Turbulence Probability using Principal Component Analysis and Support Vector Machine Approaches
or
PCA and SVM Require Rectangular Data Arrays and That Would Have Been a Herculean Task, So We Used Ensemble Trees

Kimberly L. Elmore[1] and Michael B. Richman[2]

ABSTRACT

A regression tree ensemble or forest is applied to a large data set containing 128 possible predictors for the purpose of predicting the probability of encountering moderate or greater turbulence during jet transport flights over the continental United States. Most of the predictors contain either missing data or outlier values. Because the data are not missing at random and the structure of the missing data is not known, imputation techniques are inapplicable. Because regression trees can accommodate missing data, a regression tree method was chosen to address the problem. A simple preprocessing step is applied to these data to remove outliers, replacing them with missing data flags.

Fifty different regression trees are generated by sampling a subset of the data at random without replacement. The fifty resulting trees tend to share similar structures and use only four or five of the same available predictors after cross validation pruning. An average of the predicted probability is then taken over all fifty trees for each set of observations. While no attempt is made to optimize the ensemble beyond generating the fifty regression trees, this method is found to work moderately well for predicting the probability of turbulence, with a Brier skill score of 0.237.

1. Introduction

Now in its third year, the Third Annual American Meteorological Society Artificial Intelligence competition focusses on various meteorological problems amenable to artificial intelligence techniques. The data set used here consists of 103,990 observations of conditions along jet transport flight routes (Figs. 1 and 2). The data set contains approximately 136 different variables. The data set also contains two additional parameters: a measure of eddy dissipation rate (EDR) and whether or the aircraft experienced moderate or greater turbulence (ISMOG).

Once a model is defined, it is applied to the “test” data that consists of 50,127observations from which EDR and ISMOG have been excised. The resulting forecasts are returned to the competition coordinator, who then scores the predictions using the Brier Skill Score (Wilks, 2006).

2. Method

a) Preprocessing

Not all of the variables are explained or defined, though some clearly have no predictive value, such as observation number and aircraft ID number. In addition, a qualitative inspection of the data indicates that neither time of day nor location have predictive value and so these, too, are removed. This leaves 128 possible predictors. Of these remaining predictors, easily half contain 50% or more missing data (some were missing more than 90% of the data). Inspection reveals that

1. Corresponding author: Kimberly L. Elmore, University of Oklahoma/National Severe Storms Laboratory, 120 David L. Boren Blvd, Norman, OK 73072; kim.elmore@noaa.gov, 405-325-6295.
2. University of Oklahoma School of Meteorology.

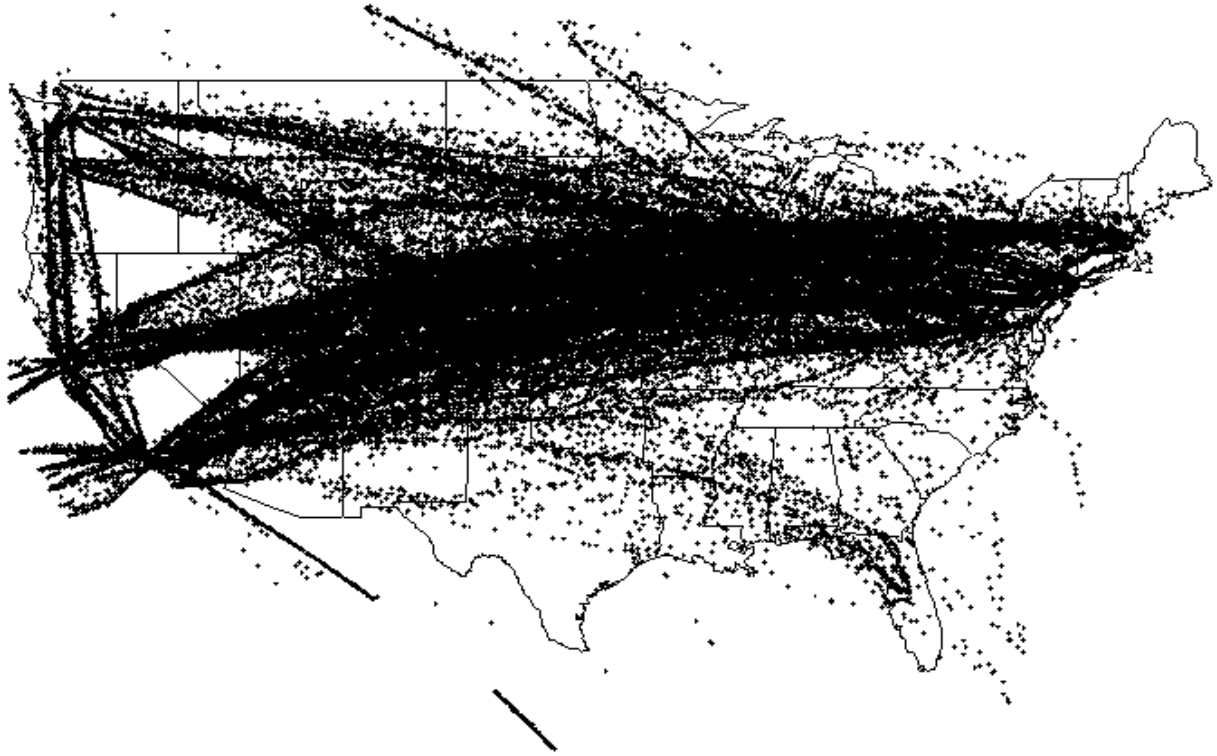

Figure 1. Locations of data points that are not associated with moderate or greater turbulence.

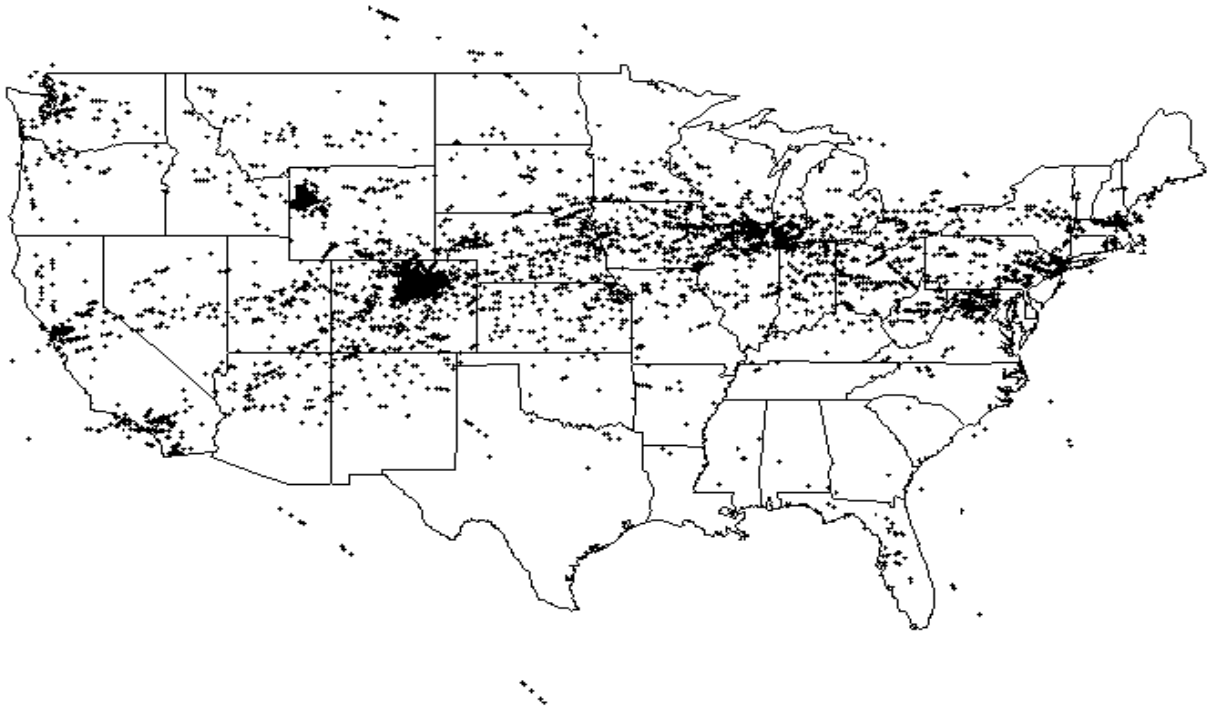

Figure 2. Data points associated with moderate or greater turbulence.

the data are not missing at random. In addition, several of the remaining predictors contain what appears to be erroneous data values.

Because the data are not missing at random, most imputation techniques are inapplicable.

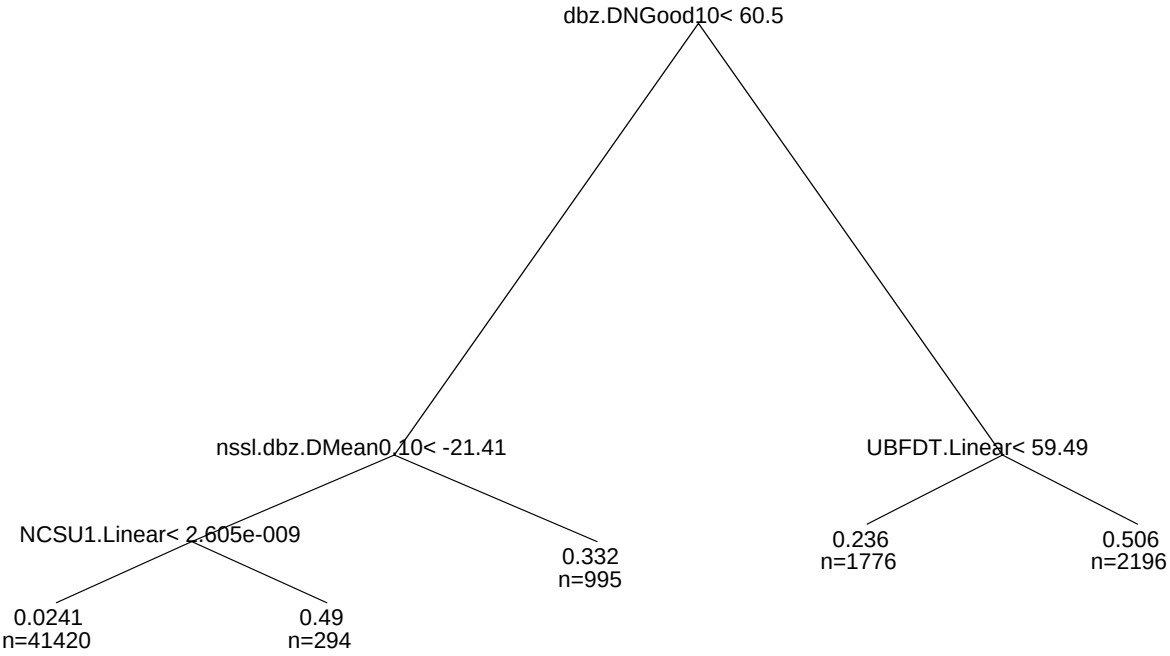

Figure 3. An example of the most prevalent tree structure within in the ensemble, dubbed the type 1 tree.

because if most of the data are missing, imputation may capture incorrect relationships, skewing the covariance or correlation structure (if a linear technique is applied) or drive the noise level high, making a nonlinear imputation technique unreliable. In some cases, it is possible to take the structure of the missing data into account and then impute the missing values, but such techniques depend on correctly assessing the missing data structure. Failure to do so leads to incorrect imputations, which can in turn lead to a model that generalizes poorly.

PCA operates on a covariance, correlation or other similarity matrix and generating a similarity matrix depends upon a rectangular data array. Given the range of scales in these data, the only obvious choice is a correlation matrix. We could compute missing data correlations (using algorithms available in S-Plus) but these use simple row deletion. If 90% of the rows are deleted, the baby is effectively discarded with the bath water. Still, it is possible to generate a series of eigenvalues and associated eigenvectors. Since 10% of ~50,000 observations is 5000 observations, we could obtain up to about 4999 positive eigenvalues. Thus, while it is possible to do this, the wisdom of mining a data set after discarding 90% of the values contained within it seems imprudent. Likewise, support vector machine (SVM) techniques expect rectangular data arrays. Because data imputation is not feasible, neither of these techniques were deemed viable. In fact, *any* technique that depends on complete data (no missing data) becomes nonviable. This seriously constrains the choices of available techniques.

In addition, we found that some variables, particularly those environmental variables with the “linear” tag in columns 97 through 128, possess statistically implausible values. We examined the quantiles of these columns in one percentile increments. Typically, we see these values range over 3 orders of magnitude in an absolute sense, but in some cases they varied over 10 or more orders of magnitude. In all cases, these implausible values were contained in the upper or lower 2-3 percent of the distribution.

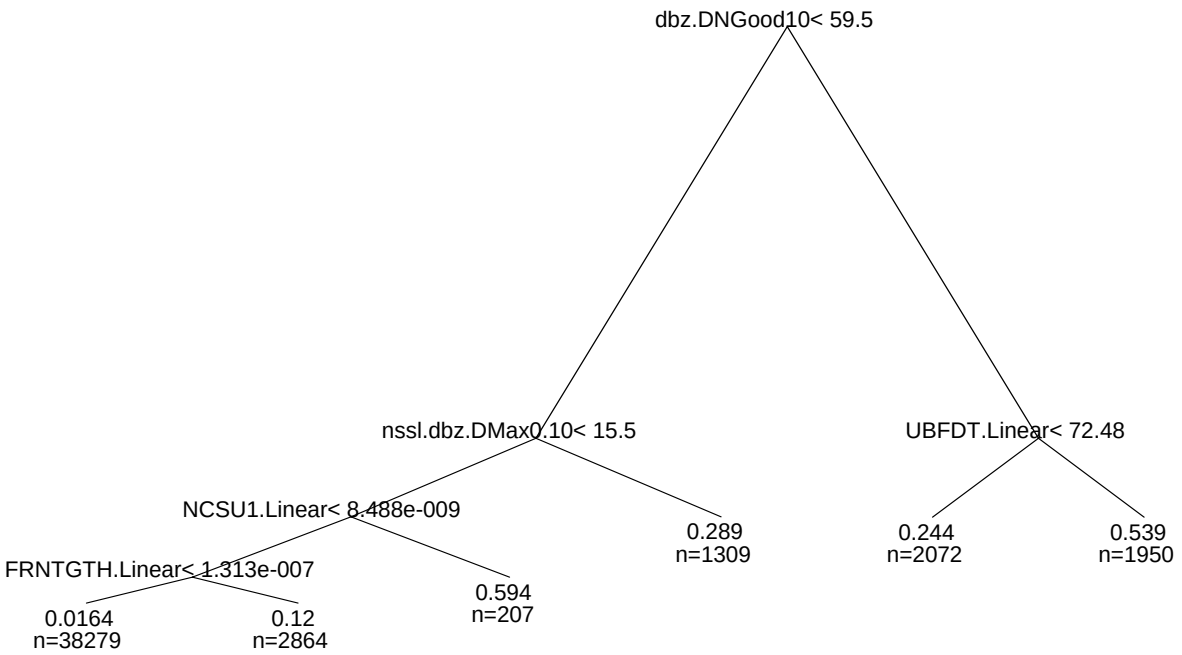

Figure 4. Example of a type 2 tree.

Because we could not be sure what these values represented, we chose to simply remove those we deemed statistically implausible. Thus, while these columns did not initially contain missing data, they did after the preprocessing step. Because only columns that did not originally contain missing data are altered by this step, the total increase in missing data values constitutes a small percentage (1-2%) though any single column may end up with as much as 5% of the values converted to missing data.

b) Method

Because of the large amount of missing data, we choose to use a method based on regression trees. A regression tree could be generated for either EDR or ISMOG, the latter being a binary variable taking on values of 0 or 1. Attempts to generate EDR values yielded poor results, so the regression was implemented on ISMOG, with the resulting values interpreted as a probability.

The particular tree method used here is part of the rpart library, available for both the open source R system and the commercial package S-Plus. The rpart library follows closely Breiman et. al. (1993). The rpart library handles missing data by replacing it with surrogate values of either all 0 or all 1, depending upon which provides the best overall performance based on cross-validation. All other data is left as is and not altered by the algorithm. In addition, rpart contains pruning algorithms that prune or simplify the resulting tree to a parsimonious model that almost certainly does not contain all available predictors. Pruning is also based on internal cross validation.

Trees are an attractive choice because they are nonparametric, nonlinear, and yet subject to easy explanation and interpretation. They accommodate hierarchical relationships, missing data, and are well-suited to mixed data types. A disadvantage is that they need large data sets, but this year's competition provided a very large data set.

The data were split such that 67% of the training data were used as an internal training set and the remaining 33% were used as an internal testing set on which the candidate model was

tested. Initial results that built a single regression tree based on 67% of the provided data appeared promising. To improve performance, we decided to implement an ensemble tree model, also known as a forest. Note that this is not a *random forest*, which is constructed using different techniques. The ensemble tree was built by randomly choosing (without replacement) 67% of the training data and building a new regression tree. For the sake of time, this was repeated only 50 times. Each model was then used to generate a probability prediction and the resulting 50 predictions averaged for each case. Due to time constraints, no attempt was made to optimize the sample size used to build the tree forest.

3. Results and Discussion.

For the data set we used, the reference Brier score is 0.061 and the resulting ensemble BSS is 0.216. This value is sensitive to the reference Brier score, which may explain why the model appears to perform better on the test data, a very unusual result.

Within the ensemble five different tree structures emerged. The most prevalent structure is

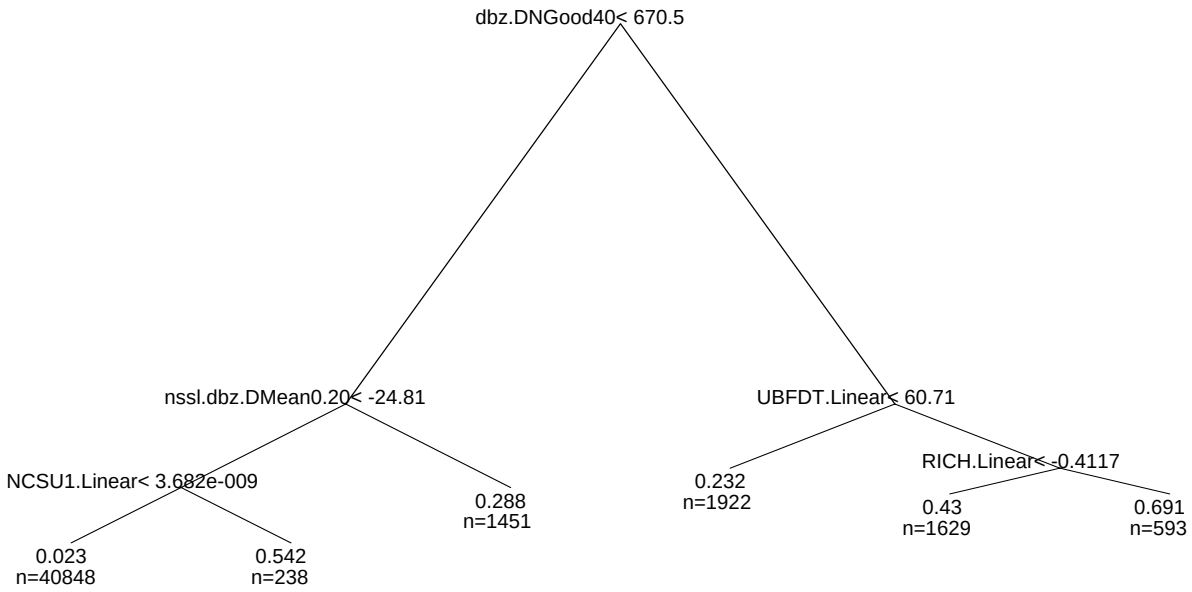

Figure 5. An example of a type 3 tree.

called type 1 (Fig. 3) and comprises 76% of the forest. Within that type, all but three members utilize the predictor called nssl.dbz.DMean0.10. The three exceptions use nssl.dbz.DMean0.20, which is similar. Type 2 (Fig. 4) trees comprise 14% ensemble forest and are similar to type 1 trees except for an extra split on the left hand side. Type 2 trees also use nssl.dbz.DMean0.10, except for two of the seven, which use nssl.dbz.DMean0.20. Type 3 (Fig. 5) and 4 (Fig. 6) trees differ only slightly and each type contains 8% of the members. A single tree of a fifth type also occurred. It is similar to a type 1 tree, but uses both the nssl.dbz.DMean0.20 and nssl.dbz.DMean0.10, but discarding the NCSU1.Linear predictor (Fig. 7).

Because we lack knowledge about what many of the predictors are, we do know which the predictors are associated with radar reflectivity. These appear in every tree and is typically associated with convection. According to the commentary associated with the competition data set notes that about 60% of the turbulence encounters are associated with convection, so the appearance of radar reflectivity near turbulence encounters is easily explained.

The similarity of all four types is likely due to the relatively large sample size used to build the regression trees. A smaller sample size is likely to display more variety, though it may not perform better. As no attempt was made to optimize the regression tree ensemble, some improvement is likely to result through more experimentation.

4. Acknowledgements

This work was funded by the National Severe Storms Laboratory and the National Oceanic and Atmospheric Administration. The authors thank Gillian Peguero and John Williams for organizing this year's competition and thereby providing an opportunity to learn more about analyzing large sparse data sets.

5. REFERENCES

Brieman, L., J. Friedman, C. J. Stone and R. A. Olshen, 1993: *Classification and Regression Trees.* Wadsworth, 358 pp.

Wilks, D. S., 2006: Statistical Methods in the Atmospheric Sciences, Second Edition. Academic Press, 627 pp.

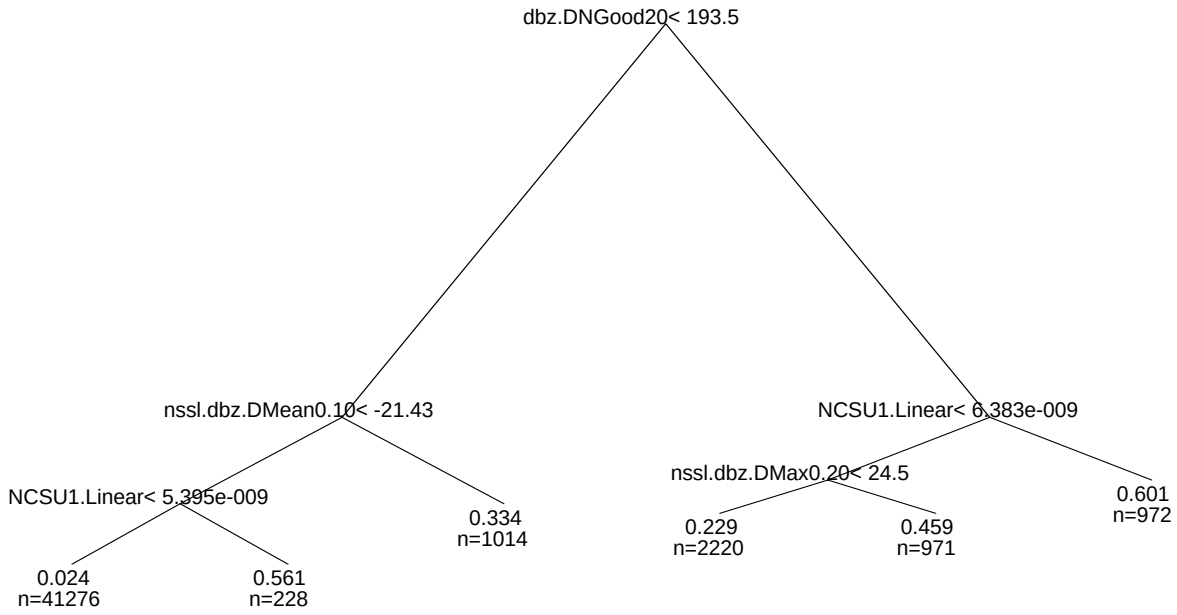

Figure 6. An example of a type 4 tree.

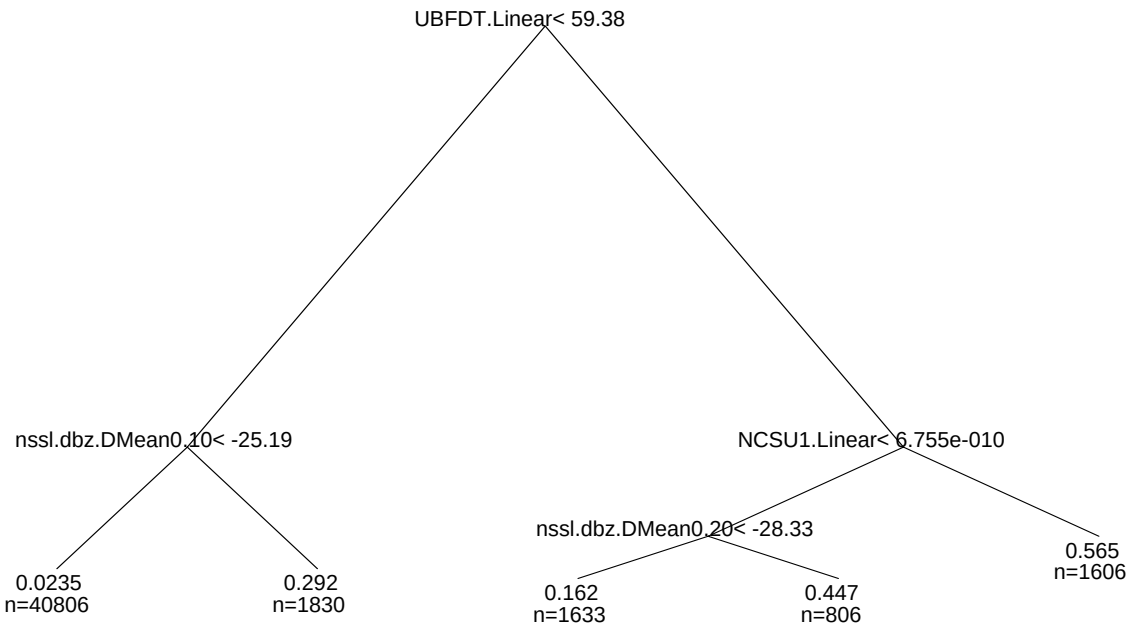

Figure 7. The fifth type of regression, only one which was constructed by the regression tree ensemble.