Random forests

Turn taking data:

Switchboard corpus of conversations

Duration between turns

Speech rate

Are people laughing?

Sex of speaker

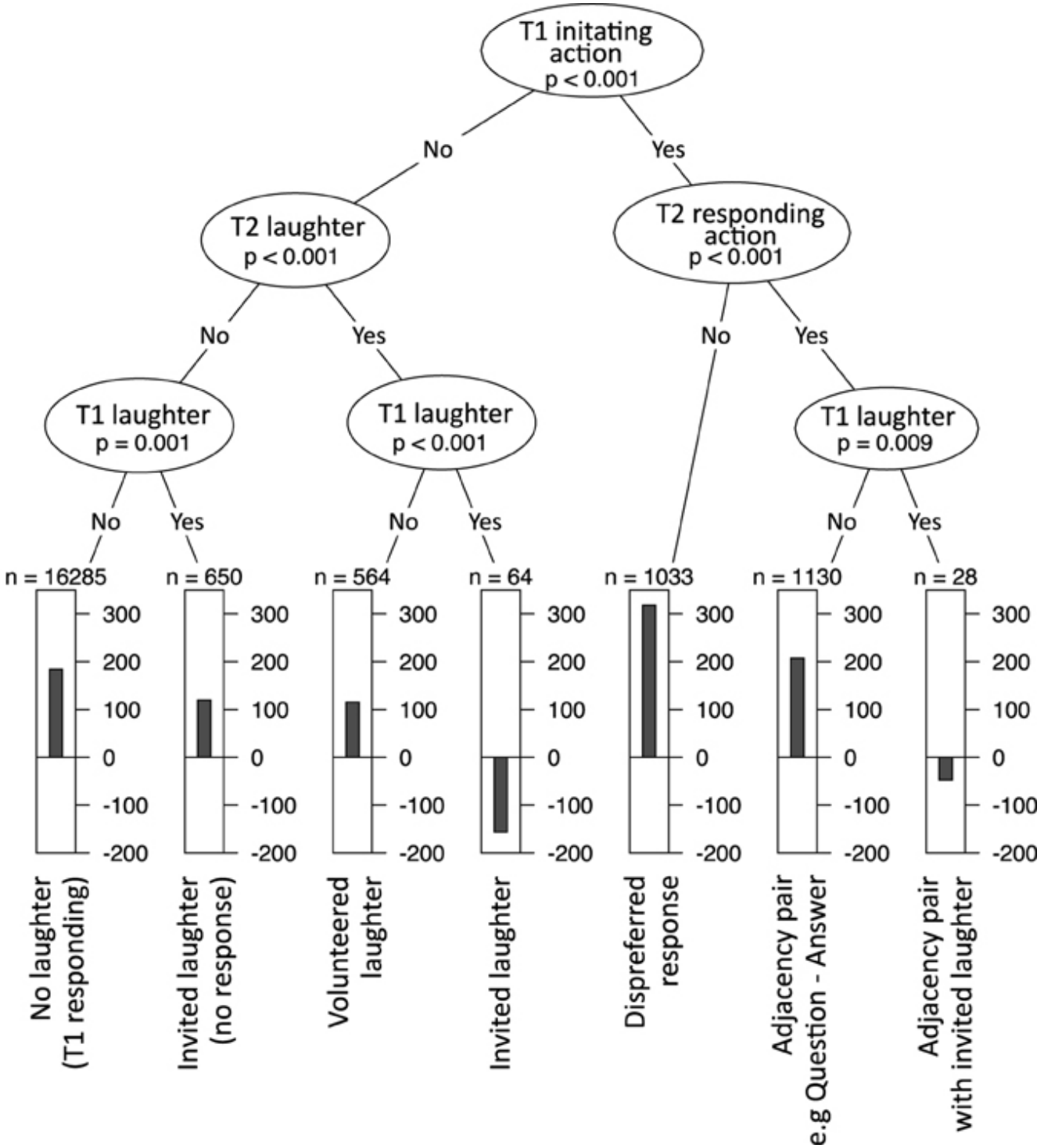
Speech act of turn:

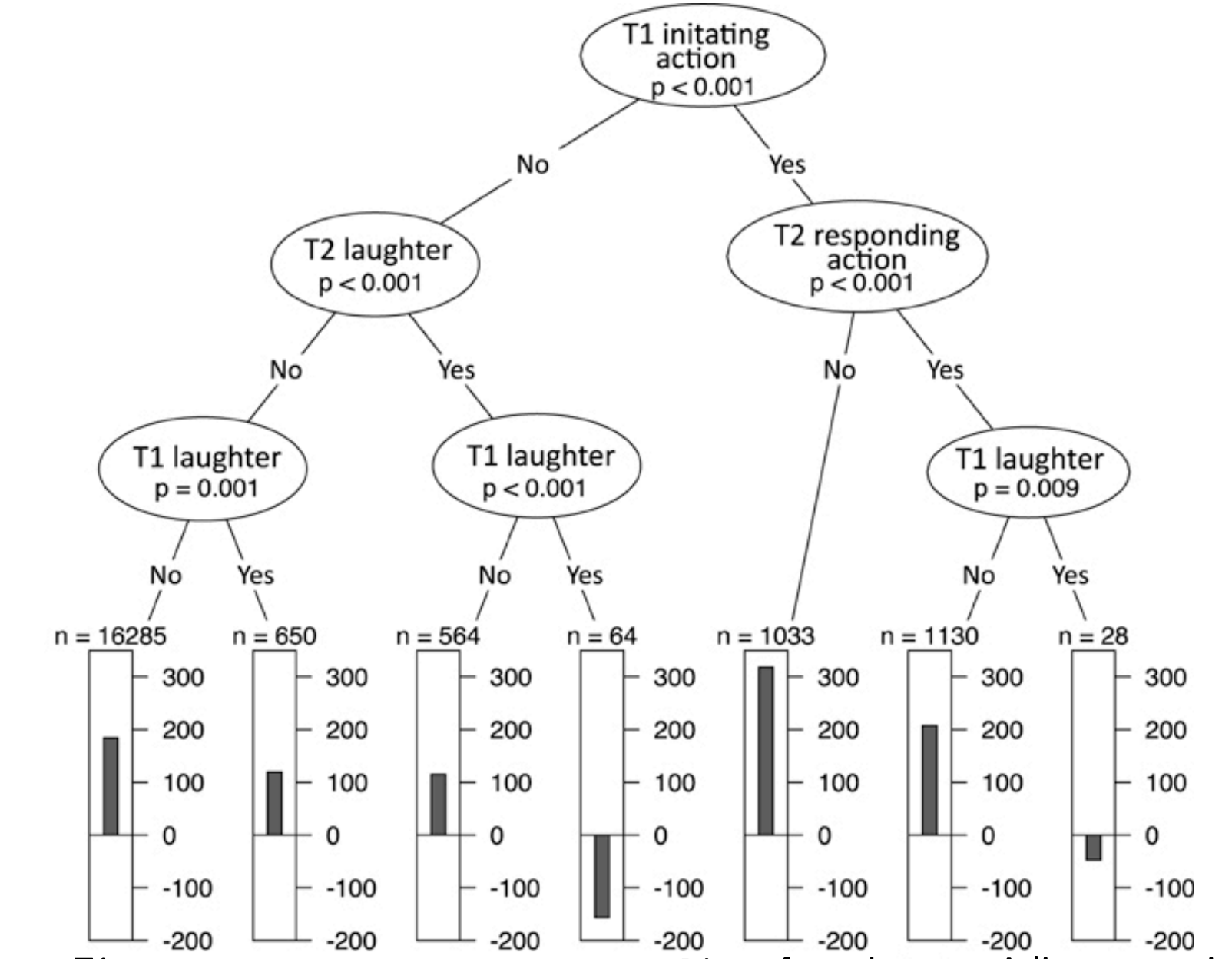
statement, question, backchannel etc.

Multicolinearity

	T1 Concreteness	T2 Concreteness	T1 Frequency	T2 Frequency	T1 Speech rate	T2 Speech rate	T1 Info density	T2 Info density	T1 Turn FTO	T2 Turn FTO	T1 Surprisal
T1 Concreteness		-0.23 **	0.31 **	-0.1 **	-0.36 **	0.15 **	-0.01	0.03 **	0.37 **	-0.21 **	0.01
T2 Concreteness			-0.07 **	0.33 **	0.14 **	-0.33 **	0.08 **	0.01	-0.27 **	0.39 **	0
T1 Frequency				-0.02 *	0.06 **	0.01	0.08 **	0.02 *	0.12 **	-0.08 **	0
T2 Frequency					0.04 **	0.07 **	0.04 **	0.11 **	-0.12 **	0.11 **	0
T1 Speech rate						-0.07 **	0.04 **	0	-0.21 **	0.09 **	0
T2 Speech rate							-0.02 *	0.05 **	0.11 **	-0.22 **	-0.01
T1 Info density								0.02 *	-0.29 **	0.02 *	0.03 **
T2 Info density									0.01	-0.26 **	0.02 *
T1 Turn FTO										-0.16 **	0
T2 Turn FTO											0
T1 Surprisal											

Table 2. The covariance between different variables in the data. Tests include pearson correlation, anova or chi-square, as appropriate. * = significant at 0.05, ** = significant at 0.001


T1 initating action
p < 0.001
No
Yes
T2 laughter
p < 0.001
T2 responding action
p < 0.001
No
Yes
No
Yes
T1 laughter
p = 0.001
T1 laughter
p < 0.001
T1 laughter
p = 0.009
No
Yes
No
Yes
No
Yes
n = 16285
n = 650
n = 564
n = 64
n = 1033
n = 1130
n = 28
300
200
100
0
-100
-200
No laughter (T1 responding)
Invited laughter (no response)
Volunteered laughter
Invited laughter
Dispreferred response
Adjacency pair e.g Question - Answer
Adjacency pair with invited laughter


T1 initating action
p < 0.001
No
Yes
T2 laughter
p < 0.001
T2 responding action
p < 0.001
No
Yes
No
Yes
T1 laughter
p = 0.001
T1 laughter
p < 0.001
T1 laughter
p = 0.009
No
Yes
No
Yes
No
Yes
n = 16285
n = 650
n = 564
n = 64
n = 1033
n = 1130
n = 28
300
200
100
0
-100
-200
T1 Responding
Invited Laughter (no response)
Volunteered Laughter
Invited Laughter
Dispreferred Response
Adjacency Pair
Adjacency pair with invited laughter

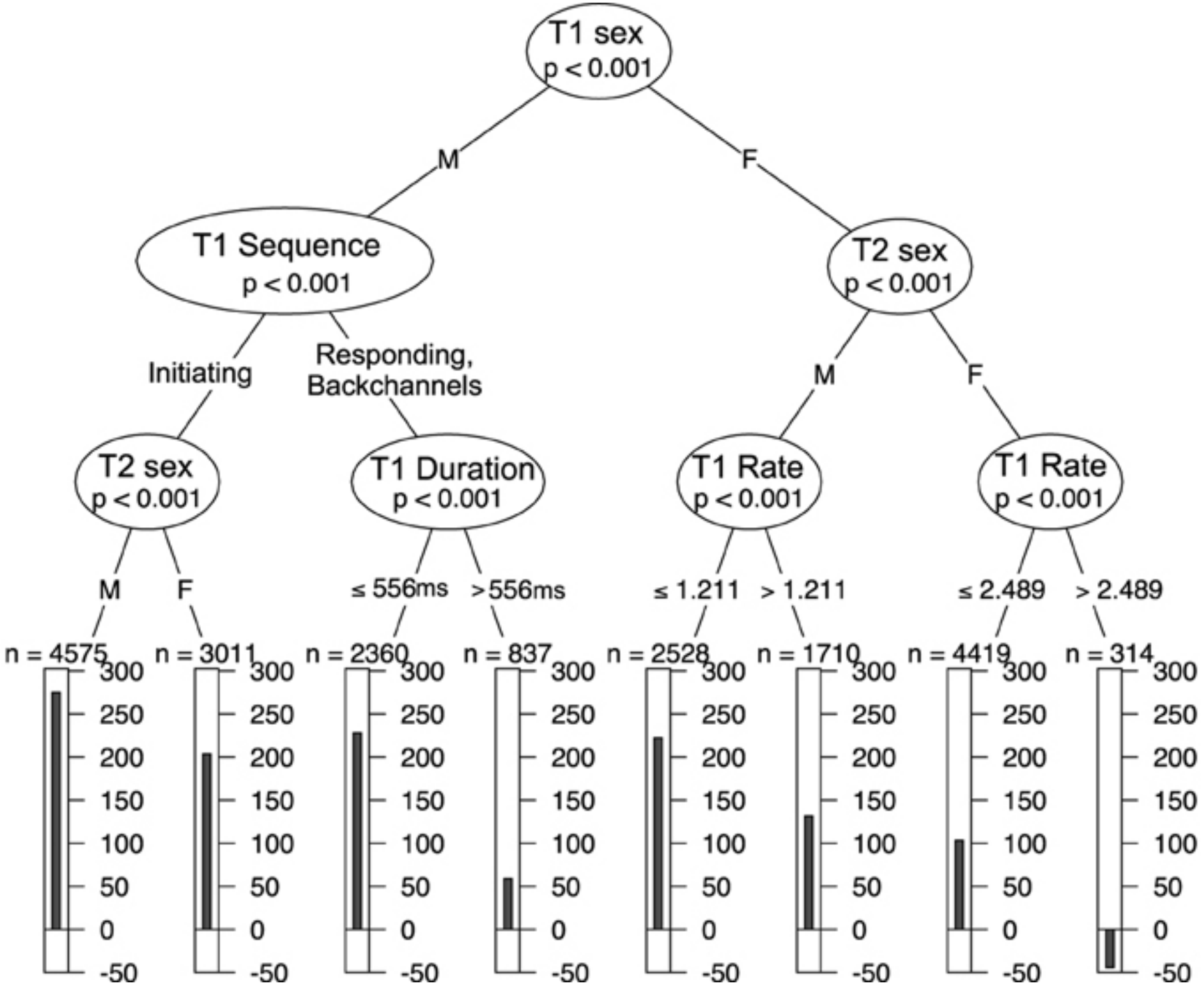
Building a tree

The strength of association between each predictor variable and FTO is determined by a statistical test of independence.

The variable with the strongest association is chosen as the first node in the tree.

The data is divided according to this variable into two sub-sets.

The process repeats recursively with each sub-set until all predictor variables are statistically independent from FTO in each leaf of the tree.


T1 sex
p < 0.001
M
F
T1 Sequence
p < 0.001
T2 sex
p < 0.001
Initiating
Responding, Backchannels
M
F
T2 sex
p < 0.001
T1 Duration
p < 0.001
T1 Rate
p < 0.001
T1 Rate
p < 0.001
M
F
≤ 556ms
>556ms
≤ 1.211
> 1.211
≤ 2.489
> 2.489
n = 4575
n = 3011
n = 2360
n = 837
n = 2528
n = 1710
n = 4419
n = 314
300
250
200
150
100
50
0
-50

Variable importance

How much influence a variable has over the fit of the real data to the predicted data

Problem: Single trees are sensitive

- Choice of first variable can change the order of the next ones

Solution:

Generate a ‘forest’ of trees from randomly selected sub-sets of variables.

Aggregate the variable importance

Importance measures

Standard mean decrease in classification accuracy when a variable is permuted (see Breiman, 2001).

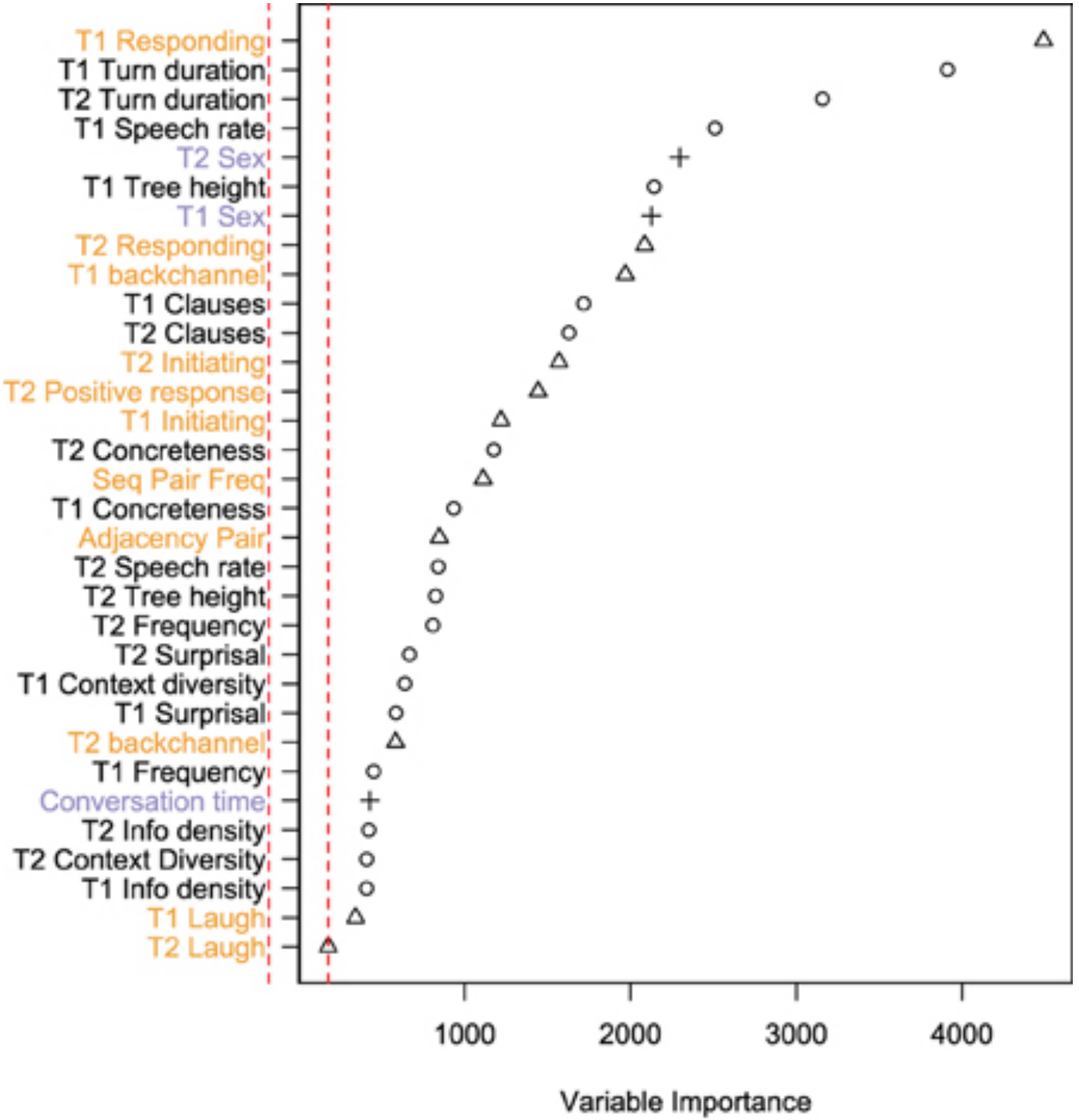
For each tree in the forest, the prediction error (mean squared error) is calculated by comparing the true values of FTO to the values predicted by the tree.

Permute the test variable, re-calculate prediction error.

The difference between the two errors gives a measure of how influential the variable is for prediction of FTO.

The difference in errors are calculated for all trees.

The importance measure is then the mean of these differences normalized by the standard deviation of the differences.


T1 Responding
T1 Turn duration
T2 Turn duration
T1 Speech rate
T2 Sex
T1 Tree height
T1 Sex
T2 Responding
T1 backchannel
T1 Clauses
T2 Clauses
T2 Initiating
T2 Positive response
T1 Initiating
T2 Concreteness
Seq Pair Freq
T1 Concreteness
Adjacency Pair
T2 Speech rate
T2 Tree height
T2 Frequency
T2 Surprisal
T1 Context diversity
T1 Surprisal
T2 backchannel
T1 Frequency
Conversation time
T2 Info density
T2 Context Diversity
T1 Info density
T1 Laugh
T2 Laugh
1000
2000
3000
4000
Variable Importance

Application to diversity

Greenberg diversity index

Population size

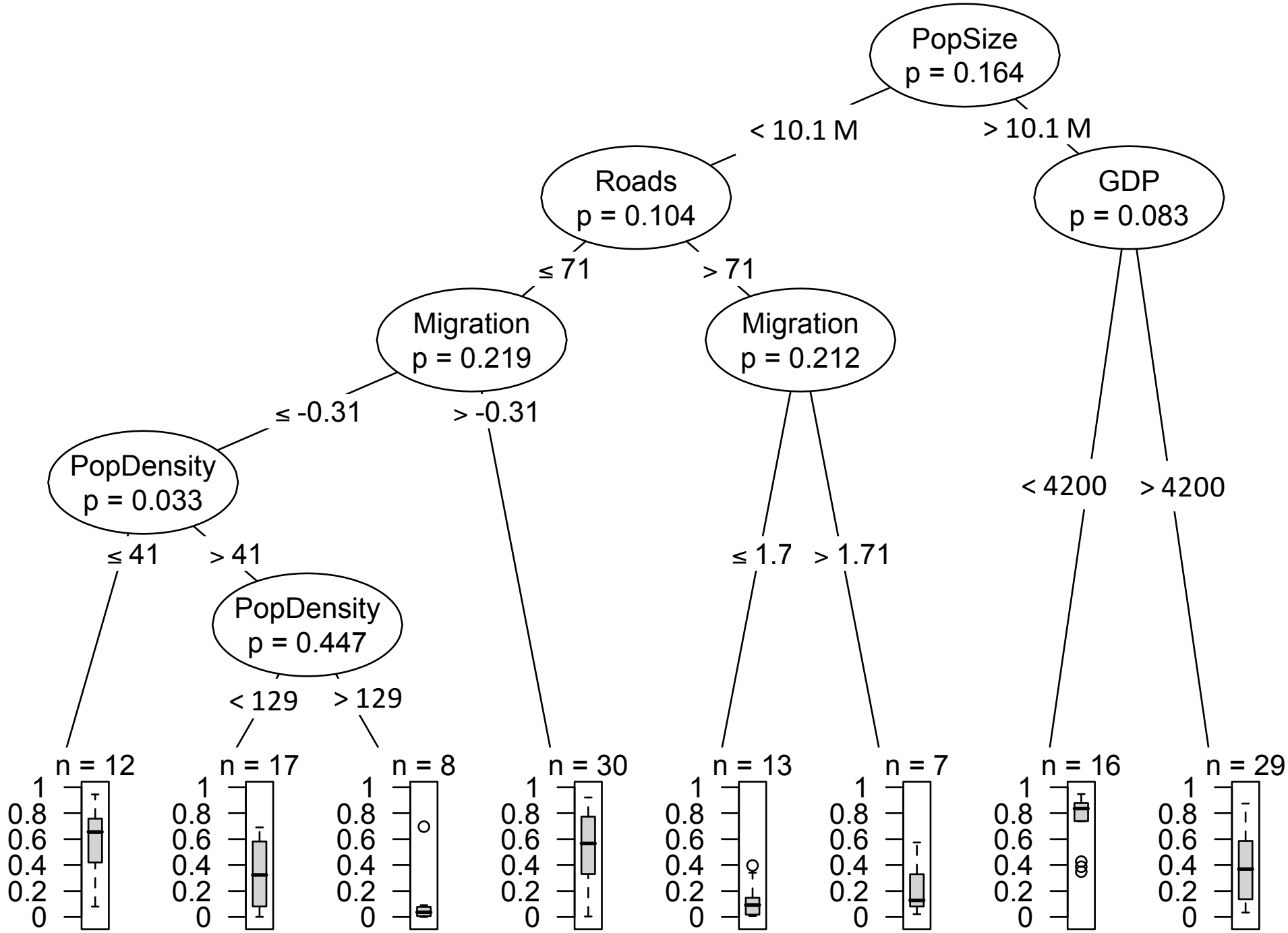
Population density

GDP

Migration rates

Km roads

132 countries


PopSize
p = 0.164
< 10.1 M
> 10.1 M
Roads
p = 0.104
GDP
p = 0.083
≤ 71
> 71
Migration
p = 0.219
Migration
p = 0.212
≤ -0.31
> -0.31
PopDensity
p = 0.033
< 4200
> 4200
≤ 41
> 41
≤ 1.7
> 1.71
PopDensity
p = 0.447
< 129
> 129
n = 12
n = 17
n = 8
n = 30
n = 13
n = 7
n = 16
n = 29
1
0.8
0.6
0.4
0.2
0

Decision trees

```
# make a single tree:
ct = ctree(dur~.,data=dx)
# make a tree with just 3 branches:
ct2 = ctree(dur~.,data=dx, controls=ctree_control(maxdepth=3))
# plot it (this will make a decision tree, like in our paper)
plot(ct)
#or a bit more tidy:
plot(ct,inner_panel=node_inner(ct,id=F),terminal_panel=node_barplot)


# This can be pretty difficult to read if you have lots of factors, you can look at the details like this:
ctreex@tree
```

Random Forests

```
# build a random forest:
# see ?cforest_unbiased for list of parameters
data.controls <- cforest_unbiased(ntree=1000, mtry=5)
# run the random forest
d.cforest <- cforest(dur~.,data=dx,
controls=data.controls)
# work out ranking of variables
d.varimp <- varimp(d.cforest, conditional=T)
print(sort(d.varimp))
```

Prediction

```
# make a prediction based on new data:

predictions = predict(d.cforest,newDatainDataFrame)

# (if you run this without new data, you'll just try to
predict the values that the model was trained on.  This
can be insightful, too.)


# compare predictions to actual values (a kind of
measure of model fit)

cor.test(predictions, dx$dur)

#  You can also run a forest with a random 90% of the
data, then predict the other 10% and see how well the
model fits
```

References

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organisation on the timing of turn taking: A corpus study. Frontiers in Psychology, 6: 509. doi:10.3389/fpsyg.2015.00509.

Bürki, A., Alario, F. X., and Frauenfelder, U. H. (2011). Lexical representation of phonological variants: evidence from pseudohomophone effects in different regiolects. *J. Mem. Lang*. 64, 424–442. doi: 10.1016/j.jml.2011.01.002

Tagliamonte, S. A., and Baayen, R. H. (2012). Models, forests, and trees of york english: was/were variation as a case study for statistical practice. *Lang. Variat. Change* 24, 135–178. doi: 10.1017/S0954394512000129

Plug, L., and Carter, P. (2014). Timing and tempo in spontaneous phonological error repair. *J. Phonet*. 45, 52–63. doi: 10.1016/j.wocn.2014.03.007

Sadat, J., Martin, C. D., Costa, A., and Alario, F. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cogn. Psychol*. 68, 33–58. doi: 10.1016/j.cogpsych.2013.10.001