Class 13a: Random Forests, for Model (and Predictor) Selection

1 WHICH ASPECTS OF OBSERVED DATA ARE IMPORTANT? WHAT PREDICTORS DO WE INCLUDE?

[1] Substantive reasons! (Gelman and Hill 2007: 69 have a nice discussion of some criteria)
[2] Via various methods of predictor selection

- Strategies for regression/parametric models
  - step-up/step-down model building
  - log likelihood/AIC tests
  - AIC-based model averaging/selection (e.g., Burnham and Anderson 2004; in linguistics: Kuperman and Bresnan 2012)

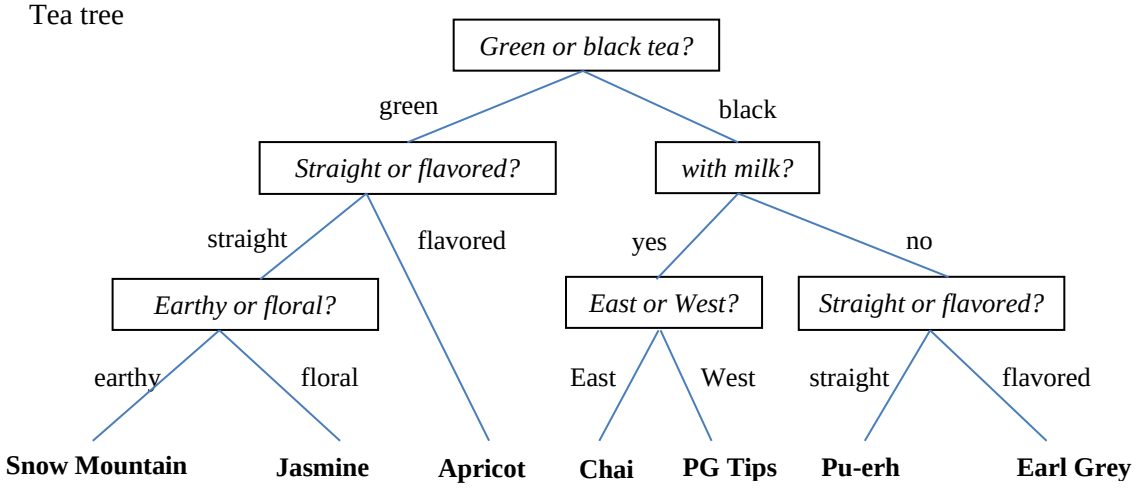
- Notable issues in model selection strategies for regression/parametric models
  - step-up/step-down model building misses interactions
  - higher-order interactions aren't easy to capture (or interpret)
  - assumption of a certain probability distribution in the data (e.g., normal, Poisson, etc.)
  - data sparseness: severe limitations when it comes to small *n*, large *p* power (e.g., general rule of thumb for logistic regression is $(\min(n_1, n_0)/10-1)$.)
  - multicollinearity and overfitting

- Outline
  - Classification (and regression) trees
  - Ensemble methods: Random forests and variable importance
  - An illustration: multicollinearity
  - Advantages, comparisons, and limitations

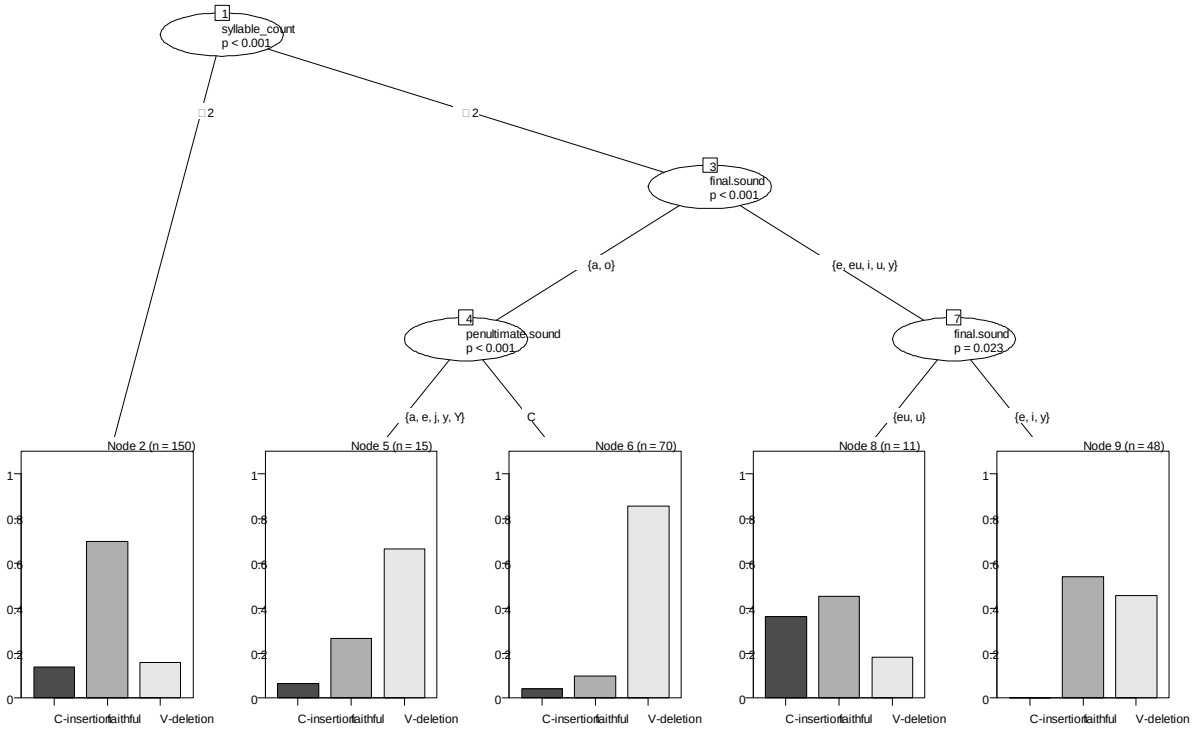
2 CLASSIFICATION (AND REGRESSION) TREES (CART)

- A non-parametric method of sorting data based on predictor variables.
- An illustration: imagine you have a guest and want to serve them tea from your (extensive) tea collection. How do you help the guest narrow down what they'd like to try?
  - Method 1. *So, what do you want?* ← highly inefficient
  - Method 2. A decision tree based on the most predictive variables, trying to eliminate as many tea choices as possible in each split. ← more efficient

(1) Tea tree



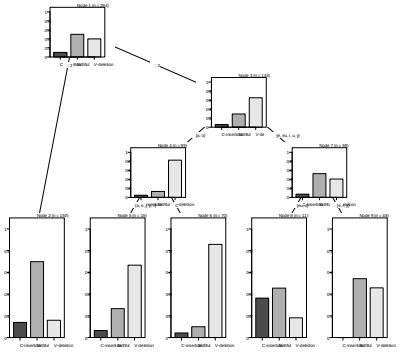
(2) Conditional inference tree for French *–esque* attachment




```
> library(party)
> set.seed(47)
> d.ctree <- ctree(outcome ~ penultimate.sound + final.sound + syllable_count,
  data = data)
> plot(d.ctree)
```

- Upshot: CART can handle multinomial dependent predictors.
- How the tree works:
  - Splits, or branches of the tree, are made by trying to maximize the purity of the data partitions (= "impurity reduction").
  - Predictor that is most strongly associated with the data partitions (i.e., makes the purist split with the smallest *p* value) is chosen as the basis of the split.

(3) Impurity reduction of ctree in (2)



```
> plot(d.ctree, inner_panel = node_barplot)
```

- Discuss: What does the tree in (2) tell us? How does it compare to the multinomial logistic regression results?

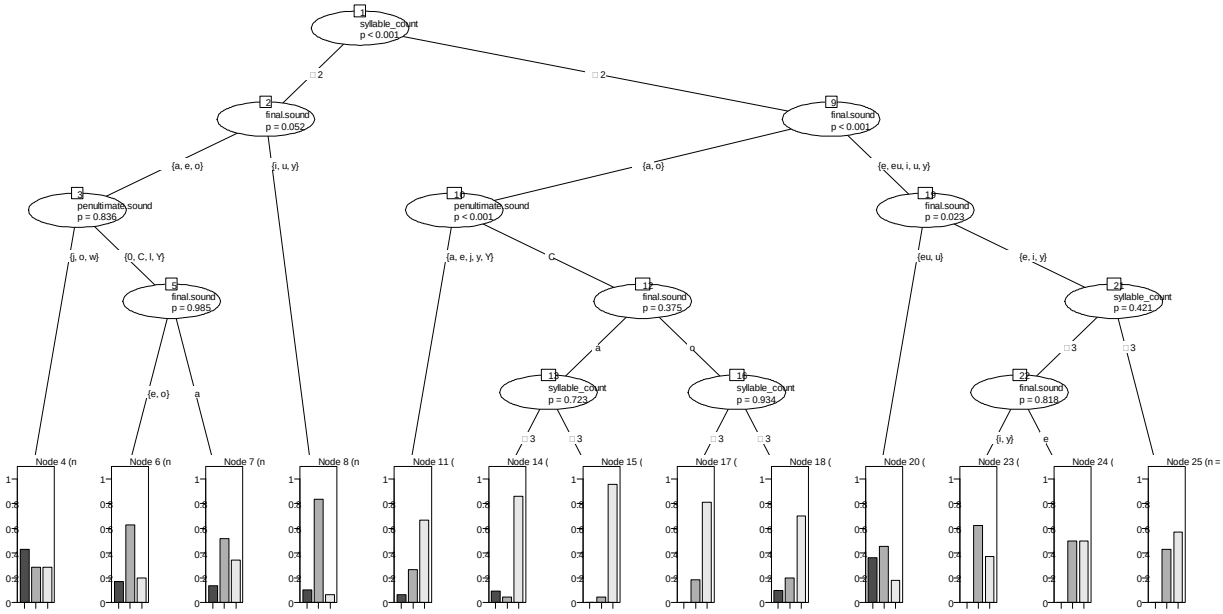
(4) % correct of the ctree model: 70.07%

```
> ctree_predict <- predict(d.ctree)
> sum(data$outcome==ctree_predict)/nrow(data)
```

2.1 AVOIDING OVERFITTING IN CARTS

- Like regression models, CARTs can also be prone to over-fitting if you're not careful.
- Traditionally, this is done post-hoc by "pruning" the tree that you produce:

(5) Unpruned/unlimited tree



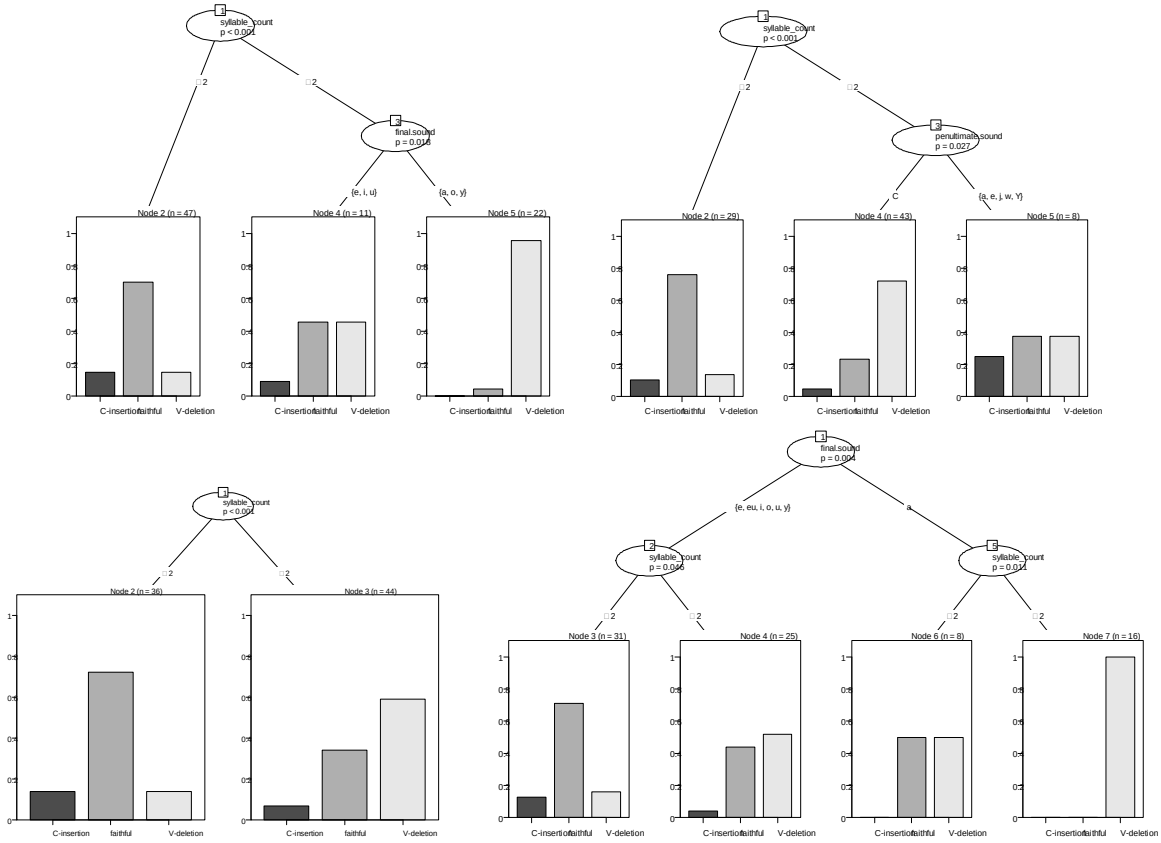

```
> d.ctree.unpruned <- ctree(outcome ~ penultimate.sound + final.sound +
  syllable_count, controls = ctree_control(mincriterion=0.01), data=data)
> plot(d.ctree_unpruned)
```

- But! the `party` package uses a different methodology to prevent overfitting = "conditional" CARTs
  - Splits in the tree are only implemented when a global null hypothesis cannot be rejected, as determined by a chosen *p*-value (by default, *p*=0.05 (`mincriterion=0.95`); see Hothorn et al. 2006 for the math).

2.2 THE PROBLEM WITH SINGLE CARTS

- Tree structure is subject to high variability depending on data (i.e., depending on the sample of learning data) because subsequent splits further down the tree depend on previous ones.

(6) Variability of trees illustrated on random samples of *–esque* data




```
> data.s1 <- data[sample(1:nrow(data), 80, replace=FALSE),]
> ds1.ctree <- ctree(outcome ~ penultimate.sound + final.sound + syllable_count,
  controls = ctree_control(mtry=2), data=data.s1)
> plot(ds1.ctree)
```

3 CONDITIONAL RANDOM FORESTS

- *Conditional Random Forests* – solution to problem laid out in §2.2 (Strobl et al. 2009a, 2009b; a.o.; in linguistics: Tagliamonte and Baayen 2012; a.o.)
  - Instead of using a single tree, use ensemble methods in a forest of trees.
  - Capitalize on the diversity of CART trees.

- Conditional random forests use
  - random subsamples of data used to build each tree
  - random restricted set of predictor variables in each tree split
- = diverse trees: variables have a greater chance of being included in the model when a stronger competitor is not (cf. regression models)

(7) Setting up and running a random forest model

STEP 1. Set up model control parameters.

```
> data.controls <- cforest_unbiased(ntree=1000, mtry=2)
```

- `ntree` = number of trees in the forest
  Use a suitably large number of trees to produce more robust and stable results. For smaller datasets with a large number of predictor variables, you will need more trees.
- `mtry` = number of randomly-selected variables used in each split
  Suggested value = $\sqrt{p}$ (square root of the total number of predictor variables)
- IMPORTANT: make sure to report model parameters. Because random forests are 'random,' model parameters are necessary for result replication.

STEP 2. Make sure response variable is a factor. If not, make it one.

```
> is.factor(data$outcome)     # if FALSE, use the next line of code.
> data$outcome <- as.factor(data$outcome)
```

STEP 3. Set the random seed, then run the random forest, using preselected model controls.

```
> set.seed(47)
> d.cforest <- cforest(outcome ~ penultimate.sound + final.sound + syllable_count,
  data = data, controls=data.controls)
```

- IMPORTANT: Random forests are 'random.' Always run the model at least 2x with different seeds to ensure the robustness and stability of the model. If the model results fluctuate (e.g., inconsistent variable importance rankings), increase the number of trees in the forest.

(8) % correct of the random forest model: 70.75%
(cf. ctree predictions in (4) and regression model results)

```
> ctree_predict <- predict(d.ctree)
> sum(data$outcome==ctree_predict)/nrow(data)
```

- Advantages of a forest of diverse trees:
  - detects contributions and behavior of predictor variables otherwise masked by competitors.
  - useful for small *n*, large *p* problems
  - greater accuracy than simple/mixed effect regression models (Strobl et al. 2008; in linguistics: Shih and Grafmiller 2011; Tagliamonte and Baayen 2012)

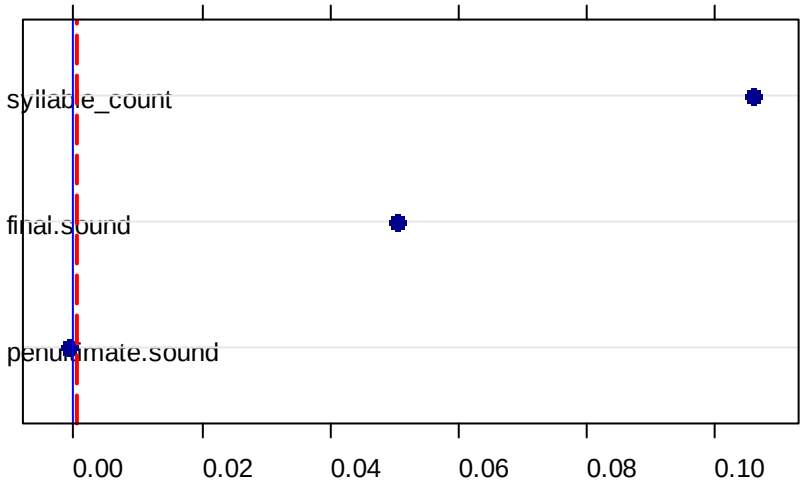
- Random forests are built on unpruned, large trees. Are they prone to overfitting? Breiman (2001) says no due to the Law of Large Numbers, but it's still debated in the current literature.

3.1 VARIABLE IMPORTANCE

- Unlike CARTs, it's not easy to read effects off a random forest model because a single predictor can show up in many different (non-nested) trees with different covariates.
- But, random forests and ensemble methods make up for this shortcoming in being a great tool for model and predictor selection because we gain more information about how each variable behaves with respect to the observed data.

- *Permutation Variable Importance*[1]
  - In a random forest model, randomly shuffle values of a predictor variable to break the association between response and predictor values.
  - Calculate the difference in model accuracy before and after shuffling. (= a standardized/scaled % correct, averaged across all trees in the forest; see Strobl et al. 2009b: 336 for the math).
  - If the predictor never had any meaningful relationship with the response, shuffling its values will produce very little change in model accuracy. On the other hand, if predictor was strongly associated with the response, permutation should create a large drop in accuracy.
- Permutation variable importance covers the individual impact of each predictor in the random forest model, including its impact in interactions (which is much harder to assess in regression models).

(9) Variable importance for *–esque* forest model




```
> d.varimp <- varimp(d.cforest, conditional=TRUE)
  # Note that this step could take a long time, depending on the size of the
  forest and computational power.
> d.varimp      # prints the results to screen

  ### to plot
> dotplot(sort(d.varimp), panel=function (x,y){
> panel.dotplot(x, y, col='darkblue', pch=16, cex=1.1)
> panel.abline(v=abs(min(d.varimp)), col = 'red', lty='longdash', lwd=2)
> panel.abline(v=0, col='blue)
> }
> )
```

- Interpreting variable importance
  - Interpret variable importance only as a ranking, not as absolute values (because random forests are random!).
  - Variables are informative and important if importance value is above the absolute value of the lowest negative-scoring variable (or zero, if there's no negative variable).
  - Irrelevant variables will vary randomly approaching zero.
- Discuss: compare variable importance results in *–esque* with results in regression models.

[1] Previous work on random forests have used a number of other types of variable importance measures, most notably the Gini score, based on impurity reduction, but Strobl et al. 2009a show that these tests are biased towards correlated predictors. So don't be fooled into working with the R package entitled `randomforest`.

4 MULTICOLLINEARITY AND THE POWER OF RANDOM FORESTS

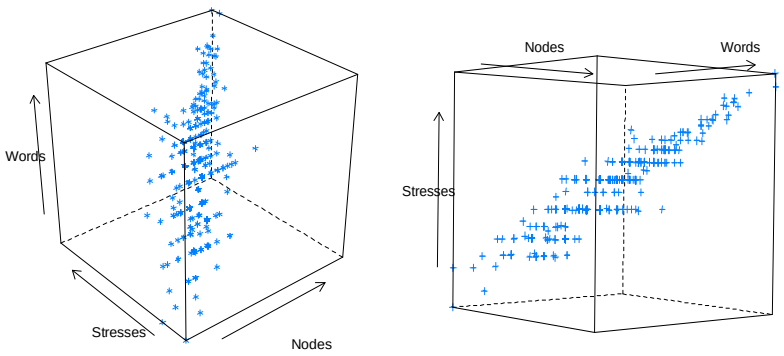
- If random forest results coincide with those of simpler, regression-based models, then regression-based models may be sufficient for many of your modeling needs.

- One thorny problem in regression modeling: *multicollinearity*
  - (Multi-)collinearity arises when two or more predictors are highly correlated, either because they measure the same thing or are attributable to a shared underlying explanation.
  - (Stephanie will draw a graphic on the board illustrating the effect of collinearity.)
  - Multicollinearity doesn't necessarily affect overall model prediction, but it does increase potential overfitting *and*
  - If we want to make inferences and interpret the explanatory power of our predictor variables, we need to be able to tease them apart.

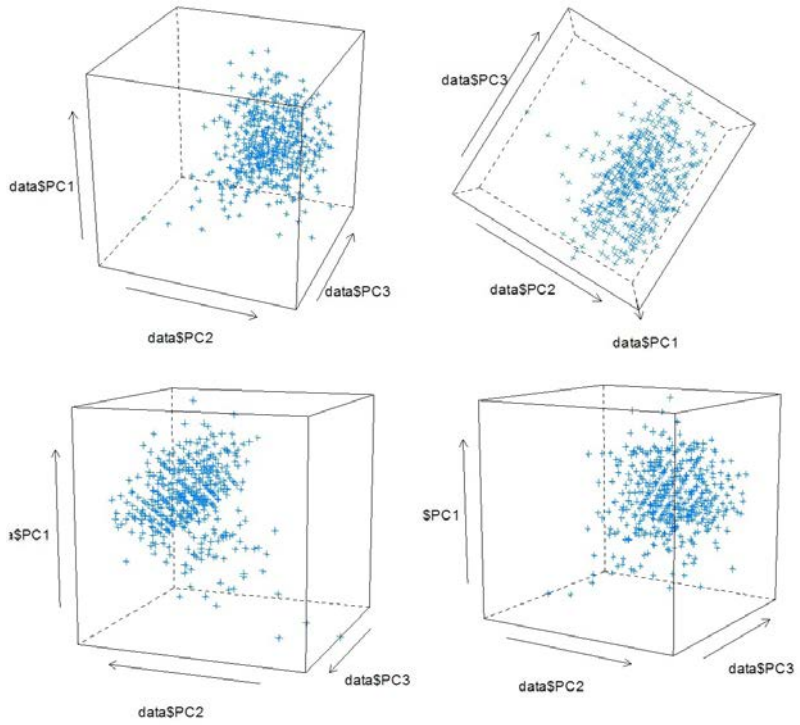
- An illustrative linguistics example: measures of end weight (based on Grafmiller and Shih 2011)
  - Principle of End Weight: "Phrases are presented in order of increasing weight." (Wasow 2002: 3; following Behagel 1909; Quirk et al. 1985; a.o.)
    e.g., *peas and carrots* > *carrots and peas*
    e.g., *the attitude of people who are really into classical music and feel that if it's not seventy-five years old, it hasn't stood the test of time* > ?*people who are really into classical music and feel that if it's not seventy-five years old, it hasn't stood the test of time's attitude*
  - What is "weight"?
    syntactic complexity (Hawkins 1994)
    processing load (Gibson 1998, 2000; Temperley 2007)
    phonological complexity (Selkirk 1984; Zec and Inkelas 1990; Anttila et al. 2010)
    phonological weight (McDonald et al. 1993; Benor and Levy 2006; a.o.)
    words (Wasow 2002; Szmrecsányi 2004; a.o.)
  - Problem: end weight measures are all highly correlated. Why?

- Data: genitive construction choice in English. (*the car's wheel* ~ *the wheel of the car*; data from Shih et al., to appear)
  - parallels some of the discussion last week on zero variants: a fix for dispreferred structure = using another construction altogether (e.g., dispreferred phonological structure = *the chur**ch's** bell* < *the bell of the church*)

(10) Correlated predictors overlap in narrow spread in 3d space



(11) Uncorrelated predictors should have a wider spread in 3d space



- Multicollinearity in regression models can cause unstable results.

(12) Logistic regression model results for English genitive construction choice

Factor	Estimate	Std. Error	t value	Pr (>|t|)	
Intercept	2.439	0.2111	11.56	<0.0001	***
Possessor animacy = inanim	-3.862	0.1998	-19.33	<0.0001	***
Rhythm	-0.189	0.1755	-1.08	0.2806	
Possessor sibilant = Y	=1.317	0.2861	-4.60	<0.0001	***
Weight = words	-0.825	0.4111	-2.01	0.0447	.
Weight = stresses	0.494	0.4319	1.14	0.2523	
Weight = syllables	-0.011	0.1673	-0.06	0.9490	
Weight = referents	-0.417	0.172	-2.43	0.0153	.
Weight = syn nodes	-1.272	0.385	-3.30	0.0010	*
Weight = content words	0.8139	0.4237	1.92	0.0547	

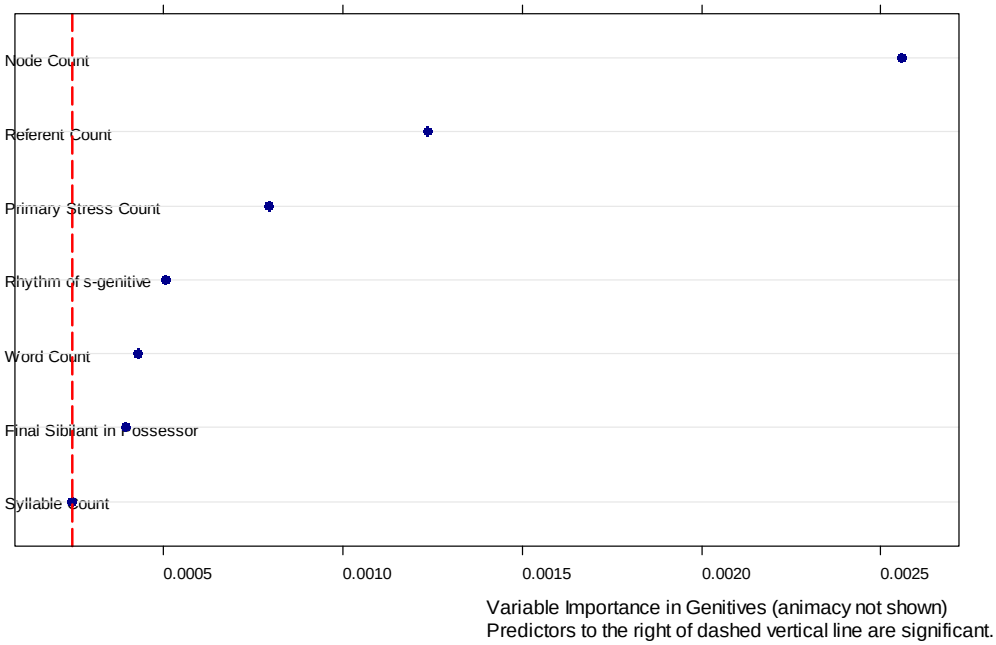
. significant at $p < 0.05$, * significant at $p < 0.01$, ** significant at $p < 0.001$, *** significant at $p < 0.0001$

- Unexpected coefficient sign errors due to collinearity. In individual models, both highlighted predictors show negative coefficient values.

4.1 RANDOM FORESTS FOR DEALING WITH COLLINEARITY

- A random forest model for English genitive construction choice
  - `ntree` = 2000, `mtry` = 3
  - Model stability verified on >2 random seeds.

(13) Permutation variable importance for genitives forest model



- Ranking of variable importance demonstrates, amongst weight measures: Syntactic nodes > Referents > Primary stresses > Words > Syllables.
- These results have been replicated in a AIC-based model averaging/selection approach.

(14) Accuracy of random forest model vs. regression model (n = 1123)
Forest: `C = 0.9441701, Dxy = 0.8883402`
Regression: `C = 0.898, Dxy = 0.796`

5 DISCUSSION

- Advantages of random forests
  - handles small n, large p problems
  - deals well with correlation and high-order interactions
  - eliminates order effects found in single CARTs
  - shown to be more accurate than CARTS or parametric regression models
- Drawbacks of random forests
  - long computing time
  - difficult to see size and direction of main effects
- Use random forests when…
  - working with highly-correlated data, data with many interactions, and/or datasets with small n. (Though note that random forests ≠ magic forests, and you still need reasonably sufficient n and well-motivated p).
  - exploring the independent effect of predictor variables (for model selection or inference).
  - you have all the time in the world to wait for `varimp` results.
- As always, choose the methodology that is best for the question at hand, but beware its weaknesses and pitfalls. Consistency of results across statistical methods and data can also be an indicator of robustness.

6 SELECTED REFERENCES

Anttila, Arto; Matthew Adams; and Michael Speriosu. 2010. The role of prosody in the English dative alternation. Language and Cognitive Processes. 25(7–9): 946–981.

Behagel, O. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. Indogermanische Forschungen. 25: 110-142.

Benor, Sarah Bunin and Roger Levy. 2006. The Chicken of the Egg? A Probabilistic Analysis of English Binomials. Language. 82(2): 233-278.

Breiman, Leo. 2001. Random Forests. Machine Learning. 45(1): 5–32.

Burnham, Kenneth P. and David R. Anderson. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological methods and research. 33(2): 261–304.

Gelman, Andrew and Jennifer Hill. 2007. Data analysis using regression and multilevel/hierarchical models. New York, NY: Cambridge University Press.

Gibson, Edward. 1998. Linguistic Complexity: locality of syntactic dependencies. Cognition. 68: 1-76.

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. in Y. Miyashita; A. Marantz; and W. O'Neil (ed). Image, Language, Brain. Cambridge, MA: MIT Press. 95-126.

Grafmiller, Jason and Stephanie Shih. 2011. New approaches to end weight. Paper presented at Variation and Typology: New trends in Syntactic Research. 25–27 Aug 2011. Helsinki, Finland. <http://stanford.edu/~stephsus/GrafmillerShihHelsinki2011.pdf>

Hawkins, John A. 1994. A Performance Theory of Order and Constituency. Cambridge: Cambridge University Press.

Hothorn, Torsten; Kurt Hornik; and Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional framework. Journal of Computational and Graphical Statistics. 15(3): 651–674.

Hothorn, Torsten; Kurt Hornik; Carolin Strobl; and Achim Zeileis. 2013. Package 'party'. R package version 1.0-6. <http://cran.r-project.org/package=party>

Kuperman, Victor and Joan Bresnan. 2012. The effects of construction probability on word durations during spontaneous incremental sentence production. Journal of Memory and Language. 66: 588–611.

McDonald, Janet L.; Kathryn Bock; and Michael H. Kelly. 1993. Word and World Order: Semantic, Phonological, and Metrical Determinants of Serial Position. Cognitive Psychology. 25: 188-230.

Quirk, Randolph; Sidney Greenbaum; Geoffrey Leech; and Jan Svartvik. 1985. A Comprehensive Grammar of the English Language. London and New York: Longman.

Selkirk, Elisabeth O. 1984. Phonology and Syntax: the Relation between Sound and Structure. Cambridge, MA: MIT Press.

Shih, Stephanie S. 2011. Random forests for classification trees and categorical dependent variables: an informal quick start R guide. <http://www.stanford.edu/~stephsus/R-randomforest-guide.pdf>

Shih, Stephanie; Jason Grafmiller; Richard Futrell; and Joan Bresnan. to appear. Rhythm's role in predicting genitive alternation choice in spoken English. In Ralf Vogel and Ruben van de Vijver (eds). Rhythm in phonetics, grammar, and cognition. <http://stanford.edu/~stephsus/DGfS-Shihetal.pdf>

Strobl, Carolin; Anne-Laure Boulesteix; Thomas Kneib; Thomas Augustin; and Achim Zeileis. 2008. Conditional variable importance for random forests. BMC Bioinformatics. 9: 307–317.

Strobl, Carolin; Torsten Hothorn; and Achim Zeileis. 2009a. Party on! A new, conditional variable importance measure for random forests available in party package. The R Journal. 1(2): 14-17.

Strobl, Carolin; James Malley; and Gerhard Tutz. 2009b. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods. 14(4): 323–348.

Szmrecsányi, Benedikt. 2004. On operationalizing syntactic complexity. Journées internationals d'Analyse statistique des Données Textualles. 7: 1031-1038.

Tagliamonte, Sali A. and R. Harald Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice.

Temperley, David. 2006. Minimization of dependency length in written English. Cognition. 105: 300-333.

Wasow, Tom. 2002. Postverbal Behavior. Stanford, CA: CSLI Publications.

Zec, Draga and Sharon Inkelas. 1990. Prosodically Constrained Syntax. in Sharon Inkelas and Draga Zec (eds). The Phonology-Syntax Connection. Stanford, CA: Center for the Study of Language and Information.

Contact: Stephanie S Shih | stephsus@stanford.edu

Departments of Linguistics, Stanford University & UC Berkeley