# Improving computational efficiency in identifying parsimonious statistical models

Joseph Valentin[5], Ken Aho[2,5], John Edwards[3,5], Dewayne Derryberry[1,5], Teri Peterson[4,5]
Dept. of Mathematics and Statistics[1], Dept. of Biological Sciences[2], Dept. of Informatics and Computer Science[3], Dept. of Management and Marketing[4], Idaho State University[5]

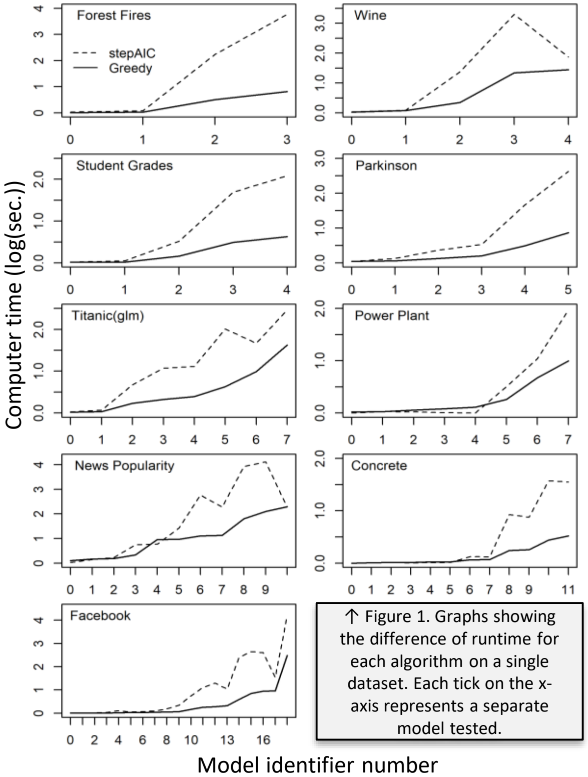
## Introduction

- **Stepwise model selection is a method of continually adding or removing predictive variables from a regression model to find the most parsimonious fit**
  - This is done by calculating a parsimony index (one that balances model complexity and fit) such as the Akaike Information Criterion(AIC) at each step and finding the optimal (lowest) value
  - StepAIC is an existing algorithm coded in the R language that performs stepwise model selection
- **StepAIC runtime is affected by many factors and can take a large amount of time to run**
  - Model complexity strongly decreases the computational efficiency of StepAIC
  - Model complexity includes dataset size, the presence of squared terms, and if interactions are being used
- **Our algorithm, named Greedy, was developed to address the issue of computational efficiency**
  - It is another stepwise model selection algorithm coded in the R language
  - Our algorithm considers models quadratic to the number of predictor variables
  - This contrasts with StepAIC, which considers models cubic to the number of predictor variables

## Materials and Methods

- **All runtimes were gathered in R Studio on the same PC**
  - Computational time was measured using an Intel i5-4690 processor and 16 GB of RAM
- **Nine datasets were obtained from the UCI Machine Learning Repository to test the algorithms**
  - The datasets ranged from 396 to 39645 observations
  - Each dataset contained 4 to 21 variables which included quantitative predictor variables and five contained categorical predictor variables
- **For each dataset, a group of linear models or generalized linear models were used**
  - The models were combinations of predictor variables and squared quantitative predictor variables
  - For each model (when applicable), the model was tested using first, second, third, or fourth order interactions
  - The model was then passed into each algorithm and the time to produce the best model was recorded

Runtime by algorithm



↑ Figure 1. Graphs showing the difference of runtime for each algorithm on a single dataset. Each tick on the x-axis represents a separate model tested.

↓ Table 1. Differences in AIC values of the models produced by both algorithms.

| Dataset | Largest AIC Difference | Average AIC Difference |
|---|---|---|
| Facebook | 29.476 | 5.72 |
| Titanic(glm) | 14.851 | 4.34 |
| Wine | 175.5 | 38.64 |
| Parkinsons | 3.72 | 0.66 |
| Concrete | 19.793 | 3.90 |
| Powerplant | 34.88 | 7.06 |
| News Popularity | 161.2 | 40.64 |
| Student Grades | 67.275 | 17.85 |
| Forest Fires | 1.494 | 0.37 |

## Acknowledgements

**This work was made possible by the NSF Idaho EPSCoR Program and by the National Science Foundation under award number IIA-1301792.**

## Results

- **For most of the models tested, Greedy had a much faster runtime (fig. 1)**
  - 78 models were tested, Greedy was faster in 60, StepAIC was faster in 13, and both had the same runtime in 5
  - Out of the 13 where StepAIC was faster, only 2 instances happened when the runtimes of each algorithm exceeded one second
  - The most StepAIC was faster by was 3.3 seconds and the most Greedy was faster by was 4.2 hours
- **Models produced by Greedy were generally as parsimonious as the ones produced by StepAIC (table 1)**
  - The largest difference in AIC values between the algorithms was 175.5 or 3.279%
  - The average difference in AIC values across all the datasets used was 10.924 or 0.206%
  - There were 25 instances of both algorithms producing the same AIC value

## Conclusion

Although the Greedy algorithm is still under development, the preliminary results are promising. Greedy consistently outperforms StepAIC in computational time while identifying models with nearly identical or similar parsimony. As the algorithm develops more, we hope to get the AIC difference to a lower level to ensure it is producing the best possible models. Once the algorithm is completed, it may replace StepAIC as the method of doing stepwise model selection in the R environment.

## Works Cited

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.