**Random Forest Documentation for 'Centrarchid Density'**

**1. Modeling Procedure**

We used classification Random Forests (RF) analysis to explore associations among SC Stream Assessment response variables and NFHAP spatial predictor variables (Breiman 2001, Cutler et al. 2007). Machine learning techniques such as RF provide an alternative modeling paradigm to traditional statistics, where no a priori model is defined, and complex data structures (non-normal distributions, interactions) are accommodated. Machine learning techniques use an algorithm to learn the relationship between the response and its predictors by identifying dominant patterns in the dataset (Breiman 2001, Elith et al. 2008). Random Forests represent an advance in machine learning techniques that have increased the accuracy and prediction power of single classification and regression trees by the creation of an ensemble of trees (Breiman 2001). Random forests are non-parametric, can handle both categorical and continuous data as either predictor and/or response variables, can handle high-order interactions, are insensitive to outliers, and can accommodate missing data by using surrogates (Breiman 2001, De'ath and Fabricius 2000, Urban 2002). Categorical Random Forests fit an ensemble of trees to a dataset, where each individual tree in the forest is built using a randomly selected bootstrap sample of the training dataset. In addition, only a random subset of predictor variables is considered for node and splitpoint selection (Amit and German 1997). In this way, two elements of randomness are injected into the procedure. Observations not included in the bootstrap samples are passed down their respective trees, and each tree's terminal nodes contain a predicted categorical response to different combinations of observed values among predictor variable pathways. Each tree has a 'vote' in the most important predictive variables to split on, and on the categorical responses of different values of input combinations; and the majority of votes among the ensemble of trees 'wins'. Therefore, we can a) predict and rank variables that most strongly influence an outcome (variable importance plot), and b) isolate and examine the behavior of individual predictors on the outcome, while holding the effect of all other predictive variables constant (partial dependence plots).

RF modeling was conducted by building 5000 trees using default values for other parameters in the randomForest package in the R programming environment (R Core Team 2012). RF models have known biases in variable importance selection for highly correlated predictor variables; therefore we conducted a preliminary screening of our abiotic variables to eliminate highly correlated variables. Correlations remaining in models are listed in documentation item 6.

**Literature Cited**

Amit, Y., and D. German. 1997. Shape quantization and recognition with randomized trees. Neural Computation **9**:1545-1588.

Breiman, L. 2001. Random Forests. Machine Learning **45**:5-32.

Cutler, D., T. Edwards, K. Beard, A. Cutler, K. Hess, J. Gibson, and J. Lawler. 2007. Random forests for classification in ecology. Ecology 88: 2783 – 2792.

De'ath, G., and K.E. Fabricus. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. . Ecology **81**:3178-3192.

Elith, J., Leathwick, J. R., Hastie, T. 2008. A working guide to boosted regression trees. Journal of Animal Ecology **77**:802-813.

R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Urban, D. L. 2002. Classification and regression trees.*in* B. a. J. B. G. McCune, editor. Analysis of Ecological Communities. MjM Software Design, Oregon

**2. Error Estimates Procedure**

The RF algorithm builds trees based on repeated randomized samples of the dataset, hence it is not essential to hold back data for testing after model creation to obtain an unbiased estimate of error. Model performance was evaluated with three accuracy measures calculated using the resubstitution method (Theodoridis and Kourtroumbas 2006). The three measures were the Proportion Correctly Classified (PCC), Cohen's weighted Kappa statistic (weighted $_K$), and the area under the receiver operating curve (AUC). Both PCC and weighted $_K$ are derived from the model confusion matrix, which gives the number of actual versus predicted classifications of group membership. PCC performance measures are given in two forms: 1) an overall PCC percentage (accuracy) representing the number ofcorrectly classified cases divided by the total number of cases across all outcome classes, and 2) a measure of accuracy for a specific outcome class (precision). Weighted $_K$ corrects the overall PCC for agreement caused by chance, and gives a value ranging from -1 to 1 (Cohen 1968). A positive value indicates greater agreement between modeled and measured classifications than expected by chance alone, and a negative value indicates less agreement than expected by chance alone (Table 2). Cohen's weighted $_K$ was calculated using the vcd package in R (R Core Team 2012). The AUC is derived from plotting the true positive rate (sensitivity) against the false positive rate (specificity), with each point plotted representing a sensitivity/specificity pair. The area under the resulting plot is a measure of how well the model correctly classifies groups. AUC values range from 0 to 1, with values > 0.5 indicating better model performance than expected by chance alone (Swets 1988). We used the ordROC function in the nonbinROC R package in R to calculate AUC values (http://cran.rproject.org/web/packages/nonbinROC/index.html).

**Literature Cited**

Cohen, J. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin. 70:213

R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. Science **240**:1285-1293.

Theodoridis, S., and K. Kourtroumbas. 2009. Pattern Recognition. 4th edition. Elsevier, New York.

**3. Response Variable Definition**

This metric measures the density of all centrarchid species. Centrarchids are cosmopolitan generalists that tend to increase in southeastern streams that have experienced anthropogenic disturbances such as: flow alteration, nutrient enrichment, urbanization, agriculture, and point and non-point source water quality degradation (Meyer et al. 1999, Walser and Bart 1999, Marion 2008, Detenbeck et al. 1992). Centrarchids are habitat generalists that can utilize a wide variety of habitats, velocities, and substrates, but are often found in pools and in undercut banks along stream margins. Centrarchids feed throughout the water column opportunistically; however their diet commonly consists of invertebrates and fish. Centrarchids construct large saucer-shaped nests in shallow water over a variety of substrates (sand, gravel, organic substrates). Streams with increased densities of centrarchids tend to have degraded habitat quality.

Fish data used to calculate this metric were collected as part of the South Carolina Stream Assessment. This metric was calculated using data from 167 wadeable freshwater stream sample locations in South Carolina's upstate, a region which includes four major drainage basins (Catawba/Wateree Rivers, Broad River, Saluda River, Savannah River), and four Level IV ecoregions (Blue Ridge, Inner Piedmont, Outer Piedmont, Slate Belt; Omernik 1987). Centrarchid Density is initially calculated as the number of centrarchids per 100m² stream area. This calculation standardizes centrarchid abundance measurements across sites while accounting for differences in stream sizes sampled. Several subsequent steps were taken to transform Centrarchid Density into a three-category metric, with designations for 1) high centrarchid density, 2) intermediate centrarchid density, and 3) low centrarchid density. High centrarchid densities reflect poor quality streams, and lower centrarchid densities reflect higher quality streams. Centrarchid densities were plotted in graphic array, and natural breaks in attribute data values were identified (Figure 1). Centrarchid densities less than 3.9 scored 'high' quality, densities between 4 and 7.9 scored 'intermediate' quality, and densities greater than 8 scored 'low' quality.

**Literature Cited**

Detenbeck, N. E., P.W. DeVore, G.J. Niemi, A. Lima. 1992. Recovery of temperate-stream fish communities from disturbance: a review of case studies and synthesis of theory. Environmental Management **16**:33-53.

Marion, C.A., and M.C. Scott. 2008. Interrelationships of land use and fish assemblage integrity among the tributaries of the Reedy River, South Carolina. SCDNR Statewide Fisheries Research Technical Document F-6.

Meyer, J.L., A.B. Sutherland, K.H. Barnes, D.M. Walters, and B.J. Freeman. 1999. A scientific basis for erosion and sedimentation standards in the Blue Ridge physiographic province. Pages 321-324 in K.J. Hatcher, ed. *1999 Georgia Water Resources Conference*. Institute of Ecology, University of Georgia, Athens, GA.

Omernik, J. M. 1987. Ecoregions of the conterminous United States. Annals of the Association of American Geographers 77(1): 118-125.

Walser, C. A., Bart. H.L. 1999. Influence of agriculture on in-stream habitat and fish community structure in Piedmont watersheds of the Chattahoochee River System. Ecology of Freshwater Fish **8**:237-246.

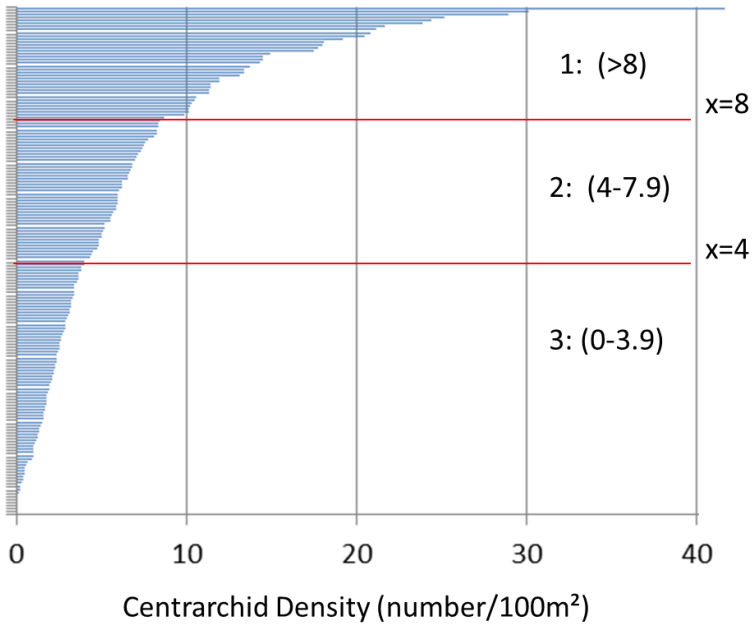



**Figure 1**. Centrarchid densities plotted in a graphic array; natural breaks identified at densities of 4 and 8. N=167. Number ranges on right side of figure denote score values (classification values) for different density ranges on plot.

**4. Variables Retained in Model**

| | | | |
|---|---|---|---|
| [1] "AREASQKM" | "Dam_coC" | "Road_lenC" | "Epa_303dC" |
| [5] "TriC" | "Mine_coC" | "C_H20_01" | "C_URBAN_01" |
| [9] "C_BARREN_01" | "C_FOREST_01" | "C_AGRICULTURE_01" | "C_WETLAND_01" |
| [13] "C_ROWCROP_01" | "SoilhygC" | "SoilpermC" | "Elev_maxC" |
| [17] "U_Score_CentDens" | | | |

**5. Model Call**

* randomForest(formula = U_Score_CentDens ~ ., data = CDS[, -1], keep.forest = TRUE, importance = TRUE, ntree = 5000, do.trace = TRUE, type = 2)
Type of random forest: classification
Number of trees: 5000
No. of variables tried at each split: 4

* The original call to randomForest (R package 'randomForest'; v4.5-36)

**6. Correlations Remaining in Model**

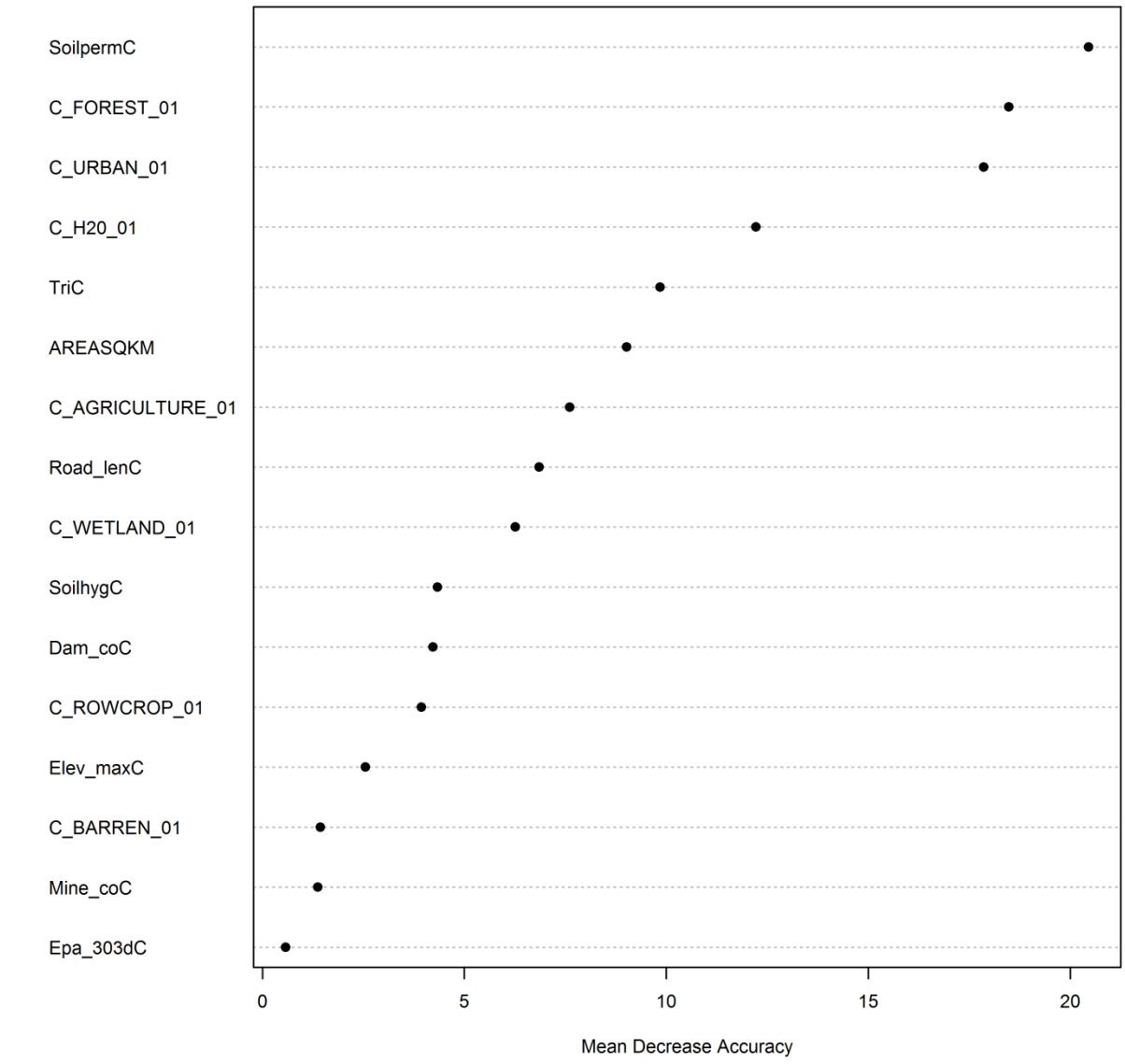
"i= 1 variables C_URBAN_01 and C_FOREST_01 correlation= -0.702056051893182"
"i= 2 variables Road_lenC and Epa_303dC correlation= 0.661663181754727"
"i= 3 variables Dam_coC and Road_lenC correlation= 0.608786954769638"
"i= 4 variables Road_lenC and TriC correlation= 0.504938383794838"
"i= 5 variables C_FOREST_01 and C_AGRICULTURE_01 correlation= -0.477143655581278"
"i= 6 variables TriC and C_URBAN_01 correlation= 0.454927254145352"
"i= 7 variables C_WETLAND_01 and C_ROWCROP_01 correlation= 0.410295771088565"
"i= 8 variables TriC and C_FOREST_01 correlation= -0.398874225408399"
"i= 9 variables C_BARREN_01 and C_WETLAND_01 correlation= 0.382913116254108"
"i= 10 variables C_ROWCROP_01 and SoilpermC correlation= 0.342768913712269"

**7. List of Important Variables (Mean Decrease Accuracy)**

**Table 1.** List of important variables in descending order of importance. Table denotes mean decrease in accuracy, and a weighting factor based on the percentage of importance explained relative to the most important predictor variable (SoilpermC).

| Predictor Variable | MeanDecreaseAccuracy | Weighting Factor |
|---|---|---|
| SoilpermC | 20.458 | 1.000 |
| C_FOREST_01 | 18.483 | 0.903 |
| C_URBAN_01 | 17.861 | 0.873 |
| C_H20_01 | 12.224 | 0.598 |
| TriC | 9.848 | 0.481 |
| AREASQKM | 9.020 | 0.441 |
| C_AGRICULTURE_01 | 7.610 | 0.372 |
| Road_lenC | 6.853 | 0.335 |
| C_WETLAND_01 | 6.266 | 0.306 |
| SoilhygC | 4.337 | 0.212 |
| Dam_coC | 4.222 | 0.206 |
| C_ROWCROP_01 | 3.940 | 0.193 |
| Elev_maxC | 2.554 | 0.125 |
| C_BARREN_01 | 1.437 | 0.070 |
| Mine_coC | 1.371 | 0.067 |
| Epa_303dC | 0.575 | 0.028 |

**8. Variable Importance Plot**

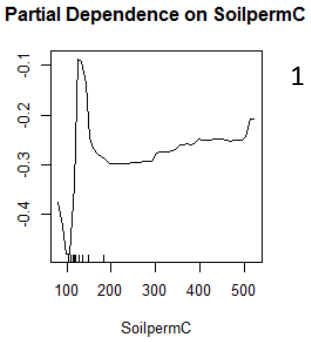



`

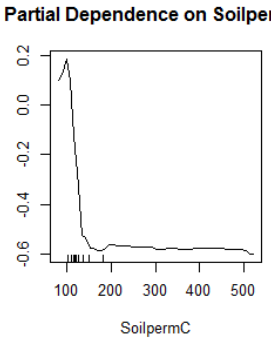
**Figure 2.** Top ranked variables from random forests classification for predicting Centrarchid Density in Upstate South Carolina.

**9. Partial Dependence Plots**

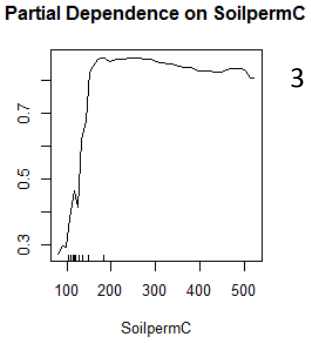
Partial dependence plots isolate and examine the relationships between top ranked predictors and Darter Richness (1=high centrarchid richness, 2=intermediate centrarchid richness, 3=low centrarchid richness) while holding the effect of all other predictive variables constant;



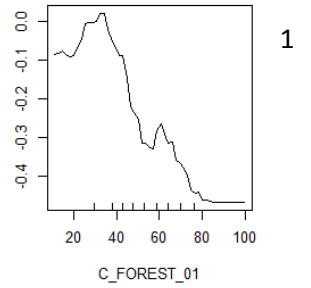

1



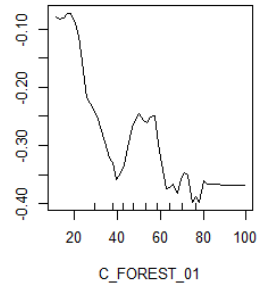

2



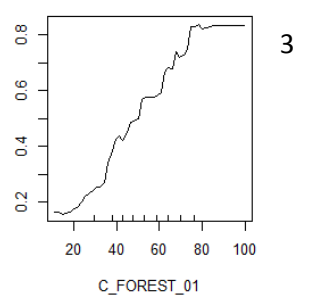

3




1

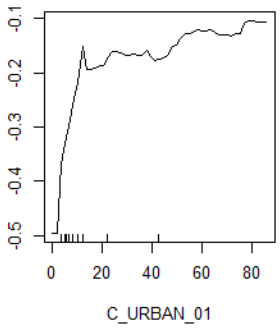



2



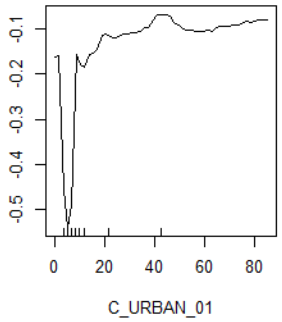

3

Partial Dependence on C_URBAN_01



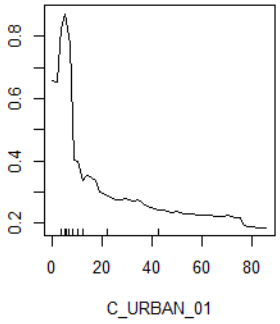

1

Partial Dependence on C_URBAN_01



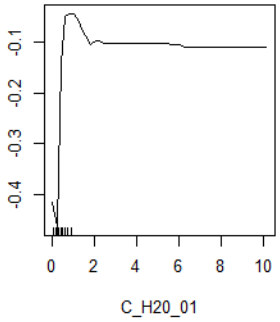

2

Partial Dependence on C_URBAN_01



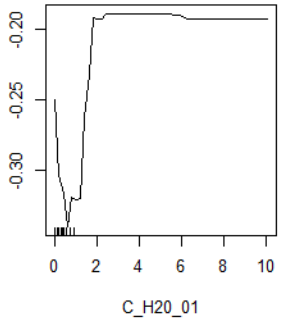

3

Partial Dependence on C_H20_01



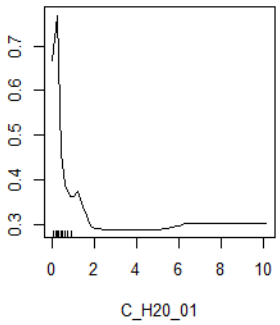

1

Partial Dependence on C_H20_01



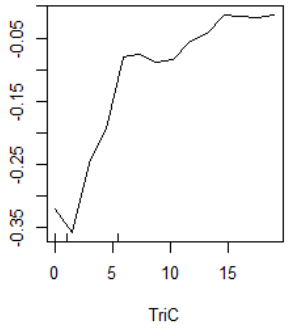

2

Partial Dependence on C_H20_01



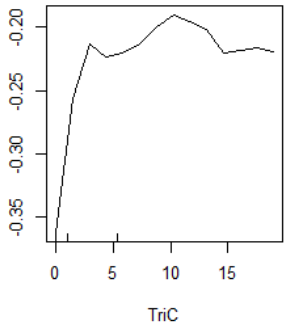

3

Partial Dependence on TriC



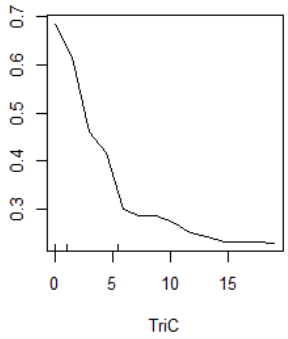

1

Partial Dependence on TriC



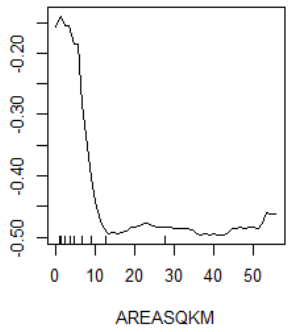

2

Partial Dependence on TriC



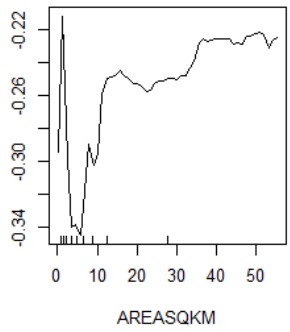

3

Partial Dependence on AREASQKM



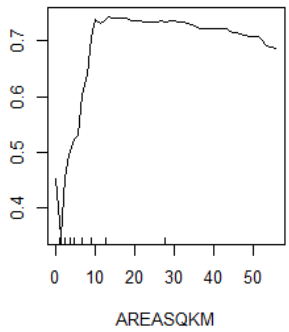

1

Partial Dependence on AREASQKM




2

Partial Dependence on AREASQKM




3

**10. PCC Result and Confusion Matrix**

| Percent Correctly Classified: 50.90 % | |
|---|---|
| Group 1 Accuracy: | 25.58 % |
| Group 2 Accuracy: | 10.00 % |
| Group 3 Accuracy: | 83.33 % |

| Confusion Matrix | | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | 11 | 8 | 24 |
| 2 | 14 | 4 | 22 |
| 3 | 9 | 5 | 70 |

**11. Weighted K Result**

*w*K = 0.19 (Slight Strength of Agreement)

**12. AUC Result**

AUC = 0.61 (Model performance better than expected by chance alone