# A Method for Calculating a Weight Averaged Prediction from Multiple Linear Regression Equations

*Fei Pan***
*Leonard R. Johnson***
*Christopher J. Williams*

**ABSTRACT**

This study proposes a method for combining regression equations using a relevance network model, a weight generating function, and a generalized mixed operator. The combination of these methods puts a relative weight on the predictions of each of the individual equations and then calculates a weighted average estimate. The method was validated using computer simulation structured within the Statistical Analysis System (SAS). The simulation tests demonstrated that the method is capable of making a prediction that is not significantly different from the true prediction provided the input values for the combined model fall within the valid range of at least one variable. The mean difference between the predictions using the proposed method and the prediction from the true models was less than 9.5 percent of the true model predictions for the complete set of randomized simulations. Prediction accuracy can be improved by increasing the number of variables in an equation and by broadening the width of the variable valid interval, but not necessarily by increasing the number of equations in an equation set. Individually, the number of variables is more influential than variable valid interval width on prediction accuracy.

**Keywords:** *linear regression, prediction, simulation, timber harvesting, forest operations*

## Introduction

Linear regression equations represent one tool for statistical analysis of time study data of forest harvesting operations. They provide an indication of variables that have a statistically significant effect on timber harvesting production and can be used to predict harvesting operation cycle times and production rates. There is often a high degree of uncertainty, however, when applying harvesting equations developed from a case study to other site conditions. Furthermore, harvesting equations are often developed over finite intervals for the values of the independent variables, which limit their applications for the out-of-interval conditions.

The ability to use developed regression equations from multiple sites should result in more confidence in the resulting estimates. This process, however, will only be effective if the method of developing predictions from multiple regression equations allows for ranking of multiple regression equations so that appropriate weights can be allocated to each equation.

Multi-attributes decision-making (MADM) problems deal with the rating and ranking of competing courses of action (Ribeiro 1996, Pereira and Ribeiro 2003). Attributes can be physical, economic, or have any other characteristic of the alternatives that the decision maker considers as relevant criteria for alternative selection. MADM problems are usually modeled by choosing a set of relevant attributes that characterize a finite number of alternatives or courses of action and by eliciting their relative weights (Pereira and Ribeiro 2003). This provides the theoretical basis for the problem of ranking multiple linear regression equations.

The objectives of this research were:

1. to formulate a method that uses the MADM theory to evaluate the relative weights of multiple regression equations so that a weight-averaged prediction can be generated, and
2. to validate the method through computer simulations so that the proposed method will be statistically reliable.

## Methodology

Suppose two regression Equations [**1**] and [**2**] have been developed from two case studies for skidding operations (**Table 1**):

$$\text{Skidding cycle time} = 0.4\,(\text{skidding dist.}) + 0.5\,(\#\text{ of trees per cycle}) \quad [\mathbf{1}]$$

$$\text{Skidding cycle time} = 0.2\,(\text{slope}) + 0.6\,(\text{skidding dist.}) \quad [\mathbf{2}]$$

All of the regression coefficients are standardized.

The calculation for a weight-averaged prediction involves six steps (the detailed calculations for each step are summarized in **Table 2**):

1. Use a relevance network model and the input value $X_{ij}$ for the independent variable to generate the variable satisfac-

The authors are, respectively, Graduate Research Assistant (feipan@vandals.uidaho.edu) and Professor Emeritus (ljohnson@uidaho.edu), Dept. of Forest Products; and Professor (chrisw@ uidaho.edu), Dept. of Statistics, Univ. of Idaho, Moscow, ID. This paper was received for publication in February 2008.

* Forest Products Society member.
International Journal of Forest Engineering 20(1): 9-16.

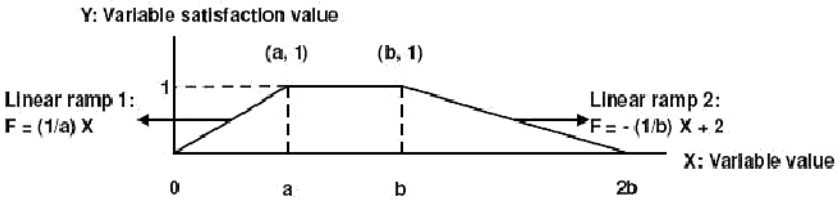


**Figure 1.** ~ *The structure of the relevance network model when the maximum allowed value is 2b.*

**Table 1.** ~ *Summary of skidding harvesting equations parameters used as illustrative example.*

| | Equation | Slope | Skidding distance | Number of trees per cycle |
|---|---|---|---|---|
| Standardized coefficient | 1 | $\alpha = 0.0$ | $\alpha = 0.4$[a] | $\alpha = 0.5$ |
| | 2 | $\alpha = 0.2$ | $\alpha = 0.6$ | $\alpha = 0.0$ |
| Variable data range | 1 | N/A | [300, 1000] | [10, 20] |
| | 2 | [5,15] | [600, 1500] | N/A |
| User input value | 1 | N/A | 1400 | 4 |
| | 2 | 27 | 1400 | N/A |

[a] $\alpha$ = the standardized regression coefficient.

tion value $F_{ij}$ for all of the variables with $0 \le F_{ij} \le 1$. $i$ and $j$ represent the $i^{th}$ variable and the $j^{th}$ equation, respectively.

The variable satisfaction value $F_{ij}$ is determined from the following relevance network model (Merida and Rollon 2000, Hartsough et al. 2001) (**Fig. 1**). If the input value $X$ is in the range of the observed values used to develop the regression equation [a, b], a value of 1 will be assigned. If the input value $X$ is more than twice the range of limits ([2b, + ∞]) or less than 0 ([–∞, 0]), a value of 0 will be assigned. If the input value $X$ is between 0 and the interval lower limit ([0, a]) or between the interval upper limit and twice the upper limit ([b, 2b]), a value between 0 and 1 will be generated through the linear ramps (**Fig. 1**) and assigned. In this relevance network model, 0 and 2b are the allowed interval boundaries and are defined as the minimum and maximum allowed value, respectively.

2. Use the variable satisfaction value $F_{ij}$ and the absolute value of standardized regression coefficient $\alpha_{ij}$ to generate the unnomalized variable weight $W_{ij}$ by a weight generating function (3) with the $\beta$ parameter equal to 1 (Pereira and Ribeiro 2003).

$$W_{ij}(F_{ij}) = \alpha_{ij}\frac{1+\beta F_{ij}}{1+\beta} \qquad [3]$$

where:

$\alpha_{ij}$ = the relative importance of the $i^{th}$ variable among all of the variables in the $j^{th}$ equation, $0 < \alpha_{ij} \le 1$;

$\beta$ = the weight dependence on the $F_{ij}$s, $0 \le \beta \le 1$.

When applying the weight generating function (3) to the regression equations, the values of $\alpha_{ij}$ can be reflected by the absolute value of the regression equation standardized coefficients as they provide an indication of the relative importance of the

**Table 2.** ~ *Calculation of weight-averaged prediction for skidding and felling equations used in the illustrative example.*

| | | | |
|---|---|---|---|
| **Step 1: Calculate the variable satisfaction values $F_{ij}$** | | | |
| | Slope | Skidding distance | Number of trees per cycle |
| Equation 1 | $F_{11}$[a] = 0 | $F_{21} = (-1400/1000) + 2 = 0.6$ | $F_{31} = 4/10 = 0.4$ |
| Equation 2 | $F_{12} = (-27/15) + 2 = 0.2$ | $F_{22} = 1.0$ | $F_{32} = 0$ |
| **Step 2: Calculate the unnormalized independent variable weights $W_{ij}$** | | | |
| | Slope | Skidding distance | Number of trees per cycle |
| Equation 1 | $W_{11} = 0(1 + 0)/2 = 0$ | $W_{21} = 0.4(1 + 0.6)/2 = 0.32$ | $W_{31} = 0.5(1 + 0.4)/2 = 0.35$ |
| Equation 2 | $W_{12} = 0.2(1 + 0.2)/2 = 0.12$ | $W_{22} = 0.6(1+1)/2 = 0.6$ | $W_{32} = 0(1 + 0)/2 = 0$ |
| **Step 3: Calculate the normalized the independent variable weights $W_{ij}^*$** | | | |
| | Slope | Skidding distance | Number of trees per cycle |
| Equation 1 | $W_{11}^* = 0/(0 + 0.32 + 0.35) = 0$ | $W_{21}^* = 0.32/(0 + 0.32 + 0.35) = 0.48$ | $W_{31}^* = 0.35/(0 + 0.32 + 0.35) = 0.52$ |
| Equation 2 | $W_{12}^* = 0.12/(0.12 + 0.6 + 0) = 0.17$ | $W_{22}^* = 0.6/(0.12 + 0.6 + 0) = 0.83$ | $W_{32}^* = 0/(0.12 + 0.6 + 0) = 0$ |
| **Step 4: Calculate the unnormalized equation weights $T_j$** | | | |
| Equation 1 | $T_1 = 0 \cdot 0 + 0.6 \cdot 0.48 + 0.4 \cdot 0.52 = 0.496$ | | |
| Equation 2 | $T_2 = 0.2 \cdot 0.17 + 1.0 \cdot 0.83 + 0 \cdot 0 = 0.864$ | | |
| **Step 5: Calculate the normalized equation weights $T_j^*$** | | | |
| Equation 1 | $T_1^* = 0.496/(0.496 + 0.864) = 0.3647$ | | |
| Equation 2 | $T_2^* = 0.864/(0.496 + 0.864) = 0.6353$ | | |
| **Step 6: Calculate the final weight-averaged prediction V** | | | |
| Combined | $V = 5.62$[b] $\cdot 0.3647 + 8.45b \cdot 0.6353 = 7.42$ | | |

[a] $i$ and $j$ represent the $i^{th}$ variable and the $j^{th}$ equation, respectively.
[b] Equation predicted values calculated using user input values.

independent variables in the prediction. Because $0 < \alpha_{ij} \leq 1$, the absolute values of the standardized coefficients are required to be normalized. In the real application, this was proven to be mathematically unnecessary.

The value of the parameter $\beta$ will be set at 1, meaning the weight of a regression equation will completely rely on the variable satisfaction values. Factors such as operator skill level, equipment type, operating season, and location might be potential factors influencing the weight generating function. These need to be considered and used in grouping the equations that can be combined. For example, the set of equations associated with operations in dry, summer conditions should not normally be combined with equations developed during the winter operating season.

3. Normalize the variable weights ($W_{ij}$), so that:

$$\sum_{i=1}^{n} W_{ij}^{*} = 1$$

where:

$W_{ij}^{*}$ = the normalized variable weight, and

$n$ = the number of variables in the $j^{th}$ equation.

4. Use the variable satisfaction value ($F_{ij}$) and the normalized variable weight ($W_{ij}^{*}$) to generate the unnormalized equation weight ($T_j$) by a generalized mixed operator. The formula is:

$$T_j = \sum_{i=1}^{n} W_{ij}^{*} F_{ij}$$

The classical weighted averaging operator is defined as:

$$WA(F)_j = \sum_{i=1}^{n} (NW_{ij}^{*})\, F_{ij}$$

where:

$WA(F)_j$ = the weight-averaged weight score for the $j^{th}$ equation and

$NW_{ij}^{*}$ = the normalized, user-assigned variable weight score (Pereira and Ribeiro 2003).

Generalized mixed operators extend the classical numerical weights to weighting functions $W_{ij}^{*}(F_{ij})$ (Pereira 2000). It is defined as:

$$W(F)_j = \sum_{i=1}^{n} W_{ij}^{*}(F_{ij}) F_{ij}.$$

5. Normalize the equation weights $T_j$, so that:

$$\sum_{j=1}^{m} T_j^{*} = 1$$

where:

$T_j^{*}$ = the normalized equation weight and

$m$ = the total number of equations.

6. Calculate the final result ($V$) using a generalized mixed operator. The formula is:

$$V = \sum_{j=1}^{n} T_j^{*} \hat{Y}_j$$

where:

$\hat{Y}_j$ = the prediction from $j^{th}$ equation using input value $X_{ij}$.

## Method Validation

### Validation Method

The proposed method for using multiple regression equations to make predictions was validated using computer simulations through the SAS 9.0 program (SAS 2003). Harvesting production regression equations were selected and used as true models. Forest harvesting operations can be subdivided into two major categories: felling-processing and transportation operations. Transportation operations (also called skidding and forwarding operations) are primarily influenced by travel distance, while the felling-processing operation is less impacted by travel distance but more affected by the size of the trees. Accordingly, skidding and felling harvesting equations from previous studies by Halbrook and Han (2005) and Adebayo (2006) were selected and used as the true models. A felling-processing equation (Schroder 1996) was selected to validate the proposed method under the circumstance of transformed variables (**Table 3**).

The original data used to develop the true felling and skidding equations were examined to determine the probability distributions of the independent variables. The resulting variable probability distributions and simple statistics were pro-

**Table 3.** ~ *Harvesting equations used for the simulation.*

| Linear regressions used for simulation | Standardized coefficient | Variable data range | $r^2$ | MSE |
|---|---|---|---|---|
| Skidding cycle time in min = 3.396 + | | | 0.66 | 1.46 |
| 0.006 (loaded travel distance) + | 0.639 | 330 to 1700 | | |
| 0.054 (re-grapple distance) + | 0.512 | 0 to 150 | | |
| 0.092 (slope %) | 0.188 | 11 to 32 | | |
| Felling cycle time in centi-min = 36.656 + | | | 0.71 | 17.61 |
| 0.908 (diameter at breast height) + | 0.091 | 3 to 21 | | |
| 0.916 (travel empty distance) | 0.836 | 0 to 400 | | |
| Square root (felling-processing cycle time in min) = –0.14852 + | | | 0.89 | 0.178 |
| 0.13053 (Ln (Tree volume per stop)) + | 0.270 | 1.66 to 209.30 | | |
| 0.62580 (Square root (Number of trees per stop)) | 0.758 | 1 to 15 | | |

grammed into the SAS random variable generating functions to recreate a simulated population that would produce the true models. A set of study site conditions was also randomly generated and designed to mimic the input values of the independent variables. These input values were constrained by the variable data range (**Table 3**) so that the true models could always make reliable predictions.

Data subsets from the simulated population were randomly selected and were used to develop regression equations from the simulated subsets. This mimics a situation where the available equations are based on a subset of the true population and where variables have a more limited range than the true model. The subset data range for an independent variable was defined as the variable valid interval [a, b]. Since equations developed over different ranges of data generally have different levels of accuracy in predicting true population values, various variable valid intervals, including narrow, medium, and wide ranges, were considered. Variable valid intervals with wide, medium, and narrow ranges were defined to cover 90 percent, 60 percent, and 30 percent of the true variable range, respectively. For example, if the true range of a variable is [330, 1700], a narrow range subset data of this variable will have an interval width of 30 percent · (1700 – 330) = 411. SAS randomly selected the lower end of a variable valid interval and used the width of the interval to determine the upper end. The result obtained was a random variable valid interval. Maximum allowed values of 2b and 3b (the minimum allowed value is 0 for both conditions) in the relevance network model were simultaneously built into the simulation code to determine which could produce a more accurate prediction.

The simulated equations from the data subsets were combined using the proposed method to make a weight-averaged prediction. The minimum number of equations (or NE) that were combined in the proposed method was two (NE = 2), since if only one equation is used to predict the dependent variable, the normalized weight of this equation will always be 1 and the function of weighted averaging disappears. When different variable valid interval widths, number of independent variables, and number of equations were considered simultaneously, the conditions that needed to be tested became so large that validation of this method by checking every condition was not a realistic option. For example, there are 63 possible equations for the simulated skidding regression equation if the number of variables and the three variable valid interval widths are considered simultaneously. The possible combinations for 2 – 63 equations would be *H*, where:

$$H = \sum_{k=2}^{63} \binom{63}{k} = 2^{63} - 64, \binom{63}{k} = 63!/[2!(63-2)!]$$

where:

! = the factorial function.

To resolve this issue, a separated random equation selection procedure in SAS was used to generate random combinations of equations (equation sets).

For each equation set generated, SAS used the proposed method to make predictions when the maximum allowed values were 2b and 3b. Meanwhile, SAS used the true model to make a prediction using the same input values. Therefore, three predictions resulted simultaneously. The process of random selection of the subsets data and calculation of the predicted values was repeated for this randomly generated set of equations 100 times. The predictions resulting from simulated equations were then compared to the true model estimates using a paired two-sample t-test. For the skidding and felling true models, the process was repeated 32 times for each "number of equations" choice (e.g., the number of equations in an equation set = 8, or NE = 8). Among the 32 equation sets, 30 were generated by random equation selection; the other two were considered as the poorest conditions where only the least significant variable was included in the simulated equation or all of the variables had narrow valid intervals. For the felling-processing equation, the process was repeated 11 times for each "number-of-equations" choice, including 10 randomly generated equation sets and one poorest condition where only the least significant variable was included. This sequence of tests was first repeated when using NE = 8, 6, 4, 3, and 2 skidding equations. If the simulation had shown that the number of equations had a significant impact on the prediction accuracy, NE = 8, 6, 4, 3, and 2 would also have been used for the felling and the felling-processing equations. The number of equations did not have a significant impact on prediction accuracy so NE = 3 and NE = 2 was used for the felling and felling-processing equations, respectively. The entire process of randomized simulations is outlined in **Figure 2**.

In order to the test the minimum variable range width for the method to make a valid prediction, simulations were con-

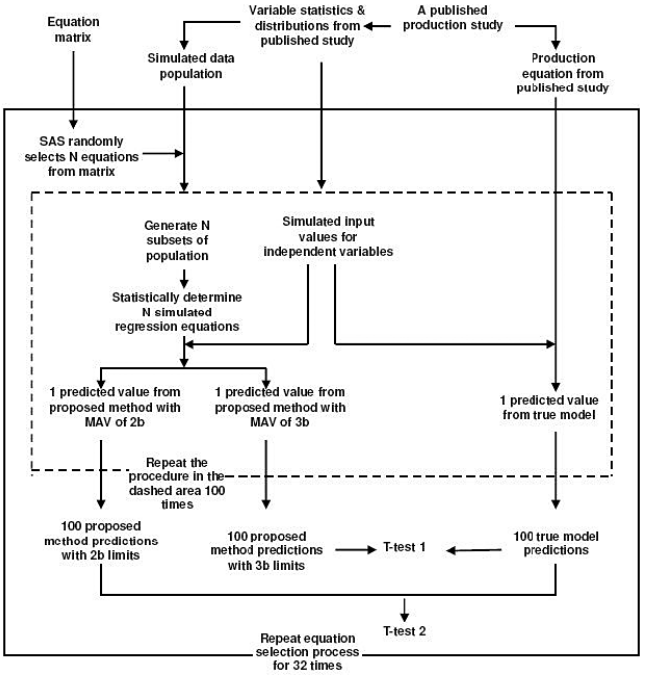


**Figure 2.** ~ *Validation process for the N random equations selected by SAS.*

ducted for the equation sets that only included the least significant variable. The variable range was initially set at narrow or 30 percent of the true variable range. Each time of simulation, the variable range width was reduced for 1 percent until the time when the method produced a prediction that was significantly different ($p < 0.05$) from the true model prediction. This variable valid interval width was regarded as the minimum range width to make a valid prediction.

## Simulation Results

### Prediction Accuracy

For the selected skidding, felling, and felling-processing equations, all of the *p*-values from the randomized simulations were greater than 0.05 (**Table 4**), meaning that the predicted values using the proposed method were not significantly different from the predictions from the true model ($\alpha = 0.05$).

For the skidding equation, the mean difference (sample size = 100) between the proposed method prediction and the true model prediction ranged from –0.257 to 0.2497 minutes based on a true model prediction range of 6.388 to 24.64 minutes. The mean difference was less than 4 percent of the true model prediction. For the felling equation, the mean difference (sample size = 100) ranged from –3.784 to 2.1162 centi-minutes based on a true model prediction range of 39.38 to 422.24 centi-minutes. The mean difference was less than 5.4 percent of the true model prediction. For the felling-processing equation, the mean difference (sample size = 100) ranged from –0.052 to 0.00032 minutes based on a true model prediction range of 0.546 to 2.983 minutes. The mean difference was less than 9.5 percent of the true model prediction.

### Effects of Number of Equations, Number of Variables, and Variable Valid Interval Width

Multiple regression analyses were performed based on the simulation results to detect the impact of the number of equations, number of variables, and variable valid intervals on prediction accuracy. Prediction accuracy (the absolute value of the mean difference between the proposed method prediction and the true model prediction) in these tests was a function of these three factors. The *p*-values associated with the regression coefficients indicate the significance of these three factors in explaining prediction accuracy. In the regression analysis, the absolute value of the mean difference acted as the dependent variable and these three factors were first analyzed as the independent variables individually. If the factors were shown to be individually significant, they were then analyzed jointly to detect their relative importance.

In the regression analysis, the linguistic range definitions were assigned range scores to make them numerical:

1. a score of 0.9 was assigned if the variable valid interval was wide,
2. a score of 0.6 was assigned if the variable valid interval was medium,
3. a score of 0.3 was assigned if the variable valid interval was narrow; and
4. a score of 0 was assigned if the equation did not include this variable.

The score for the range of a specific variable in an equation set was the average value of all of the range scores assigned to this variable. The overall average of all of the individual scores was used as the range score of the equation set.

The simulation using the skidding equation showed that increasing the number of equations did not necessarily improve prediction accuracy. Regression Equation [**4**] was developed using the results from the skidding equation simulation to evaluate the significance of the number of equations. The *p*-value of the estimated coefficient $\beta_1$ was 0.1027, indicating that number of equations was not significant in explaining the prediction accuracy ($\alpha = 0.05$).

$$\left|Mean\ difference\right| = \hat{\beta}_0 + \hat{\beta}_1 (NE) \qquad [4]$$

**Table 4.** ~ *Complete randomized simulation results ($\alpha = 0.05$).*

| | Mean difference[a] | | Standard deviation of mean difference | | *p*-value | |
|---|---|---|---|---|---|---|
| Number of equations | Min. | Max. | Min. | Max. | Min. | Max. |
| | | | Skidding equation | | | |
| 2 | –0.257 | 0.1337 | 0.0978 | 2.1077 | 0.2249 | 0.7945 |
| 3 | –0.239 | 0.1265 | 0.1147 | 1.9888 | 0.2117 | 0.9496 |
| 4 | –0.131 | 0.1595 | 0.1162 | 2.0556 | 0.2775 | 0.9045 |
| 6 | –0.187 | 0.0877 | 0.2217 | 2.1676 | 0.2640 | 0.8507 |
| 8 | –0.110 | 0.2497 | 0.1565 | 2.0357 | 0.1928 | 0.9023 |
| | | | Felling equation | | | |
| 2 | –2.749 | 1.4679 | 0.9944 | 35.262 | 0.1789 | 0.8452 |
| 3 | –3.784 | 2.1162 | 0.7354 | 36.182 | 0.1295 | 0.7207 |
| | | | Felling-processing equation | | | |
| 2 | –0.042 | 0.00032 | 0.0057 | 0.3602 | 0.1209 | 0.5790 |
| 3 | –0.052 | –0.00047 | 0.0039 | 0.3700 | 0.1452 | 0.5355 |

[a] The true skidding equation has predictions ranging from 6.388 to 24.64 (min.); the true felling equation has predictions ranging from 39.38 to 422.24 (centi-minutes); and the true felling-processing equation has predictions ranging from 0.546 to 2.983 (min.).

where:

NE = 2, 3, 4, 6, and 8.

Simulations on the three selected equations showed that when fewer variables were included in an equation, particularly when the most significant variable was absent, the absolute value of the mean difference was higher. Regression model [5] was developed in each "number-of-equations" group to detect the impact of number of the variables. The *p*-values for most of the estimated coefficients were less than 0.05 ($\alpha = 0.05$), indicating that the effect of the number of variables on the prediction accuracy is significant.

$$|Mean\ difference| = \hat{\beta}_0 + \sum_{m=1}^{3} \hat{\beta}_m (NX_m) \quad [5]$$

where:

$NX_m$ = number of $X_m$.

The simulation results showed that equations developed from a wide range of data can generate more accurate predictions than equations coming from narrow variable valid intervals. Regression model [6] was developed in each "number-of-equations" group to examine the effects of the variable valid interval. Most *p*-values for the estimated coefficients were less than 0.05 ($\alpha = 0.05$), implying that variable range has a significant influence on the prediction accuracy.

$$|Mean\ difference| = \hat{\beta}_0 + \sum_{m=1}^{3} \hat{\beta}_m (RX_m) \quad [6]$$

where:

$RX_m$ = range of $X_m$.

Simulation using the skidding and felling equations showed that the effect of the valid interval range on prediction accuracy was not as strong as the effect of the number of variables. Regression Equation [**7**] was developed in each "number-of-equations" set to compare the relative importance between number of variables and variable valid intervals. The results showed that the variable ranges mostly became insignificant ($p > 0.05$, $\alpha = 0.05$) while the number of variables was significant ($p < 0.05$).

$$|Mean\ difference| =$$

$$\hat{\beta}_0 + \sum_{m=1}^{3} \hat{\beta}_m (NX_m) + \hat{\beta}_4 (RXs) + \sum_{n=1}^{3} \hat{\beta}_{n+4} (NX_n \cdot RXs) \quad [7]$$

where:

$RXs$ = range of Xs or the equation set variable range score.

## The Poorest Cases

Simulation using the skidding equation showed that when only the least significant variable $X_3$ (slope %) was included in the simulated equation, the variable valid interval for $X_3$ was required to be wider than 20 percent of the true variable range. If it was not, the proposed method did not provide an accurate prediction (**Table 5**).

Another poor condition could occur when all of the input values were out of the valid interval ranges for all of the independent variables. The simulation results showed that the *p*-value of the t-test could be less than 0.05 under this circumstance, indicating the proposed method could make an inaccurate prediction. Simulation, however, using the same equation sets and the variable valid intervals showed that if the input values were in the valid interval of at least one variable, the proposed method was valid for making a reliable prediction (**Table 6**).

## Comparison of Various Maximum Allowed Values

Because the randomized simulation process did not universally show whether 2b or 3b was more advantageous for an accurate prediction, multiple maximum allowed values (MAVs) were simultaneously built into the SAS code to detect their probabilities in making the best and poorest predictions. The options included 3b, 2b, (2b–a), (3b–2a), and (a+b) for the

**Table 5.** ~ *Simulation results for the condition when only the least significant variable was included in the simulated equation ($\alpha = 0.05$).*

| Variable valid interval width | Mean difference | | Standard deviation of mean difference | | *p*-value | |
|---|---|---|---|---|---|---|
| | 2b[a] | 3b[a] | 2b | 3b | 2b | 3b |
| 25% | 0.271 | 0.272 | 2.0369 | 2.0371 | **0.1862**[b] | **0.1849** |
| 20% | –0.360 | –0.358 | 2.0518 | 2.0535 | **0.0820** | **0.0844** |
| 15% | –0.552 | –0.552 | 2.3394 | 2.3386 | 0.0202 | 0.0203 |
| 10% | –0.528 | –0.522 | 2.3851 | 2.3932 | 0.0292 | 0.0317 |
| 5% | –0.835 | –0.838 | 2.9224 | 2.9780 | 0.0052 | 0.0059 |

[a] Maximum allowed values used in the simulation.
[b] **Bold** indicates differences that are not significant.

**Table 6.** ~ *Simulation results for the circumstance when all but one input value was out of the variable valid interval ($\alpha = 0.05$).*

| Equation used | Number of equations | Mean difference | | Standard deviation of mean difference | | *p*-value | |
|---|---|---|---|---|---|---|---|
| | | Min. | Max. | Min. | Max. | Min. | Max. |
| Skidding | 2 | –0.165 | 0.3175 | 0.1732 | 2.8033 | 0.0574 | 0.8444 |
| Felling | 2 | –0.770 | 7.4704 | 1.6013 | 43.363 | 0.0947 | 0.8889 |

**Table 7.** ~ *Summary of the performance of various maximum allowed values.*

| Maximum allowed values | The probability to produce the best prediction | | The probability to produce the worst prediction | |
|---|---|---|---|---|
| | Skidding equation | Felling equation | Skidding equation | Felling equation |
| 2b | 1[a]/26[b] | 1/10 | 0 | 0 |
| 3b | 5/26 | 2/10 | 10/26 | 2/10 |
| 2b–a | 11/26 | 2/10 | 4/26 | 5/10 |
| 3b–2a | 2/26 | 2/10 | 5/26 | 1/10 |
| b+a | 7/26 | 3/10 | 7/26 | 2/10 |

[a] The total number of times that the specific maximum allowed value had the best predictions.
[b] The total number of experiments.

MAVs (0, 0, 2a–b, 3a–2b, and 0 for the minimum allowed values correspondingly). In the testing of these options, the random input values were constrained to be in the valid interval of only one variable because the effects of the linear ramp is only reflected by the out-of-range input values. Meanwhile, the subset data range was no wider than 60 percent of the original variable range. If the data range was wide (e.g., 90% of the original variable range), the SAS randomized data generation process might fail to generate sufficient data out of the data range to make a viable test. The true skidding model with six typical equation sets and the true felling model with three typical equation sets were used in this restricted randomized simulation.

The restricted randomized simulation results showed that all of the proposed MAVs could achieve the best prediction, but that four of them could also produce the worst prediction (**Table 7**). A favorable MAV is judged to be the result expected to have the highest probability to produce the best prediction and the lowest probability to produce the worst prediction. **Table 7** shows that the MAVs of (2b–a) and (b+a) had the highest probability to produce the best prediction but also were associated with the opportunities to make a worst prediction. The MAV of 2b never generated the worst prediction, but also had the lowest probability to produce the best prediction. Therefore, if the lowest probability to produce the worst prediction is desired, the MAV of 2b should be used; if the highest probability to produce the best prediction is preferred, the MAVs of (2b–a) or (b+a) should be applied.

## Discussion

The simulation tests showed that increasing the number of equations does not necessarily improve the prediction accuracy of the method, due to the fact that random equation generation may introduce a worse equation to an existed equation set with better equations. But, if the additional equations introduced more significant variables or broadened the range of variables included in the equation sets, they might help improve prediction accuracy. The simulation results concerning the effect of number of equations on prediction accuracy should be regarded as the outcome from the simulation, but not necessarily a strict guideline in real applications. In real applications, one of the criteria used to select equations will be that regression equations with fewer numbers of significant variables and with extremely biased variable ranges not be used. Practically, the r-square value is a good representation of the equation quality.

The simulation showed that the proposed method required that the variable valid interval width be wider than 20 percent of the true variable range to make an accurate prediction. When the variable valid interval becomes too narrow, the data from this interval will be biased enough to make the variable insignificant in prediction. In harvesting operation equations, variables that may have a narrow range include slope and diameter at breast height (DBH). For systems often observed and modeled, a slope limit of 35 percent and a cutting limit of 28 inches in DBH often represent the physical operation limits of the machine. Twenty percent of these limits will be 7 percent and 5.6 inches, respectively, and would mean that the range of slope and DBH observed in collecting the information would be less than these values. In reality, it would be difficult to find a significant relationship between this small change in the variable value and the dependent variable being modeled.

The proposed method to make predictions from multiple linear regression equations needs more research for future improvements. First, in addition to a point estimate, the proposed method should be associated with an error estimate for the point estimate as each equation used and the associated regression coefficients have their own standard error. Second, when the variable satisfaction values are all equal to 1, then it appears that the estimator is a simple average of the individual predictions. This is intuitively appealing, but it is less efficient than a weighted average based on the variances of the individual predictions.

## Conclusion

This study proposes a method for using a relevance network model, weight generating function, and generalized mixed operator to make predictions from multiple linear regression equations. Validation of this method used computer simulation to show that as long as the input values were within the valid intervals of at least one variable in one of the equations, the proposed method was effective in making predictions that were not significantly different from the true prediction. When only the least significant variable was included in the simulated equations, the variable valid interval had to be wider than 20 percent of the true variable range to ensure a valid prediction. The mean difference between the prediction using the proposed method and the prediction from the true models was less than 9.5 percent of the true model predictions for the complete randomized simulations.

The number of equations in an equation set does not necessarily improve prediction accuracy, but the number of variables in an equation and the variable valid interval are both influential for making an accurate prediction. Taken individually, the number of variables had a stronger impact on prediction accuracy than the variable valid interval. Although all five proposed maximum allowed values could result in the best prediction, the maximum allowed value of 2b was verified to be the safest choice since it never produced the worst prediction.

Use of the proposed method to make valid predictions from multiple equations still requires care in the equation selection. The selection process should avoid using equations developed from heavily biased variable ranges and with few significant variables. Further research is needed in evaluating the prediction error of this method and improving the efficiency of this method by incorporating the variance of the individual predictions.

## Acknowledgment

This study was funded by a grant from the National Fire Plan through the USDA Forest Service, Rocky Mountain Research Station.

## Literature Cited

Adebayo, A.B. 2006. Productivity and cost of cut-to-length and whole-tree harvesting in a mixed-conifer stand. MS thesis. Univ. of Idaho, Moscow, ID. 45 pp.

Halbrook, J. and H.-S. Han. 2005. Costs and constraints of fuel reduction treatments in a recreational area. USDA Forest Serv., Gen. Tech. Rep. PSW-GTR-194. 13 pp.

Hartsough, B.R., X. Zhang, and R.D. Fight. 2001. Harvesting cost model for small trees in natural stands in the Interior Northwest. Forest Prod. J. 51(4): 54-61.

Pereira, R.A.M. 2000. The orness of mixture operators: The exponential case. *In*: Proc. of the 8th International Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Madrid, Spain. 4 pp.

Pereira, R.A.M and R.A. Ribeiro. 2003. Aggregation with generalized mixture operators using weighting functions. Fuzzy Sets and Systems. 137: 43-58.

Merida, C.C. and E. Rollon. 2000. Using a relevance model for performing feature weighting. Accessed online Jan. 27, 2008 at www.lsi.upc.es/~dmerida/CCIA03/Merida-Rollon.pdf.

Ribeiro, R.A. 1996. Fuzzy multiple attribute decision making: A review and new preference elicitation techniques. Fuzzy Sets and Systems. 78: 155-181.

SAS Institute Inc. (SAS). 2003. SAS/STAT User's Guide, Version 9.0. SAS, Cary, NC.

Schroder, P.C. 1996. Small scale systems for applications to overstocked small-diameter stands. MS thesis. Univ. of Idaho, Moscow, ID. 162 pp.