Comparative Analysis of Regression Algorithms for Determining the Age of Green Ash Trees

Ashwin Raj
Computer Science and Engineering
University of Kerala
Thiruvananthapuram, India
thisisashwinraj@gmail.com

Shijida Shain
Computer Science and Engineering
University of Kerala
Thiruvananthapuram, India
shiji.shain@gmail.com

Vishnu RK
Computer Science and Engineering
University of Kerala
Thiruvananthapuram, India
visnhurk650@gmail.com

***Abstract*—This article comparatively analyses the different regression algorithms used to evaluate the possibility of determining tree age, based simultaneously on diameter at breast height, ash canopy condition rating scale, and the crown ratio using the Green ash (*Fraxinus pennsylvanica*) species. The first step included the visualization and statistical analysis of the available data. Various regression models were fit on the data, and their error rates were calculated. In the majority cases, the mean squared error was found to be in the range of 10-2. For graphic interpretation of the regression models, scatter plots were applied. The resulting model was applied to unrelated groups of trees of known age. The difference between the actual age and the predicted age calculated by all of the models was less than ±25%. This analysis shall be useful to determine the choice of regression model to be applied.**

Keywords— Green Ash Tree, Calving Pasture Woodland, Big Slough Woodland, Tree Age Regression Model, Southeast Mormon Woodland Comparative Analysis, DBH, Crown Ratio

I. INTRODUCTION (*HEADING 1*)

Tree age is an important information to scientists and horticulturists. In the forestry industry it can be used to determine the average age of a stand of trees, provide information on how quickly trees are growing and predict the future monetary value of the site when harvesting trees In addition to that, data concerning the relationship of tree age and dendrometric parameters is very important for determining monetary value of trees and costs of tree replacement [5]. It also allow foresters to predict the productivity of the land for trees and wildlife. Results of tree growth research prove that savings resulting from the presence of trees in towns can be more than three times the cost of tree maintenance [1] [6].

It is possible to determine age by using annual rings in the perennial parts of many herbaceous dicotyledonous plants, such as in the secondary root xylem [2]. However simply counting rings will not provide us with the exact age of a tree because some years trees produce 'two rings', called false rings. Other times, they will not produce a ring where we sample a tree. This can be a major issue as it effects the estimated age significantly. The best way to determine a tree's age in a non-destructive way is to core a tree using an increment borer. However, boring trees from bark to pith is often difficult and replete with problems as high-density hardwood trees with large diameters sometimes damage manually operated cutting increment borers. It is necessary to say that there is hardly any research available to present the same issue from the reverse perspective, estimating age based on dendrometric parameters [4]. When a tree cannot be cored, or the obtained cores are unusable, then age estimates must be achieved through indirect methods, such as the regression of age upon stem diameter. Still lacking is an accurate in situ method based on the relationship between age and tree size [4]. Such a method should provide an acceptable margin of error and enable rapid results. Were one available, the work of arborists would be facilitated [4]. The regression allows users to calculate the best fit of model parameters to data [7]. The present research is focused towards comparatively analyzing the different regression methods used for forecasting the age of selected street side tree species based on dendrometric parameters: DBH, Ash Canopy Condition Rating Scale and Crown Ratio. This paper presents a comparative analysis of regression methods for predicting tree age and their graphic interpretation.

II. OBJECTIVES AND METHOD OF RESEARCH

This research is aimed towards the comparative analyses of five different regression methods for actualizing the relationship between the tree age and the three dendrometric parameters: Diameter at Breast Height (DBH), Ash Canopy Condition Rating Scale, and the Crown Ratio. The research work was divided into three different stages as described:

Stage 1: Preprocessing and statistical analysis of the dataset.
Stage 2: Model fitting and validation on group with known age to verify the extent the age forecast comply with reality.
Stage 3: Comparison of error-rate of each predictive model.

III. DATA ANALYSIS

The dataset provides current locations and inventory of the physical condition of 30 living Green Ash trees in the central Platte River Valley from 2016 [10]. The dataset was sampled from three riparian woodlands totaling 15.5 ha in size (10 trees per woodland). Strategically sampling of Green Ash along three 100–125 m length transects was done to best represent the diversity of age classes present within each woodland habitat. Physical symptoms related to potential EAB damage and other metrics related to the Green Ash dominance in each woodland were recorded. The data demonstrates a variable age among woodlands with Calving Pasture Woodland (CPW) and Big Slough Woodland (BSW) averaging 18 years older and about 13 cm wider in diameter at breast height (DBH) than the Southeast Mormon Woodland (SEMW). Accordingly, a higher percent of trees in CPW (50%) and BSW (40%) were categorized as dominant within the forest canopy compared to SEMW (20%) [10]. Descriptive summaries of measured variables for

individual woodlands and all woodlands combined are presented in Table I.

TABLE I. DENDROMETRIC FEATURES OF GREEN ASH TREE

Features	Green Ash Species		
	Calving Pasture Woodland (CPW)	***Big Slough Woodland (BSW)***	***Southeast Mormon Woodland (SEMW)***
Mean DBH	42.3	42.1	29.1
SD DBH	9.0	15.4	12.1
Mean Age	58	58	40
SD Age	12.5	21.3	16.6
Mean CCRS[a]	2.2	2.3	1.8
SD CCRS[a]	1.0	0.9	0.9
Mean CR[b]	0.54	0.62	0.73
SD CR[b]	0.21	0.13	0.13

[a.] CCRS: Canopy Condition Rating Scale [b.]CR: Crown Ratio

Now each of the three woodland tree species were plotted simultaneously on a heat map and pair plot to find the correlation between the different dendrometric parameters. Certain features were plotted on an lm plot to get a better instinct of their correlation. It is intended as a convenient interface to fit regression models across conditional subsets of a dataset.

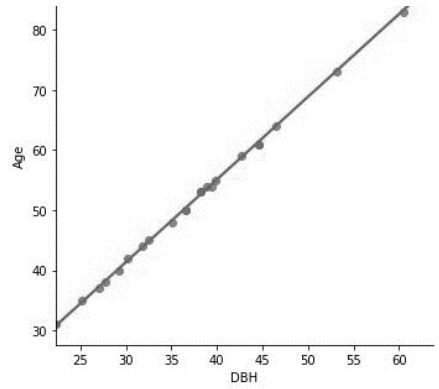

Fig. 1. Relation between the tree age and diameter at breast height

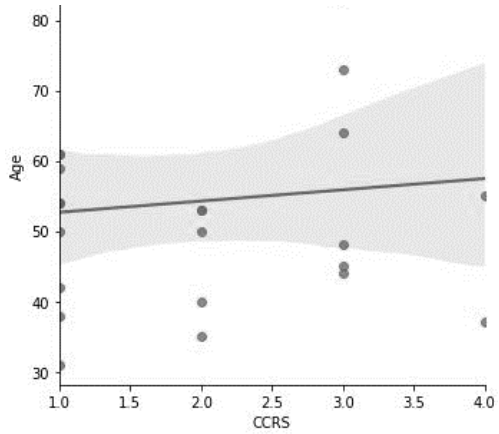

Fig. 2. Relation between tree age and canopy condition rating scale

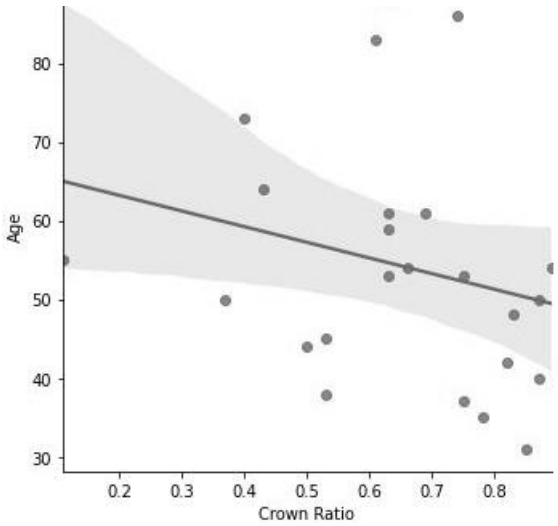

Fig. 3. Relation between tree age and crown ratio

The linear line across our plot is the best available fit for the trend of the age of the tree with respect to the different dendrometric parameters. The data points that we see at extreme top or bottom which are far away from this line are known as outliers in the dataset. We may think of outliers as exceptions. The shaded band is a point wise 95% confidence interval on the fitted values (the line). It shows us the uncertainty inherent in our estimate of the true relationship between the response and the predictor variable.

IV. MODEL VALIDATION

A. Linear Regression

A linear regression model was fit to the training data after pre-processing the data. The model was validated on the test data. The mean absolute error was found to be 0.01851954340355495 while the root mean squared error value was calculated to be around 0.019531874505177904.

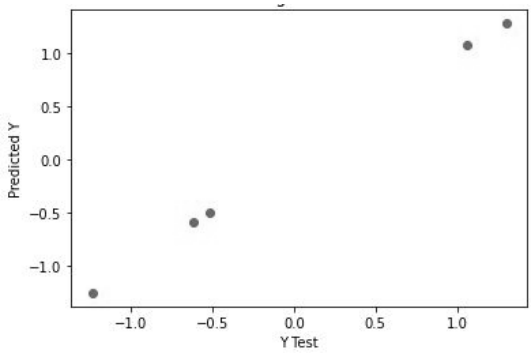

Fig. 4. Predicted values vs actual values in linear regression

B. Gradient Boosting Regression

Next the gradient boosting regression model was fit to the training data. Usually, boosting shines when there is no terse functional form around. The model was validated on the test data and the mean absolute error was found to be 0.21027609853552423. The root mean squared error value was found to be equal to 0.32145324942474596. The predicted values were then plot against known values of the test dataset.

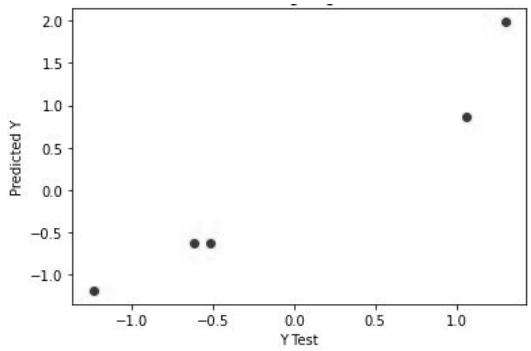

Fig. 5. Predicted values vs actual values in gradient boosting regression

C. Decisionn Tree Regression

The decision tree regression model was next to be fit to the training data. The model was validated on the test data and the mean absolute error was found to be 0.21027609853552423 while the root mean squared error value was found to be equal to 0.32145324942474596.

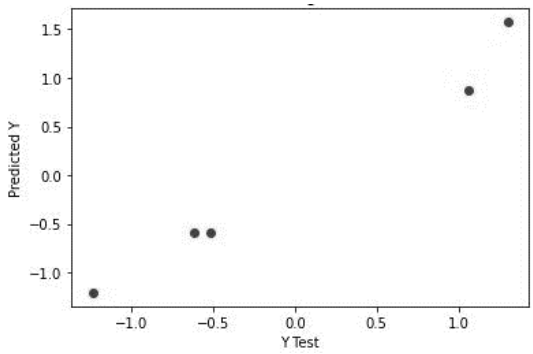

Fig. 6. Predicted values vs actual values in decision tree regression

D. Support Vector Machine Regression

The SVM regression model was fit to the training data. The model was validated on the test data and the mean absolute error was found to be 0.18883471537409952 while the root mean squared error value was found to be equal to 0.24129313473062383.

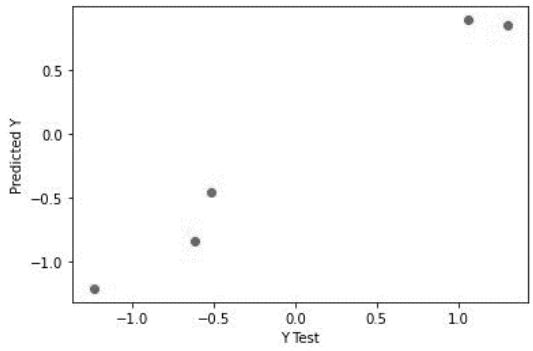

Fig. 7. Predicted values vs actual values in decision tree regression

E. Random Forest Regression

A Random Forest Regression model was fit to the training data. The model was validated on the test data and the Mean Absolute Error was found to be 0.13131186975723533 while the Root mean squared error value was found to be equal to 0.13741634521859056. These predicted values were then plot against known values.

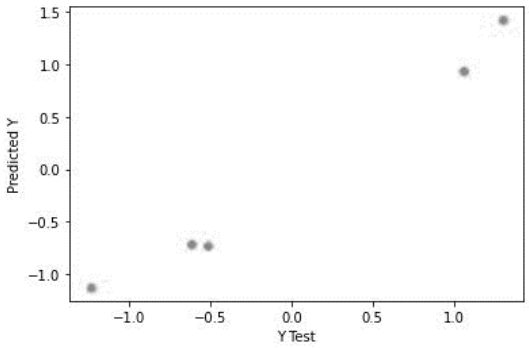

Fig. 8. Predicted values vs actual values in random forest regression

V. Using the Template

After fitting these models to the training data and validating them over our test data, error rates for the different algorithms were calculated. These are as given in Table II.

TABLE II. Error rates of different regression algorithms

Algorithm	Error Rate
Linear Regression	0.00038149
Gradient boosting Regression	0.10333219
Decision Tree Regression	0.02276836
SVM Regression	0.05822238
Random Forest Regression	0.01888325

A better way to understand the resultant outcome of error rate is by visualizing it. The plot indicating the error rate of all regression models is shown in Figure 9. Linear Regression has the lowest error rate while gradient boosting regression has the highest error rate among all regression models. Decision tree regression and Random Forest Regression were found to have close values for error rates.

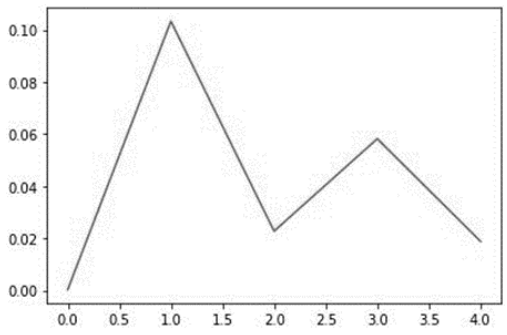

Fig. 9. Error rate of various algorithms, each corner representing error rate of a particular algorithm in the order: linear regression, gradient boosting, decision tree regression, svm regression and random forest regression respt.

VI. Error Rates of Different Algorithms

After processing the different features, we started testing the models to predict age of a tree on the Green Ash tree dataset. The obtained values for error rates is an attempt to comparatively analyze the various regression models by using three dendrometric features: DBH, CCRS and Crown

Ratio. Machine Learning models for the species were developed using the five widely used regression models. Table 2 lists the error rates estimated for each model. Based from the results, the five regression models produced different error rate estimates, even if the five models used the same dataset for the species with the exception of Decision Tree Regression and Random Forest Regression that have close values for error rates. Potential expansion of research shall take into account different planting locations and site conditions that significantly affects the dendrometric features of a tree. Besides DBH CCRS and Crown Ratio, further research shall take into account the possibility of the prediction of tree crown radius with age, because this is a useful parameter in urban designs [4].

CONCLUSION AND FUTURE WORK

The comparative analysis of the different regression model can find its use in situations when we have a positive knowledge about the dendrometric features of trees but the age remains unknown. The difference between the actual age and the predicted age calculated with the models were less than ±25%. Although the usefulness of the analysis presented in this article is limited to the tree species growing in sites and climate similar to those reported here, the methods have a general application to tree population in any city. This research serves as a basis for developing a more universal method. The analysis can be expanded with additional tree species. Currently obtained results should serve as a starting point for more extensive research to determine the relation between their age and dendrometric parameters and other environmental factors. If data is available it is a good idea to introduce more features, such as tree height, soil type, number of branches etc.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

This research was supported by the University College of Engineering, Kariavattom. We thank our colleagues from University of Kerala who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. We thank Bisharathu Beevi A, Principal, University College of Engineering, Kariavattom for assistance with the institutional resources, and Khadira Safar, Assistant Professor, University College of Engineering, Kariavattom for comments that greatly improved the manuscript. We would also like to show our gratitude to the CSE faculty at University College of Engineering, Kariavattom for sharing their pearls of wisdom with us during the course of this research, and we thank our “anonymous” reviewers for the insights they provided. We are also immensely grateful to the laboratory staff for their comments on an earlier version of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed persons.

REFERENCES

[1] Greenberg, C.H., and R.W. Simons. 1999. Age composition and stand structure of old-growth oak sites in the Florida high pine landscape: Implications for ecosystem management and restoration. Natural Areas Journal 19:30–40

[2] Hansjörg Dietz, Isolde Ullmann, Age-determination of Dicotyledonous Herbaceous Perennials by Means of Annual Rings: Exception or Rule?, Annals of Botany, Volume 80, Issue 3, September 1997, Pages 377–379, https://doi.org/10.1006/anbo.1997.0423

[3] Lomnicki, A. 2002. Introduction to Statistics for Naturalists. PWN, Warsaw, Poland.

[4] Łukaszkiewicz, Jan & Kosmala, Marek. (2008). Determining the Age of Streetside Trees with Diameter at Breast Height-based Multifactorial Model. Arboriculture and Urban Forestry. 34.

[5] McPherson, E.G. 2006. I-Tree: Demonstrating that Trees Pay Us Back. USDA Forest Service. Center for Urban Forest Research. www. fs.fed.us/psw/programs/cufr/products/powerpoint/cufr_650_48_ Webcast_04-25-06a.swf (accessed 10/10/2006). — 2007. Benefit-based tree valuation. Arboriculture and Urban Forestry 33:1–11.

[6] McPherson, E.G., S.E. Maco, J.R. Simpson, P.J. Peper, Q. Xiao, and A.E. Van der Zanden. 2002. North Western Washington and Oregon Community Tree Guide: Benefits. Costs and Strategic Planting. International Society of Arboriculture, Pacific Northwest Chapter, Silverton

[7] Motulsky, H.J., and A. Christopoulos. 2003. Fitting Models to Biological Data Using Linear and Nonlinear Regression. A Practical Guide to Curve Fitting. GraphPad Software Inc., San Diego CA. www. graphpad.com (accessed 4/11/2006).

[8] Neter, J., and W. Wassermann. 1974. Applied Linear Statistical Models. Richard D. Irwin, Inc., Homewood, IL.

[9] Sokal, R.R., and F.J. Rohlf. 1981. Biometry. The Principles and Practice of Statistics in Biological Research. 2nd Ed. W.H. Freeman and Company, New York.

[10] Wiese JD, Caven AJ. Dataset of the physical conditions of Green Ash (*Fraxinus pennsylvanica*) in riparian woodlands along the central Platte River. *Data Brief*. 2018; 21:948–952. Published 2018 Oct 24. doi:10.1016/j.dib.2018.10.063