Random Forest Documentation for ‘Coastal Plain Taxonomic Groups’

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 $_{\kappa}$), and the area under the receiver operating curve (AUC). Both PCC and weighted $_{\kappa}$ 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 of correctly 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 $_{\kappa}$ 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 $_{\kappa}$ 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

Fish data for this model were collected as part of the South Carolina Stream Assessment. The 'Coastal Plain Taxonomic Groups' metric was created using fish community data from 208 wadeable freshwater stream sample locations in South Carolina's coastal plain, a 45,773 km² area which includes four major drainage basins (Pee Dee River, Lower Santee River, ACE basin, Savannah River), and three Level IV ecoregions (Sand Hills, Atlantic Southern Loam Plains, Carolina Flatwoods (Omernik 1987). We performed a hierarchical agglomerative cluster analysis on our fish density matrix (number of species/100m stream length) using the Wards linkage algorithm to identify groups of species that commonly co-occur in coastal plain wadeable streams. A $\log_{10}(x+1)$ transformation was performed on the species density matrix to reduce the effect of large differences in fish densities among sample localities. The Euclidean distance measure was utilized to calculate species similarity. We used an indicator species analysis to assist in describing the faunal composition of each classification group and to identify the number of appropriate species groups (communities) for use in subsequent analyses (Dufrêne and Legendre 1997).

Four predominant coastal plain fish community groups were identified with hierarchical cluster analysis, and described by an indicator species analysis (Marion et al., *to be published*). The indicator species analysis found thirty-nine species as significant at the α=0.05 level, and membership was partitioned accordingly into the four fish community groups (Table 1). We assigned names to each of the four fish communities based on the prevailing taxonomic and/or ecological attributes of the species with the highest indicator values for each group.   Group one was named the fluvial community (n=12), since flow is a primary underlying ecological requirement for all of its constituents (Rhode et al. 2009). Group 2 was named the eastern mudminnow community (n=1), since *Umbra pygmaea* was the sole constituent. Group 3 was named the centrarchid community (n=10); several cosmopolitan centrarchids dominated this group including *Lepomis auritus, Lepomis macrochirus, Lepomis punctatus, and Micropterus salmoides*. Group 4 was named the non-fluvial group (n=16), since all of its members either thrive in or require low to no flow environments (Rhode et al. 2009).

Literature Cited

Dufrêne, M. and P. Legendre. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs 67(3):345-366.

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

Marion,C.A., M.C. Scott, K.M.Kubach. Multi-scale influences on SC coastal plain fish communities.

Rhode, F. C., R.G. Arndt, J.W. Foltz, and J.M. Quattro. 2009. Freshwater Fishes of South Carolina. The University of South Carolina Press, Columbia, SC.

Table 1. Species membership of four Coastal Plain Taxonomic Groups.

Scientific Name	Common Name	Group Number	Community Name
Notropis cummingsae	Dusky Shiner	1	Fluvial
Pteronotropis hypselopterus	Lowland Shiner	1	Fluvial
Etheostoma olmstedi	Tesselated Darter	1	Fluvial
Percina nigrofasciata	Blackbanded Darter	1	Fluvial
Noturus leptacanthus	Speckled Madtom	1	Fluvial
Noturus insignis	Margined Madtom	1	Fluvial
Etheostoma fricksium	Savannah Darter	1	Fluvial
Nocomis leptocephalus	Bluehead Chub	1	Fluvial
Notropis lutipinnis	Yellowfin Shiner	1	Fluvial
Ameiurus brunneus	Snail Bullhead	1	Fluvial
Notropis petersoni	Coastal Shiner	1	Fluvial
Opsopoeodus emiliae	Pugnose Minnow	1	Fluvial
Umbra pygmaea	Eastern Mudminnow	2	Eastern Mudminnow
Lepomis auritus	Redbreast Sunfish	3	Centrarchid
Lepomis macrochirus	Bluegill	3	Centrarchid
Lepomis punctatus	Spotted Sunfish	3	Centrarchid
Micropterus salmoides	Largemouth Bass	3	Centrarchid
Anguilla rostrata	American Eel	3	Centrarchid
Labidesthes sicculus	Brook Silverside	3	Centrarchid
Lepomis microlophus	Redear Sunfish	3	Centrarchid
Minytrema melanops	Spotted Sucker	3	Centrarchid
Perca flavescens	Yellow Perch	3	Centrarchid
Ameiurus platycephalus	Flat Bullhead	3	Centrarchid
Notemigonus crysoleucas	Golden Shiner	4	Non-Fluvial
Centrarchus macropterus	Flier	4	Non-Fluvial
Enneacanthus gloriosus	Bluespotted Sunfish	4	Non-Fluvial
Esox americanus	Redfin Pickerel	4	Non-Fluvial
Aphredoderus sayanus	Pirate Perch	4	Non-Fluvial
Lepomis gulosus	Warmouth	4	Non-Fluvial
Lepomis gulosus	Mud Sunfish	4	Non-Fluvial
Ameiurus natalis	Yellow Bullhead	4	Non-Fluvial
Erimyzon oblongus	Creek Chubsucker	4	Non-Fluvial
Lepomis marginatus	Dollar Sunfish	4	Non-Fluvial
Lepomis gibbosus	Pumpkinseed	4	Non-Fluvial
Erymyzon sucetta	Lake Chubsucker	4	Non-Fluvial
Enneacanthus obesus	Banded Sunfish	4	Non-Fluvial
Elassoma zonatum	Banded Pygmy Sunfish	4	Non-Fluvial
Amia calva	Bowfin	4	Non-Fluvial
Etheostoma fusiforme	Swamp Darter	4	Non-Fluvial

4. Variables Retained in Model

```
[1] "COMID"            "AreasqkmC"         "Dam_coC"         "Road_crC"
[5] "Epa_303dC"        "PcsC"              "TriC"            "ImpervC"
[9] "C_H20_01"         "C_GRASSHRUB_01"    "C_WETLAND_01"    "C_DECIDUOUS_01"
[13] "C_EVERGREEN_01"  "C_ROWCROP_01"      "LENGTHKM"        "SLOPE"
[17] "SoilpermC"       "CP_Ward4"
```

5. Model Call

```
* randomForest(formula = CP_Ward4 ~ ., data = CP_Groups, 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_DECIDUOUS_01 and SLOPE correlation= 0.71577984245247"
"i= 2 variables Dam_coC and Road_crC correlation= 0.659878048218154"
"i= 3 variables AreasqkmC and Road_crC correlation= 0.626436193530121"
"i= 4 variables Road_crC and TriC correlation= 0.620631006003958"
"i= 5 variables C_EVERGREEN_01 and C_ROWCROP_01 correlation= -0.611977618824569"
"i= 6 variables C_WETLAND_01 and C_DECIDUOUS_01 correlation= -0.572085002103961"
"i= 7 variables C_WETLAND_01 and SLOPE correlation= -0.569060135528113"
"i= 8 variables TriC and ImpervC correlation= 0.54907903262654"
"i= 9 variables SLOPE and SoilpermC correlation= 0.509703800420614"
"i= 10 variables Dam_coC and C_H20_01 correlation= 0.481882347775236"
```

7. List of Important Variables (Mean Decrease Accuracy)

Table 2. 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 (C_DECIDUOUS_01).

Predictor Variable	MeanDecreaseAccuracy	Weighting Factor
C_DECIDUOUS_01	60.596	1.000
SLOPE	59.540	0.983
C_WETLAND_01	56.300	0.929
Road_crC	41.995	0.693
AreasqkmC	25.232	0.416
C_H20_01	20.992	0.346
ImpervC	16.065	0.265
Dam_coC	15.891	0.262
SoilpermC	14.474	0.239
LENGTHKM	12.231	0.202
C_ROWCROP_01	11.754	0.194
C_GRASSHRUB_01	9.041	0.149
TriC	7.971	0.132
C_EVERGREEN_01	5.729	0.095
Epa_303dC	4.510	0.074
PcsC	1.192	0.020

8. Variable Importance Plot

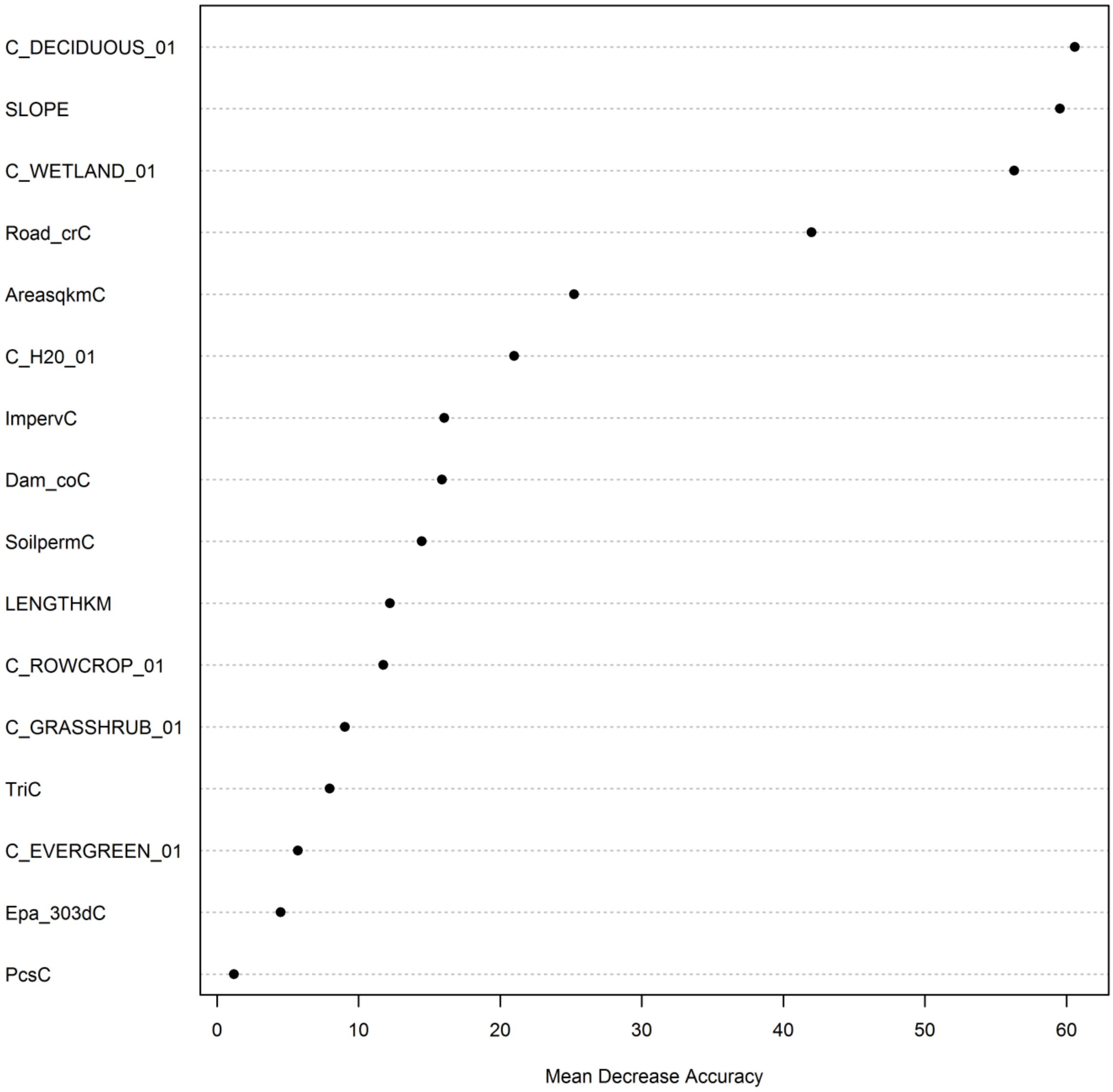
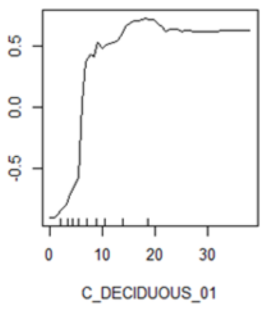


Figure 1. Top ranked variables from random forests classification for predicting South Carolina Coastal Plain Taxonomic Groups.

9. Partial Dependence Plots

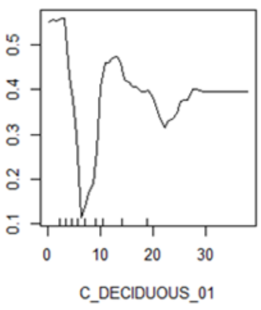
Partial dependence plots isolate and examine the relationships between top ranked predictors and Coastal Plain Taxonomic Groups (1=fluvial, 2=Eastern mudminnow, 3=centrarchid, 4=non-fluvial) while holding the effect of all other predictive variables constant.

Partial Dependence on C_DECIDUOUS_01



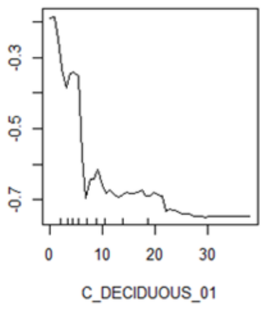
1

Partial Dependence on C_DECIDUOUS_01



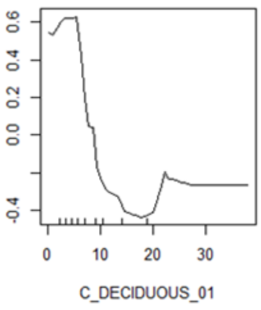
2

Partial Dependence on C_DECIDUOUS_01



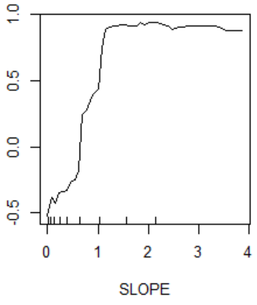
3

Partial Dependence on C_DECIDUOUS_01



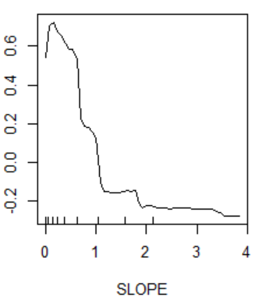
4

Partial Dependence on SLOPE



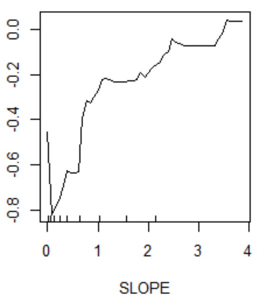
1

Partial Dependence on SLOPE



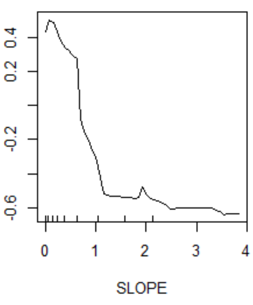
2

Partial Dependence on SLOPE

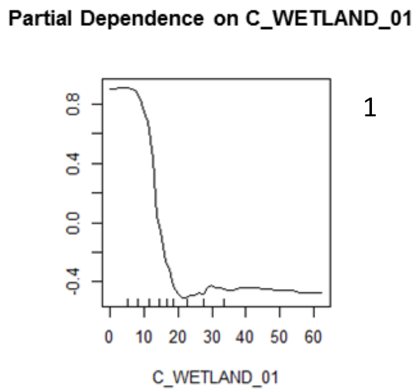


3

Partial Dependence on SLOPE

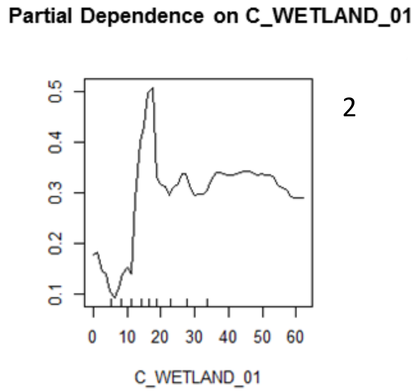


4



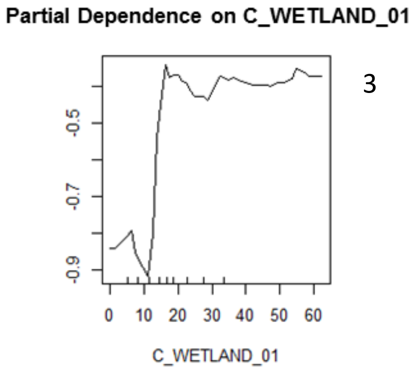
Partial Dependence on C_WETLAND_01

1



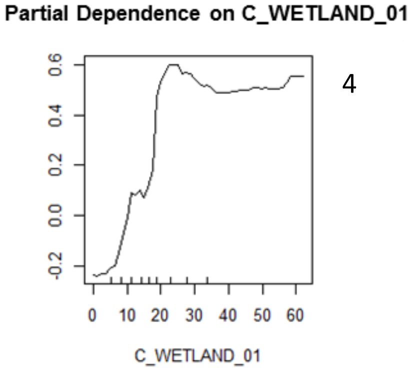
Partial Dependence on C_WETLAND_01

2



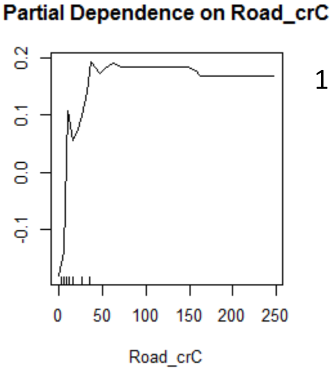
Partial Dependence on C_WETLAND_01

3



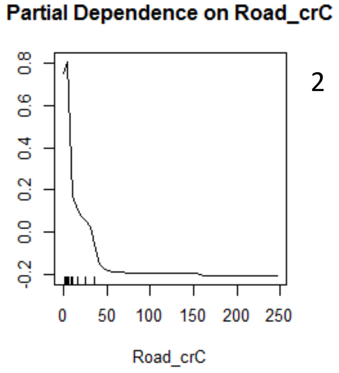
Partial Dependence on C_WETLAND_01

4



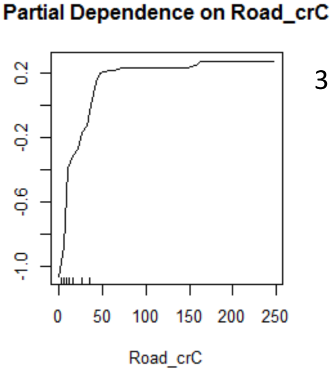
Partial Dependence on Road_crC

1



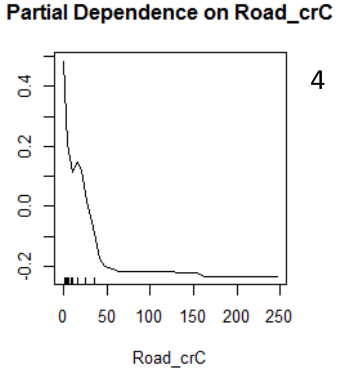
Partial Dependence on Road_crC

2



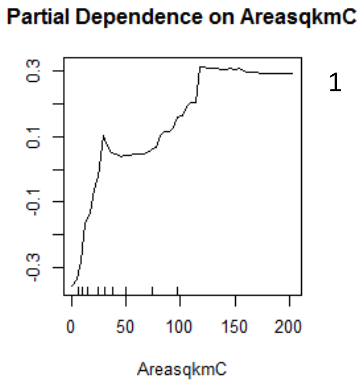
Partial Dependence on Road_crC

3

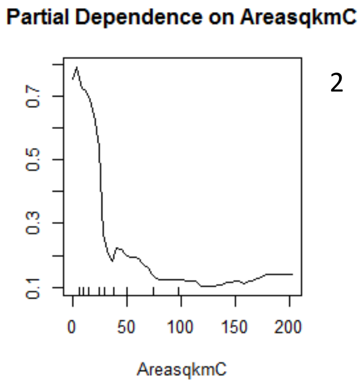


Partial Dependence on Road_crC

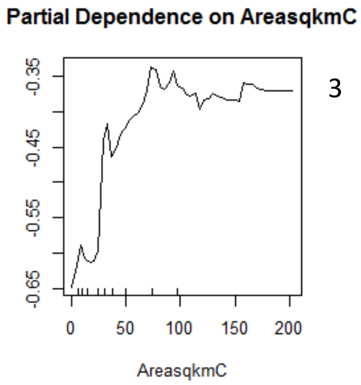
4


Partial Dependence on AreasqkmC
0.3
0.1
-0.1
-0.3
0
50
100
150
200
AreasqkmC

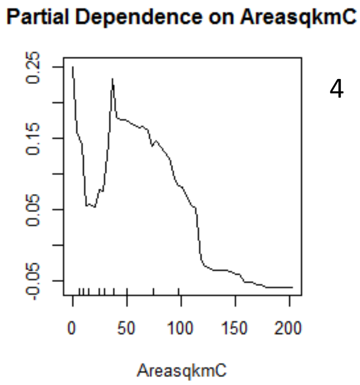
1


Partial Dependence on AreasqkmC
0.7
0.5
0.3
0.1
0
50
100
150
200
AreasqkmC

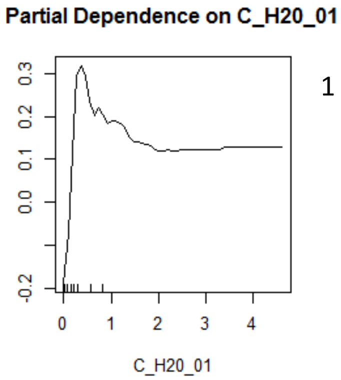
2


Partial Dependence on AreasqkmC
-0.35
-0.45
-0.55
-0.65
0
50
100
150
200
AreasqkmC

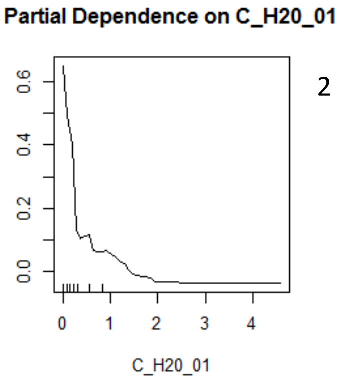
3


Partial Dependence on AreasqkmC
0.25
0.15
0.05
-0.05
0
50
100
150
200
AreasqkmC

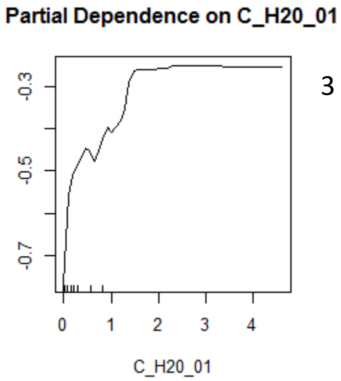
4


Partial Dependence on C_H20_01
0.3
0.2
0.1
0.0
-0.2
0
1
2
3
4
C_H20_01

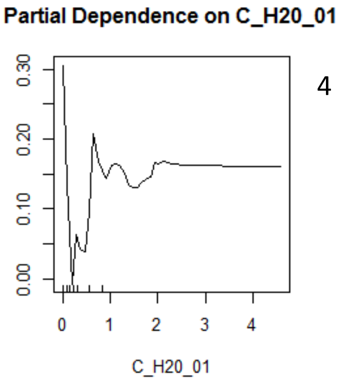
1


Partial Dependence on C_H20_01
0.6
0.4
0.2
0.0
0
1
2
3
4
C_H20_01

2

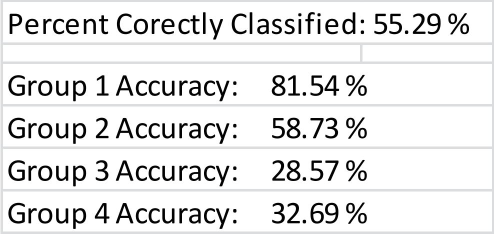

Partial Dependence on C_H20_01
-0.3
-0.5
-0.7
0
1
2
3
4
C_H20_01

3

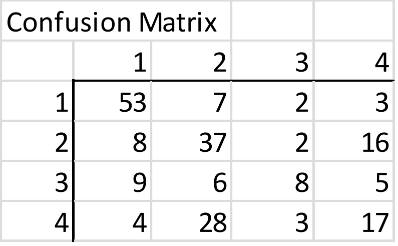

Partial Dependence on C_H20_01
0.30
0.20
0.10
0.00
0
1
2
3
4
C_H20_01

4

10. Error Estimate: PCC Result and Confusion Matrix



Percent Corectly Classified: 55.29 %	
Group 1 Accuracy:	81.54 %
Group 2 Accuracy:	58.73 %
Group 3 Accuracy:	28.57 %
Group 4 Accuracy:	32.69 %



Confusion Matrix				
	1	2	3	4
1	53	7	2	3
2	8	37	2	16
3	9	6	8	5
4	4	28	3	17

11. Error Estimate: Weighted K result

*w*K = 0. 37 (Fair Strength of Agreement)

12. Error Estimate: AUC Result

AUC = 0.70 (Model performance better than expected by chance alone)