Mohsenijam, Arash; University of Alberta, Civil and Environmental Engineering Lu, Ming; University of Alberta, Naumets, Serhii; University of Alberta, Civil and Environmental Engineering
Keyword:	Multiple Linear Regression, Model Tree, Modified Stepwise Regression, Steel Fabrication, Concrete Slump
Is the invited manuscript for consideration in a Special Issue? :	Not applicable (regular submission)

SCHOLARONE™
Manuscripts

1 INTEGRATING MODEL TREE AND MODIFIED STEPWISE
2 REGRESSION IN CONCRETE SLUMP PREDICTION AND STEEL
3 FABRICATION ESTIMATING

4
5 Arash Mohsenijam, Ph.D.¹

6 Ming Lu, Ph.D.²

7 Serhii Naumets³

8
9
10 ¹ Estimator and Process Designer, Supreme Group, 28169 96 Ave., Acheson, AB T7X 6J7,
11 Canada. E-mail: arash.mohsenijam@supremegroup.com.

12 ² Professor, Construction Engineering and Management, Department of Civil & Environmental
13 Engineering, University of Alberta. E-mail: mlu6@ualberta.ca (corresponding author).

14 ³ Ph.D. Candidate and Graduate Research Assistant, Department of Civil and Environmental
15 Engineering, University of Alberta, 116 St & 85 Ave, Edmonton, AB T6G 2R3, Canada. E-mail:
16 naumets@ualberta.ca.

17
18
19 Word count: 7022 + 5*250 (figures) = 8272 words

ABSTRACT

The model tree algorithm of M5 is integrated with the multiple linear regression technique called modified stepwise regression (MSR), resulting in a new methodology for modeling complex civil engineering problems. We purposefully chose artificial neural networks (ANN) for comparison against the proposed “M5+MSR” because they fall at the two ends of the model interpretability spectrum in machine learning. In two application cases (Case 1: concrete workability and Case 2: steel fabrication estimating), the proposed “M5+MSR” gave rise to explainable regression tree models featuring substantially reduced complexities against ANN and model prediction errors comparable to ANN. In both cases, the resulting “M5+MSR” models consistently outperformed ANN in terms of model overfitting metrics by 19% in Case 1 and by 21% in Case 2, thus boasting better learning performances. The proposed new methodology will potentially find applications in tackling a wide range of complicated engineering problems that entail fitting prediction models based on laboratory or field data.

Keywords:

Multiple Linear Regression; Model Tree; Modified Stepwise Regression; Concrete Slump; Steel Fabrication; Estimating.

INTRODUCTION

Machine learning in civil engineering applications has been using artificial neural network (ANN) algorithms to calibrate a prediction model to a training dataset that contains input factors and output responses measured from the same problem domain. For instance, Al-Barqawi and Zayed (2008) combined ANN with the Analytic Hierarchy Process model to evaluate the sustainability of water mains; the weights of factors and associated subfactors were determined so as to identify determinant factors for water main deterioration. El-Abbasy et al. (2014) used ANN to predict the condition of offshore oil and gas pipelines; deterioration curves were derived and the input variables with more significant impact on predicting pipeline service condition were highlighted. In a recent study by Awolusi et al. (2019) ANN algorithms were used to model the properties of steel fibre reinforced concrete. Five different ANN training algorithms were evaluated, with the Genetic Algorithms found to deliver the highest predicting performance. Despite significant research progress, ANN applications in civil engineering practice are scarce because ANN models generally present themselves to practitioners as “black box” which features implicit thought process and remains too complicated to implement [note the term “black box” was first coined by Ashby (1956) in cybernetics].

Alternatively, the model tree (MT) algorithm is analogous to the merge of decision tree and multiple linear regression (MLR). An application framework was proposed by Mohsenijam and Lu (2019) for developing an MLR model in its simplest form for labour-hours prediction based on project design features. As immediate follow-up research, we integrate MT as nonlinear classifiers prior to applying MLR regressions in order to model complex relationships in the problem. By the new methodology, variables that play a significant part in classification and prediction are automatically selected with their threshold values analytically set. It is worth mentioning MT is as

simple and transparent as linear regression, while also sufficient to tackle problems of real-world size and complexity. In the civil engineering field, numerous machine learning algorithms (such as Fuzzy Logic, Naive Bayes, Support Vector Machine, or other regression and classification trees like Random Forest) have been researched and applied. In the present research, we purposefully chose ANN for comparison against MT because they fall at the two ends of the model interpretability spectrum in machine learning, namely, “*Complicated and Black-Boxish*” vs. “*Simple and Transparent*”. In comparison with commonly applied modeling methods (such as ANN), the proposed MT-based methodology is found to be effective to identify the “sweet spot” in terms of achieving goodness of fit on training data (model sufficiency), delivering high prediction accuracy on testing data (model accuracy), while also considerably reducing model complexity.

In the following sections, first, the relevant literature on Model Trees is reviewed. Then, the research methodology and objectives are described. The application of the proposed methodology is first demonstrated to model workability of concrete based on a concrete slump dataset from the University of California Irvine (UCI) machine learning repository, which is commonly used to cross-validate machine learning algorithms. Thereafter, a practical case of steel fabrication productivity modeling is described before conclusions are drawn for this research.

LITERATURE REVIEW

The Emergence of Model Trees

The “divide-and-conquer” paradigm offers a practical path to solving complex problems with simple tools. Its basic idea is to break down a given problem into two or more similar, but simpler, subproblems, which are solved in turn to compose the final solution (Kleinberg and Tardos, 2006). Following a similar strategy, the methodology of Model Trees (MT) had emerged to find the

middle ground between two extremities of the interpretability spectrum. The MT development started as an extension to the technique of classification trees by Morgan and Sonquist (1963), which used the automatic interaction detection (AID) method to generate Regression Trees (RT). RT creates a predictive model represented by a tree structure where the feature domain is split between tree branches; at the end of each branch, there are *tree leaves* associated with constant values. Breiman et al. (1984) improved on AID and developed Classification and Regression Tree (CART) to approximate nonlinear functions by discretizing them into piecewise models, suitable for predicting both continuous and categorical variables. Quinlan (1992) extended CART application by replacing constant values on tree leaves with linear functions, resulting in a method called *M5*; as a result, enabling CART to model nonlinear datasets with piecewise linear functions. More recently, Loh (2002) presented a new algorithm called GUIDE for Generalized, Unbiased Interaction Detection and Estimation. GUIDE divides the residuals of a linear model into negative and positive signs and uses a Chi-square test instead of a t-test to determine the best split. Alternating Trees (ATREE) is one of the latest methods proposed in connection with MT development that uses an additive forward stage-wise approach to build the tree model (Frank et al. 2015).

Figure 1 (a) provides a simplified representation of data points in a given domain. *Classification Tree* is applied to cluster data into discrete classes through binary recursive partitioning over training data and generate a tree-like structure for future prediction (Figure 1 (b)). *Regression Tree* is similar to *classification trees*; however, constant values are assigned to tree leaves instead of classes, therefore, resulting in a point prediction that is associated with a specific tree leaf (Figure 1 (c)). To replace constant values in RT, *Model Tree* attaches regression models to the ends of their branches and utilize the input-output correlation to generate future predictions (Figure 1 (d)).

MT splits the data points in such a way that similar samples are clustered for performing multiple linear regression (MLR) at each leaf of the tree model (Figure 2), resulting in a piecewise regression model, as shown in Figure 3 (Quinlan 1992).

The resulting MT model is not only more accurate than commonly applied MLR methods for numeric predictions, but also more transparent and insightful than the opaque structure and implicit formulation of ANN (Frank et al. 1998). In short, MT is an analytical method for creating predictive models with particular emphasis on the generalization of the problem.

Figure 1. Illustration of model tree evolution: (a) Data point representation, (b) Classification, (c) RT, (d) MT

Figure 2: Illustration of MT splitting, branching structure and regression formulation at end leaf

Figure 3: Illustration of MT piecewise representation of nonlinear trend in data

Model Tree Related Applications

Lee et al. (2004) quantified productivity loss due to project change orders by applying the GUIDE methodology, observing that significant higher accuracy could be obtained by using MT against other methods in construction where data often featured high intercorrelation and contained noise. Desai and Joshi (2010) applied decision trees with constant branch nodes to analyze and predict

labor productivity; they implemented a heuristic approach to select influential attributes in building the decision tree. Deshpande et al. (2014) compared the results from nonlinear regression, MT, and ANN in predicting the compressive strength of recycled aggregate concrete. They indicated that if the input variables were limited, ANN would result in the best single model; however, MT could produce a set of models with different levels of complexity and accuracy. Omran et al. (2016) compared the effectiveness in applying regression, MT, and ANN for predicting concrete compressive strength. They concluded that time consumed in training advanced models would hinder the model's application value, which should be considered as a factor in model comparison. Behnood et al. (2017) used MT (specifically M5) to predict concrete compressive strength. They observed that MT could provide more insight into data and hence achieve high prediction accuracy while maintaining modeling transparency. Afsarian et al. (2018) also implemented M5 as a transparent method to predict building energy consumption.

Modified Stepwise Regression

Many researchers had realized the challenge of having too many variables in developing a predictive model for an applied problem domain (Gardner et al. 2016, Said and Prathyaj 2017, Mohsenijam and Lu 2019). Redundant input parameters in a predictive model would increase the chances of over-fitting while potentially introducing noise (Gardner et al. 2016). Interdependencies between variables need to be reduced in order to achieve a more transparent model, and redundancies need to be eliminated. Note, interdependencies within the data would occur where a particular variable could be adequately explained and predicted by other input variables; therefore, there is no need to include such a variable in a predictive model.

Having the optimal number of input variables significantly reduces the chances of transferring noise from dataset to the calibrated predictive model such as over-fitting and collinearity (Ivanescu

et al. 2016). To streamline MLR models without lowering the prediction accuracy, Mohsenijam et al. (2016) developed a variable selection technique based on forward selection (FS) and backward elimination (BE), called Modified Stepwise Regression (MSR). Forward selection technique starts with a linear regression model that contains no input variables; input variables are added to MLR based on correlation and statistical significance. Model calibration terminates when there is no predictive gain in adding more variables (Seber and Lee 2003). Backward elimination begins with an MLR model having all input variables; insignificant variables would be iteratively removed to reach the final subset until there is a loss in model performance by removing any more variables (Wang and Jain 2003). It is noteworthy that either forward selection or backward elimination tends to produce a near-optimum subset of variables. Given large datasets, applying forward selection and backward elimination separately often results in two different sets of input variables (Mendenhall and Sincich 2015). As an enhancement, the MSR method combines the advantages of the two MLR model streamlining techniques while overcoming the respective limitations of each (Mohsenijam et al. 2016). The MSR starts with an empty set of input variables, similar to forward selection. In an iterative process, variables with high significance are added to MLR. In the same iteration, all the selected MLR variables are tested for their statistical significance (the same as the backward elimination method). Unlike forward selection and backward elimination, MSR allows previously eliminated variables to be reinstated in the MLR model, while the already selected variables can be removed in future iterations.

Mohsenijam and Lu (2019) proposed an application framework for developing an MLR model in its simplest form for labour-hours prediction based on design features of steel fabrication projects. The modified stepwise regression was found instrumental in identifying relevant input factors and effective in generalizing the predictive regression model. Nonetheless, as pointed out in the

conclusion by Mohsenijam and Lu (2019), “to tackle noisy, non-homogenous, and highly nonlinear data, the proposed model would likely fail due to inherent limitations of MLR.” As immediate follow-up research, we integrate MT as nonlinear classifiers prior to applying MLR regressions to model complex relationships in the problem. At the same time, MT maintains the simplicity of the final model delivered, which is essentially a system of streamlined MLR equations organized in a tree structure.

MODEL DEVELOPMENT AND RESEARCH METHODOLOGY

A Trade-off between Bias, Variance and Model Complexity

With various methods available to create predictive models for complex systems based on data, a critical question that needs to be addressed is how to balance the trade-off between bias, variance and model complexity (Yu et al. 2006). This section elaborates on how to assess combined M5 and MSR (“M5+MSR”) in terms of model complexity and misclassification error against other methods.

Bias is an indicator of the quality of fit, which can be defined as the learning error for the algorithm of choice. A model with a high bias oversimplifies the problem (e.g. using a linear model to represent a linearly inseparable problem.) Variance, as a measure of future prediction accuracy, denotes the sensitivity of the developed model to the training data provided. In other words, variance represents a model’s prediction performance on unseen or new cases, noting that different training data had been used for training. As a result, a predictive model with a high variance may not generalize patterns inherent in the problem but merely memorizes the training data, potentially giving rise to poor predictions on unseen or new cases.

Lastly, certain noise (i.e. irreducible error) inevitably exists in the data originating from errors inherent in measuring tools and human-induced errors in data collection process. The theoretical

formulation for the interaction of bias, variance and noise known as *bias-variance decomposition of error* is given in Equation 1 (Shakhnarovich 2011), where $y = f(x) + \varepsilon$ is the true function to be approximated in predictive modeling, ε is a normally distributed noise, and σ is the standard deviation. Assume that we assess a fitted function of $h(x)$ that has been calibrated on the training dataset. Given the test set of x' and observed values of y' , the $h(x)$ prediction expected error is calculated as Equation 1.

$$\begin{aligned} E[(h(x') - y')^2] &= E[h(x')^2 - 2h(x')y' + y'^2] = E[h(x')^2] - 2E[h(x')]E(y') + E[y'^2] = E[(h(x')^2 - \bar{h}(x'))^2] - (1 \\ &\quad (x')^2 - 2\bar{h}(x')f(x') + E[(y' - f(x'))^2] + f(x')^2 \\ &= E[(h(x')^2 - \bar{h}(x'))^2] + E[(y' - f(x'))^2] + [\bar{h}(x') - f(x')]^2 = \text{Var}(h(x')) + \sigma^2 + \text{Bias}^2 \end{aligned}$$

The expected error in prediction between a true function and a fitted model can be represented by the summation of variance, squared bias and noise. Therefore, to reduce prediction error, both the variance and the bias need to be minimized simultaneously. The bias (i.e., an error of the predictive algorithm) decreases as the model's complexity increases; in the meantime, it would cause the variance to escalate (Geman et al. 1992). Nevertheless, oversimplified models are often characterized with higher bias and lower variance in terms of the prediction error. The best predictive model would achieve a trade-off between bias and variance, thus minimizing the total prediction error. The "best model" can be assessed as the model that has the highest generalization ability by demonstrating the best performance in predicting new observations (Moody 1994).

Metrics for Model Assessment and Selection

Metrics that take bias, variance, and complexity into account for model assessment and selection are summarized as follows:

- (1) Root-Mean Square Error (*RMSE*), and coefficient of determination (R^2) (Lowe et al. 2006).

(2) Akaike Information Criteria (AIC) (Akaike 1977) and Bayesian Information Criteria (BIC) (Schwarz 1978).

(3) Final Prediction Error (FPE) (Yu et al. 2006).

The first group ($RSME$ and R^2) is the most widely used measurement of predictive model performance for indicating the goodness of fit and fit quality (Bias) (Lowe et al. 2006). $RSME$ and R^2 can be measured by Equation 2 and 3 respectively, where y' is the observation value, $\bar{y}$ is average of observations, h is the fitted function, n is the sample size, and D is the model's degrees of freedom. Degrees of freedom in predictive models can be defined as the number of independent variables and their coefficients affecting the dependent variable. Note a larger D indicates the higher complexity of the model structure.

AIC gives an estimate of the model performance in regard to future prediction, which can also be used as a model selection tool. A model with a smaller value of AIC has better chances of mitigating model over-fitting by penalizing the high number of degrees of freedom. In contrast to AIC, BIC penalizes more on the degree of freedom in smaller sample sizes. Therefore, AIC is a better indicator applicable to problems with large sample sizes, while BIC is more meaningful in dealing with smaller sample sizes. AIC and BIC can be calculated by Equation 4 and 5, where n is the number of data points, RSS is the residual sum of squares (Equation 6), and D is the model's degree of freedom. When two models perform comparably well with respect to AIC and BIC (i.e., producing similar RSS value on identical sample sizes), the one with a smaller degree of freedom (i.e., simpler model) is preferred.

FPE is a model fitness measure based on the trade-off between variance, bias, and complexity, where lower values of FPE is more desirable; the FPE formulation is given in Equation 7.

$$RMSE = \sqrt{\frac{\sum(y' - h(x'))^2}{(n - D)}} \quad (2)$$

$$R^2 = \frac{\sum(h(x') - \bar{y})^2}{\sum(y' - \bar{y})^2} \quad (3)$$

$$AIC = n + n\log(2\pi) + n\log(RSS/n) + 2D \quad (4)$$

$$BIC = n + n\log(2\pi) + n\log(RSS/n) + \log(n)(D) \quad (5)$$

$$RSS = \sum_{i=1}^n (y_i - f(x_i))^2 \quad (6)$$

$$FPE = [(n + D)/(n - D)].[Bias^2 + Var] \quad (7)$$

234 Proposed Research Methodology

235 The proposed research methodology explores three areas of the MT development process in the
 236 context of civil engineering applications. They are: (1) grouping recorded data into branches and
 237 leaves applying MT, (2) selecting influential attributes applying MSR on derived leaves and
 238 creating regression models, and (3) performing accuracy-complexity trade-off analysis in
 239 predictive modeling.

240 The MT model is created through (1) tree generation, (2) pruning the tree, and (3) replacing sub-
 241 trees with linear regression equations (Frank et al. 1998). The MT model structure consists of
 242 a root, splitting branches, and leaf nodes. It is noteworthy that M5 is the selected technique for
 243 implementing MT in the present research due to its simplicity and proven performance, as
 244 elaborated in the following sections.

245 1. **Tree Growing:** starting from the root node, the M5 algorithm develops the tree model by
 246 recursive partitioning. Branches are created at each splitting node. When there is no predictive

gain from partitioning, splitting stops at leaf nodes. The partitioning takes place with the aim to minimize intra-subset variability, measured by Equation 8, where SDR is the standard deviation reduction as a result of partitioning, SD is the standard deviation (Equation 9), T is the number of observations as a complete set. The resulting partition based on variable X_i at the value of a consists of two sets of observations: (1) observations where X_i is lesser or equal to a , (2) observations where X_i is greater than a . T_i is the number of observations for each partitioning condition, and $SD(T_i)$ is the standard deviation associated with the current subset. M5 determines the SDR value for input variables and then splits the dataset based on the specific variable at a particular value that would maximize the expected error reduction. Herein, SD is calculated for the output values by Equation 9, based on the observations subject to each branching condition. Splitting ends if only a few instances remain in the branch or expected error reduction is insignificant. Harrell (2015) suggested that ten (10) to fifteen (15) observations per predictor is the minimum required sample size for linear regression models in order to ensure accurate prediction. In this study, the number of predictors is only known after performing MSR on the split dataset. Hence, the minimum number of observations in each branch is set to be thirty (30) so as to enable ensuing variable selection by MSR and generate a reliable linear regression model.

$$SDR = SD(T) - \sum_i \frac{T_i}{T} \times SD(T_i) \quad (8)$$

$$SD = \sqrt{\frac{\sum_j (y_j - \bar{y})^2}{n}} \quad (9)$$

2. **Regression Fitting:** Based on the M5 algorithm, an MLR model is fitted for all the splitting nodes using data associated with that node and all the attributes leading to that node.

3. **Tree Pruning:** Pruning refers to the process of removing redundant branches and reducing the tree to a smaller size, which takes into account prediction accuracy and generalization gain in applying MLR at leaves. Specifically, MSR is implemented to analytically derive an MLR model in the leanest form by eliminating less relevant input factors in the dataset being analyzed. If the expected error of the MLR at a splitting node is lower than the combined results from its branches, then the branches are removed (pruned) –note this procedure starts from leaf nodes in an iterative process.

The flowchart for the proposed methodology of applying M5 and MSR to calibrate the regression model tree based on training data is shown in Figure 4.

Figure 4: “M5+MSR” application methodology flowchart

CASE STUDIES

Case 1: Concrete workability application

• *Data Collection*

The proposed “M5+MSR” methodology is first implemented based on a concrete workability dataset retrieved from the UCI machine learning repository (UCI 2020). Workability of concrete is generally perceived as the property that determines the effort required to handle freshly mixed concrete with the minimum loss of uniformity. In addition to its implication on productivity, the significance of this property also lies in the consideration that a concrete mixture that cannot be placed easily or compacted fully is not likely to yield the expected strength and durability characteristics (Mehta and Monteiro 1993). The workability is generally represented with the slump value (mm) gauged from the standard slump test.

Yeh (2006) collected one hundred and three (103) slump results of different concrete mixes, with different values for cement (X_1), blast furnace slag (X_2), fly ash (X_3), water (X_4), super-plasticizer (X_5), coarse aggregate (X_6), and fine aggregate (X_7), all in kg/m³. The dataset overview is given in supplementary material.

• *Model Development*

Note that detailed results for MLR and MSR implementation and comparison on the same dataset can be found in Mohsenijam et al. (2017); herein, the developed predictive models along with respective evaluation metrics are presented in Table 1 and 2. The “M5+MSR” modeling process in accordance with the proposed methodology is elaborated as follows:

Identifying the 1st splitting variable and setting its threshold: in this step, the dataset is sorted based on each variable, and standard deviation reduction (i.e. SDR as in Equation 8) is calculated

for each of the seven variables with their one hundred and three (103) values. A variable and its value with the highest SDR is chosen for the first split; the data set for future splitting is divided based on values of splitting variables. It is found that water (X_4) has the highest SDR value among all the variables ($X_4 = 187.0 \text{ kg/m}^3$ with SDR value of 1.706); sample calculation is demonstrated in Equations 10, 11, 12, and 13. The result of this split leads to a tree with two branches having 36 samples and 67 samples respectively.

$$SD(T) = \sqrt{\frac{\sum_j (y_j - \bar{y})^2}{n}} = \sqrt{7810.82/103} = 8.708 \quad (10)$$

$$SD(T_1) = \sqrt{3510.66/36} = 9.875 \quad (11)$$

$$SD(T_2) = \sqrt{1996.86/67} = 5.459 \quad (12)$$

$$SDR = SD(T) - \sum_i \frac{T_i}{T} \times SD(T_i) = 8.708 - [(36/103) \times 9.875 + (67/103) \times 5.459] = 1.706 \quad (13)$$

Identifying the second splitting variable and value: Performing the same method on the branch with 67 samples results in the selection of fine aggregate (X_7) with the value of 741.0 kg/m^3 being the splitting criteria. As a result, the current branch having 67 samples is further divided into two sub-branches having 31 samples and 36 samples respectively for MLR. At the end of this stage, the tree is generated and needed to be pruned by removing insignificant branches.

Regression fitting: in this step, a linear regression model for each leaf node needs to be fitted with samples reaching the particular node.

Pruning: as the regression models represented by the full tree (i.e. three branches and three leaves) have a significant standard deviation reduction, there is no need for pruning in this case.

Variable selection: the tree structure and the samples that reach the leaves resulting from the previous steps need to undergo MSR variable selection process [detailed steps can be found in (Mohsenijam et al. 2017)].

It is noteworthy that the resulting “M5+MSR” model maintains the tree structure with relevant variables selected for those regression models at the tree leaves (the “M5+MSR” model shown in Table 1).

Table 1. Contrasting “M5+MSR” MT model against MLR and MSR models generated from the slump dataset

Modeling Method	Derived model (selected factors and regression equations)	Splitting Condition
M5+MSR	$R1: \text{Workability} = -53.109 - 0.043 \times \text{Cement} - 0.113 \times \text{Furnace Slag} + 0.468 \times \text{Water}$	$\text{Water} \leq 184$
	$R2: \text{Workability} = 101.57 - 0.0395 \times \text{Furnace Slag} - 0.114 \times \text{Fine Aggregate}$	$\text{Water} > 184; \text{Fine Aggregate} \leq 741$
	$R3: \text{Workability} = 22.681$	$\text{Water} > 184; \text{Fine Aggregate} > 741$
MLR	$\text{Workability} = -88.525 + 0.010 \times \text{Cement} - 0.013 \times \text{Furnace Slag} + 0.006 \times \text{Fly Ash} + 0.259 \times \text{Water} - 0.184 \times \text{Super-plasticizer} + 0.030 \times \text{Coarse Aggregate} + 0.039 \times \text{Fine Aggregate}$	No Splitting
MSR	$\text{Workability} = -18.099 - 0.040 \times \text{Furnace Slag} + 0.199 \times \text{Water}$	No Splitting

• *Cross-validation Against Established Models*

Analyzing the same dataset, Yeh (2007) produced two predictive models for concrete slump: (1) Non-Linear MLR (NLMLR) model and (2) ANN model. The NLMLR model has twenty-nine (29) regression coefficients estimated by Equation 14, where x_i is the i^{th} independent variable, q is the total number of independent variables; β_i, β_{ij} are regression coefficients. The ANN model has seven (7) input variables, one (1) hidden layer with seven (7) hidden nodes, and fifty-six (56) ANN transformation weights. For further testing and analysis in this research, the NLMLR and ANN

were reproduced in the current research. A 10-fold cross-validation testing technique was used to determine model prediction performance metrics for the compared methods.

$$y = \sum_{i=1}^q \beta_i x_i + \sum_{i<j}^q \sum \beta_{ij} x_i x_j \tag{14}$$

In the Bias-Variance-Complexity analysis, performance indicators as defined in the previous section are determined for resulting models based on the slump dataset, given in Table 2. Although ANN demonstrates higher model-fit quality (i.e., lower RMSE and higher R^2), when compared with “M5+MSR”, it has lower values of AIC, BIC, and FPE, implying over-learning from the training dataset. “M5+MSR” ranks the second in terms of the least number of coefficients to be estimated next to MSR (i.e., 4) and the second in terms of lowest RMSE or R^2 next to ANN. As a result, “M5+MSR” is declared the best trade-off between fit quality, prediction accuracy, and complexity in this case.

Table 2. Contrasting different model’s complexity and performance metrics based on slump dataset

Modeling Method	Coefficients to be estimated	RMSE	R^2	AIC	BIC	FPE
MLR	8	7.42	0.32	375.35	375.44	83.30
MSR	2	7.44	0.29	369.40	369.44	78.86
M5+MSR	4	5.70	0.59	347.11	347.16	61.28
NLMLR	29	5.95	0.52	381.47	381.85	84.93
ANN	56	5.53	0.71	389.34	389.72	92.62

Case 2: Structural Steel Fabrication Application

With the increasingly complex design, construction projects require estimators to account for design details in predicting labor cost. Off-site fabrication shops provide a safer work environment for laborers conducive to higher productivity, while also removing uncertainties and risks associated with site conditions. Many research studies validated the cost, safety, quality, and

environmental advantages of prefabrication (Jaillon and Poon 2014; Li et al. 2014). Off-site fabrication of structural steel relies on skilled trades to make bespoke products at fabrication shops based on clients' requirements. In general, labor cost takes up from twenty to fifty percent of the project budget (Sweis et al. 2009). It would be crucial for a fabrication facility to have an accurate prediction of the labor effort required to complete a specific scope of work. In this section, a predictive model for steel fabrication labor productivity is developed utilizing the proposed methodology ("M5+MSR"). The objective is to develop a labor cost estimating model by correlating the work content for a certain steel fabrication project and the labor-hour requirement (in labor-hours) while analyzing model fit quality, prediction accuracy, and complexity.

- ***Data Collection***

Data gathered for this study included 1558 records for structural steel project, each record representing labor-hours spent on each steel fabrication division. These labor-hour records were associated with forty-two (42) project-related design features collected from BIM databases. For each project division, collected labor-hours were the summation of hours spent on activities such as handling, cutting, fitting, welding, and surface processing in the fabrication shop. Note the data used for this research was sourced from a steel fabricator in Alberta, Canada. Labor-hours in each division can vary from 100 to 7000 labor-hours. In order to conceal sensitive company productivity information while maintaining the patterns in the data, all the captured records were linearly scaled. The dataset overview is presented in supplementary material. Due to the space constraint, the overview is only provided for those relevant attributes selected in the resulting model. It is worth mentioning that the same dataset was also utilized in the demonstration of a framework for developing labor-hour prediction models from project design features by applying MLR, in which input features were selected and the MLR model was streamlined by implementing MSR

(Mohsenijam and Lu, 2019). The final MLR model resulting from Mohsenijam and Lu (2019) is compared and cross-checked against the proposed new methodology (“M5+MSR”). To gain access to the database, refer to the Data Availability Statement at the end of this paper.

- **Model Development**

The “M5+MSR” method was applied as shown in Figure 5. Note MSR method was applied at each tree leaf node, resulting in a streamlined MLR regression model. The “M5+MSR” model derived is given in Table 3. Note, $X1$ is the division weight (kg), $X2$ is angle length (m), $X3$ is channel length (m), $X6$ is miscellaneous beams (m), $X9$ is wide flange beam length (m), $X12$ is checker plate area (m²), $X14$ is the plate area (m²), $X18$ is hollow steel section length (m), $X19$ is round hollow steel sections length (m), and $X32$ is partial penetration weld length (m).

In particular, the selection of plate area ($X14$) by applying M5 method as the branching variable is well aligned with industry know-how. The amount of *plates* required in a steel project is a key indicator of the complexity of a project. Higher plate content in a steel fabrication project generally means more complex connections and more welding hours. Projects with fewer plates were branched into R1 (Figure 5). In R1, MSR method indicated that labor hours are attributable to those variables related to structural section lengths [i.e. wide flange ($X9$), hollow sections ($X18$), angels ($X2$) and channels ($X3$)]. This pattern is consistent with current practice. The steel industry refers to these types of projects as *Stick-built*. Projects with higher plate requirement were branched into R2 and R3 (Figure 5). MSR application identified weight ($X1$), weld length ($X32$), and round hollow steel section length ($X19$) as the main predictors of required labour hours in those “high plate content” projects. In practice, projects with high plate requirements are generally referred to as *Platework*. These projects require a significant amount of fitting and welding in fabrication and assembly.

Figure 5: M5 tree branching structure

Table 3. Calibrated regression equations resulting from MSR at respective tree leaf

Tree Leaf	Calibrated regression equations
R1	$Labor-Hours = 14.84 + 0.004 \times Division\ Weight + 0.69 \times Wide\ Flange\ Length + 0.29 \times Channel\ Length + 8.32 \times Miscellaneous\ Beam\ Length + 0.254 \times Angel\ length + 0.566 \times Hollow\ Section\ Length + 1.854 \times Plate\ Area + 1.930 \times Checker\ Plate\ Area + 11.29 \times Weld\ Length$
R2	$Labor-Hours = 290.43 + 0.014 \times Division\ Weight + 1826.1 \times Round\ Hollow\ Section\ Length + 9.871 \times Weld\ Length$
R3	$Labor-Hours = 371.5 + 0.016 \times Division\ Weight + 9.752 \times Hollow\ Section\ Length + 12.039 \times Weld\ Length$

• Model Validation

The results of the proposed “M5+MSR” methodology were compared with MLR, MSR, and ANN, as shown in Table 4. It is found that ANN has higher fit quality by having a higher R^2 ; however, similar to the findings in slump dataset; while, “M5+MSR” has lower values of AIC, BIC, FPE. As a result, “M5+MSR” achieved higher fit quality and pattern generalization performance against MSR or MLR, which means that when exposed to new data points (unseen cases), the model resulting from “M5+MSR” would have higher prediction accuracy while maintaining transparency.

Table 4. Contrasting different model’s complexity and performance metrics based on steel fabrication estimating dataset

Modeling Method	Degrees of freedom	RMSE	R^2	AIC	BIC	FPE
MLR	25	426.94	0.687	11036.1	11065.9	17256.0
MSR	6	426.64	0.682	11043.4	11012.6	16932.0
M5+MSR	16	369.86	0.763	10849.3	10914.3	15245.5
ANN	130	523.93	0.876	11313.1	12141.0	29297.8

To further validate the “M5+MSR” model, the percentage split method was used for model validation: ten percent of data (i.e. 150 unseen instances) were randomly selected and reserved to test the resulting model and calculate the prediction error. If the training and testing results are close to each other (within 10%), the model prediction accuracy is validated (Sargent 2013). The correlation coefficient and root mean square for the training and test data are given in Table 5. The difference on RMSE from training and test sets is found to fall within the 10 percent range.

Table 5. “M5+MSR” model validation results

Indicator	Training set	Test set
RMSE	369.86	399.96
R^2	0.763	0.689

• **Model Implementation**

Application of the derived labor cost prediction model on two new instances (detail of inputs given in Table 8) is presented in this section. The “unseen” data were taken from more recent projects and not used in model training. Note zero-value variables are not shown for ease of demonstration. Labor-hours required for Instance 1 with the plate area of 31.98 m² is predicted by R1; while labor-hour prediction for Instance 2 is predicted by R2, as X₁₄ is set at a value of 99.52 m². The results

for the two cases are 1713.38 and 1052.82 labor-hours, respectively. Against the actual labor-hours, the differences are 123.38 labor-hours (i.e. $1713.38 - 1590$) for Instance 1 and 140.82 labor-hours (i.e. $1052.82 - 912$) for Instance 2, representing 7.76% and 15.4% difference relative to actual labor-hours, respectively.

Table 8. Input values for input variables in two implementation cases

Input variables	Description	Unit	Instance 1	Instance 2
X1	Division Weight	Kg	56746	31660
X2	Angle	Meter	88.271	146.34
X9	Wide flange	Meter	246.155	438.04
X14	Plate	Meter-squared	31.981	99.52
X18	Hollow steel sections	Meter	194.325	0
X32	Partial penetration weld	Meter	98.318	32.332
Y	Actual Labor-hours	Hours	1590	912

CONCLUSIONS

The proposed methodology combines the advantages of Model Tree and Multiple Linear Regression (to be more specific: M5+MSR) in an attempt to achieve model sufficiency and accuracy in regression-based applications. The new methodology has been implemented to generate results on two cases presented in this study, namely: predicting concrete slump and estimating steel fabrication labor cost. Variables that play a significant part in classification and prediction are automatically selected with their threshold values analytically set by applying M5 and MSR. It is demonstrated the proposed methodology is capable of achieving higher fit quality and pattern generalization performances against commonly applied regression and ANN techniques; when presented with new data points (unseen cases), the derived model is able to make predictions with higher accuracy while maintaining simplicity and transparency.

In application problems where the relationships are extremely complex, there are no meaningful connections between variables (e.g. image recognition, computer vision). In general, the method for predictive modeling would be chosen simply based on the highest prediction accuracy achievable. On the other hand, for developing data-driven predictive models in civil engineering domain, the predictive model is desired to be (1) transparent in factor selection and reasoning logic, (2) straightforward to implement in practice, and (3) able to generalize patterns in training data for reliably predicting unseen cases. The proposed new methodology holds a high potential to cater to such application needs. In this research, variables selected for splitting in the obtained M5 model, and variables selected for regression modeling are cross-checked with established domain knowledge (concrete slump case) or well-aligned with industry practitioner's know-how (steel fabrication labor cost estimating case). This degree of transparency in reasoning logic would have been impossible to attain in the case of applying highly nonlinear regression models (such as ANN.)

It is noteworthy that *Explainable Artificial Intelligence* (XAI) has been lately formalized in the computing research community, which entails the development of a second model to explain the precise logic in the problem domain based on learning, while still delivering sufficient prediction accuracy as the primary model (Gunning 2016). So far, the majority of XAI research efforts revolve around problems relevant to the areas of sociology and image/text/sound recognition in an attempt to explain to the user how the "black box" of AI software functions (Rudin 2019). Despite significant research progress, ANN applications in civil engineering practice are scarce. The proposed new methodology will potentially complement ANN as the "Explainable AI" solution in tackling a wide range of engineering problems.

461 **DATA AVAILABILITY STATEMENT**

462 Dataset analyzed in this study is available at: (<https://figshare.com/s/8de57c3a0ca8f8ed37c4>).

463 **ACKNOWLEDGMENT**

464 The research was substantially funded by the National Science and Engineering Research Council
465 through a Discovery grant (RGPIN-2016-04687).

Draft

REFERENCES

- Afsarian, F., Saber, A., Pourzangbar, A., Olabi, A. G., Khanmohammadi, M. A. (2018). "Analysis of Recycled Aggregates Effect on Energy Conservation using M5' Model Tree Algorithm." *Energy*, 156, 264-277.
- Akaike, H. (1977) On Entropy Maximization Principle. In: Krishnaiah, P.R., Ed., *Applications of Statistics*, North-Holland, Amsterdam.
- Al-Barqawi, H., and Zayed, T. (2008). "Infrastructure Management: Integrated AHP/ANN Model to Evaluate Municipal Water Mains' Performance." *Journal of Infrastructure Systems*, 14(4), 305–318.
- Ashby, R. (1956). *An Introduction to Cybernetics*. Marine Biological Laboratory Library, Woods Hole, Mass.
- Awolusi, T. F., Oke, O. L., Akinkurolere, O. O., Sojobi, A. O., and Aluko, O. G. (2019). "Performance comparison of neural network training algorithms in the modeling properties of steel fiber reinforced concrete." *Heliyon*, 5(1), e01115.
- Behnood, A., Behnood, V., Modiri Gharehveran, M., Alyamac, K. E. (2017). "Prediction of the Compressive Strength of Normal and High-Performance Concretes using M5P Model Tree Algorithm." *Construction and Building Materials*, 142, 199-207.
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth, Belmont.
- Desai, V. S., and Joshi, S. (2010). "Application of decision tree technique to analyze construction project data." *Proc., Information Systems, Springer Berlin Heidelberg, ICISTM 2010: Information Systems, Technology and Management*, 304-313.

- Deshpande, N., Londhe, S., Kulkarni, S. (2014). "Modeling Compressive Strength of Recycled Aggregate Concrete by Artificial Neural Network, Model Tree and Nonlinear Regression." *International Journal of Sustainable Built Environment*, 3(2), 187-198.
- El-Abbasy, M. S., Senouci, A., Zayed, T., Mirahadi, F., and Parvizsedghy, L. (2014). "Artificial neural network models for predicting condition of offshore oil and gas pipelines." *Automation in Construction*, 45, 50–65.
- Frank, E., Mayo, M., Kramer, S. (2015). "Alternating model trees." *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, Salamanca, Spain, 871-878.
- Frank, E., Wang, Y., Inglis, S., Holmes, G., and Witten, I. H. (1998). "Using Model Trees for Classification." 14.
- Gardner, B. J., Gransberg, D. D., David, J. H. (2016). "Reducing Data-Collection Efforts for Conceptual Cost Estimating at a Highway Agency." *J. Constr. Eng. Manage.*, 142(11), 04016057.
- Geman, S., Bienenstock, E., Doursat, R. (1992). "Neural Networks and the Bias/Variance Dilemma." *Neural Comput.*, 4(1), 1-58.
- Gunning, D. (2016). "Explainable artificial intelligence (XAI)." *Defense Advanced Research Projects Agency (DARPA)*, <<https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf>>.
- Harrell, F. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in Statistics, Springer International Publishing, Cham.
- Ivanescu A. E., Li P., George B., Brown A. W., Keith S. W., Raju D., Allison D. B. (2016). "The importance of prediction model validation and assessment in obesity and nutrition research." *International Journal of Obesity*, 40, 887–894.

- 511 Jaillon, L., and Poon, C. S. (2014). "Life cycle design and prefabrication in buildings: A review
512 and case studies in Hong Kong." *Autom. Constr.*, 39, 195–202.
- 513 Kleinberg J., and Tardos É. (2006). *Algorithm design*. Pearson/Addison-Wesley, Boston.
- 514 Li, Z., Shen, G. Q., and Xue, X. (2014). "Critical review of the research on the management of
515 prefabricated construction." *Habitat Int.*, 43, 240–249.
- 516 Loh, W. Y. (2002), Regression trees with unbiased variable selection and interaction detection,
517 *Statistica Sinica*, vol. 12, 361-386.
- 518 Lowe, D. J., Emsley, M. W., Anthony, H. (2006). "Predicting Construction Cost using Multiple
519 Regression Techniques." *J. Constr. Eng. Manage.*, 132(7), 750-758.
- 520 Mehta, PK., and Monteiro, PJM. (1993) *Concrete – structure, properties, and materials*.
521 Englewood Cliffs, Prentice Hall Inc.
- 522 Mendenhall, W. M., Sincich, T. L. (2015), *Statistics for Engineering and the Sciences Sixth*
523 *Edition*. Chapman and Hall/CRC
- 524 Mohsenijam, A., and Lu, M. (2019). "Framework for Developing Labour-Hour Prediction Models
525 from Project Design Features: Case Study in Structural Steel Fabrication." *Canadian Journal of*
526 *Civil Engineering*, DOI: 10.1139/cjce-2018-0349.
- 527 Mohsenijam, A., Siu, M. F., Lu, M. (2017). "Modified Stepwise Regression Approach to
528 Streamlining Predictive Analytics for Construction Engineering Applications." *J. Comput. Civ.*
529 *Eng.*, 31(3), 04016066.
- 530 Moody, J. (1994). "Prediction risk and architecture selection for neural networks." *Proc., From*
531 *Statistics to Neural Networks*, Springer Berlin Heidelberg, 147-165.
- 532 Morgan, J. N., and Sonquist, J. A. (1963). "Problems in the Analysis of Survey Data, and a
533 Proposal." *Journal of the American Statistical Association*, 58(302), 415-434.

- 534 Omran, B. A., Qian, C., Ruoyu, J. (2016). "Comparison of Data Mining Techniques for Predicting
535 Compressive Strength of Environmentally Friendly Concrete." J. Comput. Civ. Eng., 30(6),
536 04016029.
- 537 Quinlan, J. R. (1992). "Learning with Continuous Classes." Proceedings of the 5th Australian Joint
538 Conference on Artificial Intelligence, AI'92, Pages 343-348. World Science.
- 539 Rudin, C. (2019). "Stop explaining black box machine learning models for high stakes decisions
540 and use interpretable models instead." arXiv:1811.10154 [cs, stat], 20.
- 541 Said, H. M., and Prathyaj, K. (2017). "Performance Measurement of Building Sheet-Metal
542 Ductwork Prefabrication Under Batch Production Settings." J. Constr. Eng. Manage., 144(2),
543 04017107.
- 544 Sargent, R. G. (2013). "Verification and validation of Simulation Models", Journal of Simulation,
545 7:1, 12-24
- 546 Seber, G. A. F., and Lee, A. J. (2003). Linear regression analysis, Wiley, New York.
- 547 Shakhnarovich, G. (2011). Notes on derivation of bias-variance decomposition in linear
548 regression. Toyota Technological Institute, Chicago, Illinois.
- 549 Schwarz G (1978). Estimating the Dimension of a Model. Annals of Statistics 6: 461–464.
- 550 Sweis, R. J., Sweis, G. J., Abu Hammad, A. A., Abu Rumman, M. (2009). "Modeling the
551 Variability of Labor Productivity in Masonry Construction." Jordan Journal of Civil
552 Engineering, 3(3), 197-212.
- 553 UCI (2020). "University of California Irvine machine learning respiratory",
554 <https://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test>
- 555 Wang, G. C. S., and Jain, C. L. (2003). Regression analysis: Modeling and forecasting, Graceway,
556 New York.

- 557 Yeh, I. C. (2006). "Exploring concrete slump model using artificial neural networks." Journal of
558 computing in civil engineering, 10.1061/(ASCE)0887-3801(2006)20:3(217), 217–221.
- 559 Yeh, I. C. (2007). "Modeling slump flow of concrete using second-order regressions and artificial
560 neural networks." Cement and concrete composites, 29(6), 474–480.
- 561 Yu, L., Lai, K. K., Wang, S., Huang, W. (2006). "A Bias-Variance-Complexity Trade-Off
562 Framework for Complex System Modeling." Proc., Computational Science and its Applications
563 - Iccsa 2006, Springer Berlin Heidelberg, 518-527.

Draft

564 **LIST OF FIGURES**

- 565 1. Illustration of model tree evolution: (a) Data point representation, (b) Classification, (c) RT, (d)
566 MT
- 567 2. Illustration of MT splitting, branching structure and regression formulation at end leaf
- 568 3. Illustration of MT piecewise representation of nonlinear trend in data
- 569 4. “M5+MSR” application methodology flowchart
- 570 5. M5 tree branching structure

Draft