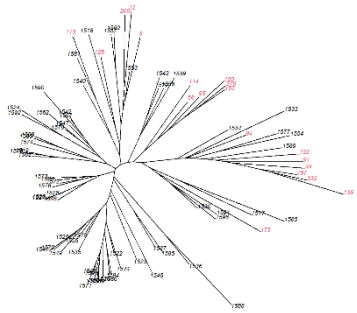
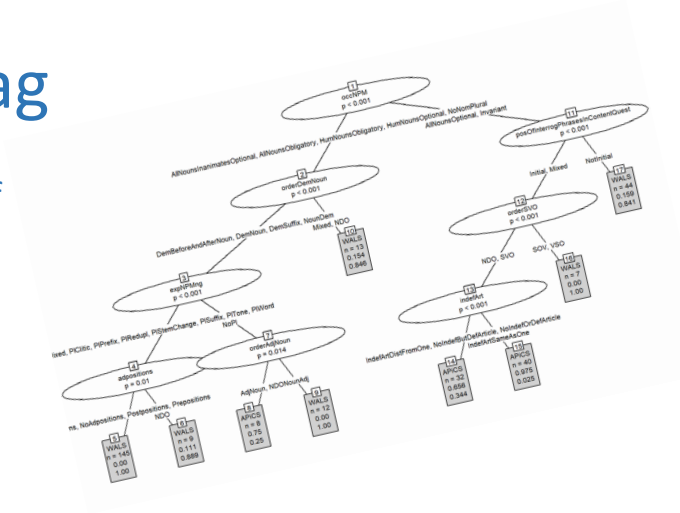
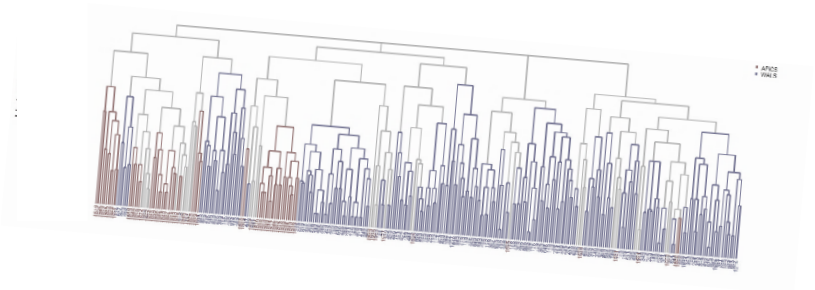
Measuring the similarity between languages: The case of creoles and non-creoles



Lara Rüter & Ingo Plag

Heinrich-Heine-Universität Düsseldorf





Introduction

Typology and comparative linguistics

- Successful rise of quantitative methods
- Assess the similarities and differences between languages by quantifying over linguistic features
- Explain "what's where why?" (Nichols 1992, Bickel 2015)

Creole studies

- Are creoles less complex than non-creoles?
- Are creoles more similar to their substrate languages than to their superstrate languages?
- **Are creoles structurally different from non-creole languages?**

(e.g. Bakker et al. 2011, Bakker et al. 2017, Daval-Markussen 2019)

Problems

Methodological

- Why compare? Why not compare?
- Sampling: features
- Sampling: languages
- Coding of features
- Missing data points (WALS: 85%)
- Choice of statistical model (so far: only Neighbor-net)

We will show that

- Different samples yield similar results.
- Different statistical models yield similar results.
- Results remain robust across methods.

This paper

- Discuss merits and problems of comparison
- Compare creoles and non-creoles
- Determine their degree of similarity
- Subset of APiCS and WALS features (nominal categories and word order)
- Explore what **different methods** can tell us about our research questions:

How [sic] different are creoles from non-creoles?

How much do they differ?

In which respect do they differ?

- Two perspectives: clustering and predicting
- Do creoles and non-creoles form different clusters?
- Can we predict the class of a given language based on the features?

Why compare? Why not compare?

- Fierce debate in P&C studies since Bakker et al. (2011)
- Comparative approaches are used across scientific disciplines (philosophy, genetics, law, musicology, computer science, sociology, history, economics, education and literary studies, ...)
- Very similar problems across disciplines
- **Reduction**
  Complex phenomena are reduced (and thereby altered) to entities that can be recognized.
- **Centrism**
  Subjective experience of the researcher, who will inevitably tend to impose their own concepts to measure the distance between the entities under study.
- **Way out**
  'Descriptive categories' vs. 'comparative concepts' (e.g. Haspelmath 2010)

Decriptive categories vs. comparative concepts

- What are the cross-linguistic entities that can be compared?

Language-particular 'descriptive categories'

- Often similar across languages, but cannot capture cross-linguistic similarities and differences
- Cannot be equated
- Work independently of comparative linguistics

'Comparative concepts'

- Concepts specifically designed for comparison that are independent of descriptive categories
- Cannot be right or wrong, only more or less useful
- Universally applicable
- Universal conceptual-semantic concepts
- Universal formal concepts
- Needed to test universal claims

How to compare

- Comparison has been a very useful method in the history of our discipline.
- The **reduction** problem and the **centrism** problem must be adequately addressed.
- This crucially holds for qualitative and quantitative research.
- WALS and APiCS do this by making use of comparative concepts:

"abstract structural features that make reference to structural properties that can be identified in any language. These can be general concepts of language form such as 'precedes/follows', 'overt/zero', 'identical/different', or semantic-pragmatic concepts like 'negation', 'question', 'focus', or more complex comparative concepts defined on the basis of such elementary formal concepts and semantic-pragmatic concepts (e.g. 'subject', 'pronoun')." (Michaelis et al., 2013,xxxvii).

- Using typological databases also eliminates one's own biases.

Methodology: Features and languages

Our sample

- WALS languages vs. creoles from APiCS
- Random sample of 21 features

	Languages	Features	Missing values (NAs)
APiCS	76	130	3 %
Creoles, 21 features	54	21	1 %
WALS	2662	192	83 %
Subset, 21 features	1863	21	50 %

Methodology: Features

Nominal categories (12)

- Gender distinctions in personal pronouns
- Incl./excl. distinction in independent personal pronouns
- Politeness distinctions in second-person pronouns
- Indefinite pronouns
- Occurrence of nominal plural markers
- Expression of nominal plural meaning
- Definite articles
- Indefinite articles
- Pronominal and adnominal demonstratives
- Distance contrasts in demonstratives
- Adnominal distributive numerals
- Sortal numeral classifiers

Word order (9)

- Order of subject, object, and verb
- Order of possessor and possessum
- Order of adjective and noun
- Order of adposition and noun phrase
- Order of demonstrative and noun
- Order of cardinal numeral and noun
- Order of relative clause and noun
- Order of degree word and adjective
- Position of interrogative phrases in content questions

Illustration: Languages and missing values

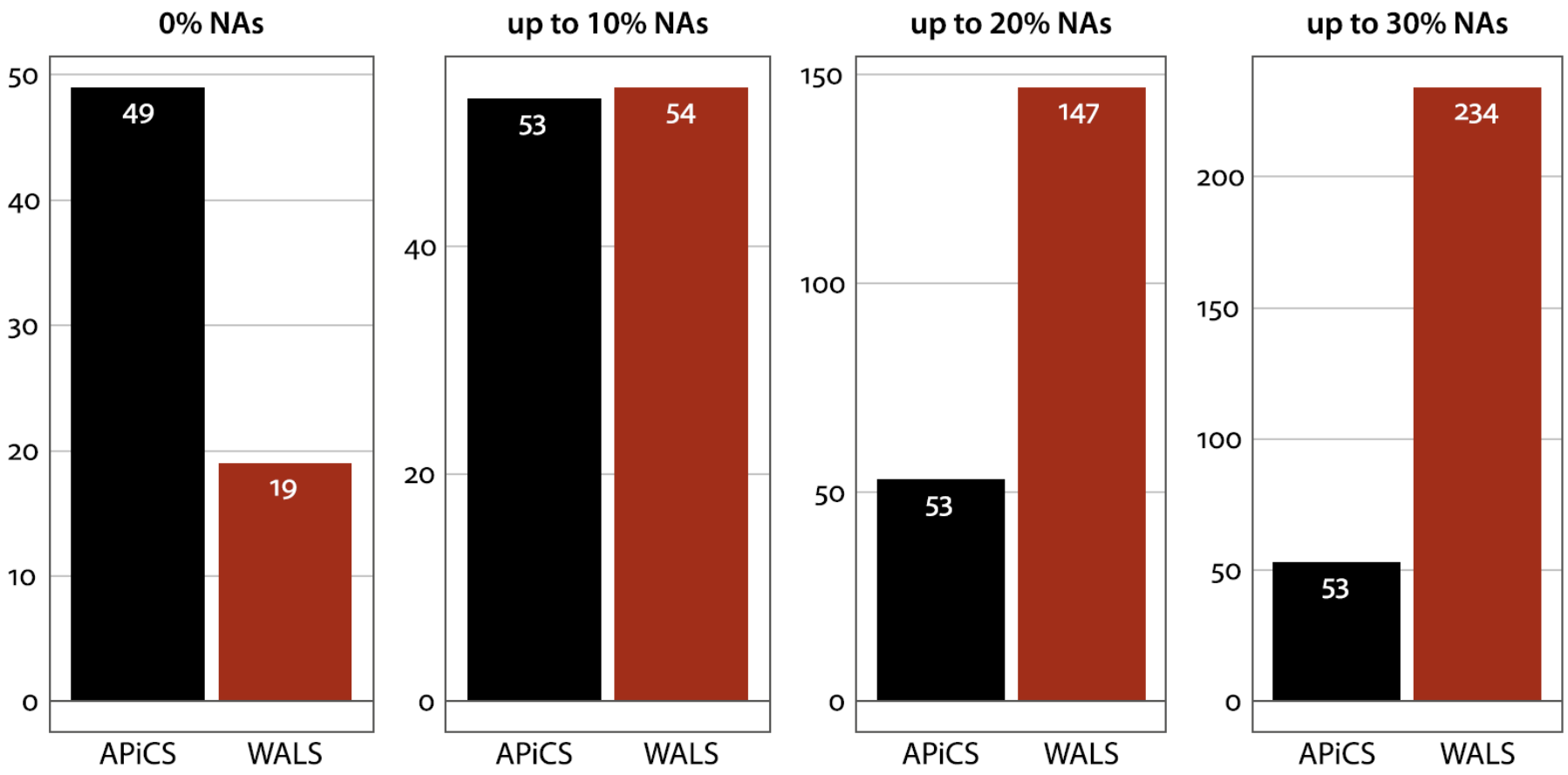
- Not all languages are coded for all features

Feature name	Area	APiCS total	WALS total
Search	--any--	Search	Search
Order of subject, object, and verb	Word order	78	1377
Order of possessor and possessum	Word order	77	1248
Order of adjective and noun	Word order	76	1366
Order of adposition and noun phrase	Word order	77	1185
Order of demonstrative and noun	Word order	79	1223
Order of cardinal numeral and noun	Word order	76	1154
Order of relative clause and noun	Word order	76	825
Order of degree word and adjective	Word order	77	481
Position of interrogative phrases in content questions	Word order	76	901

```
   orderSVO
SOV    :565
SVO    :550
NDO    :204
VSO    : 95
VOS    : 25
(Other): 15
NA's   :498
```

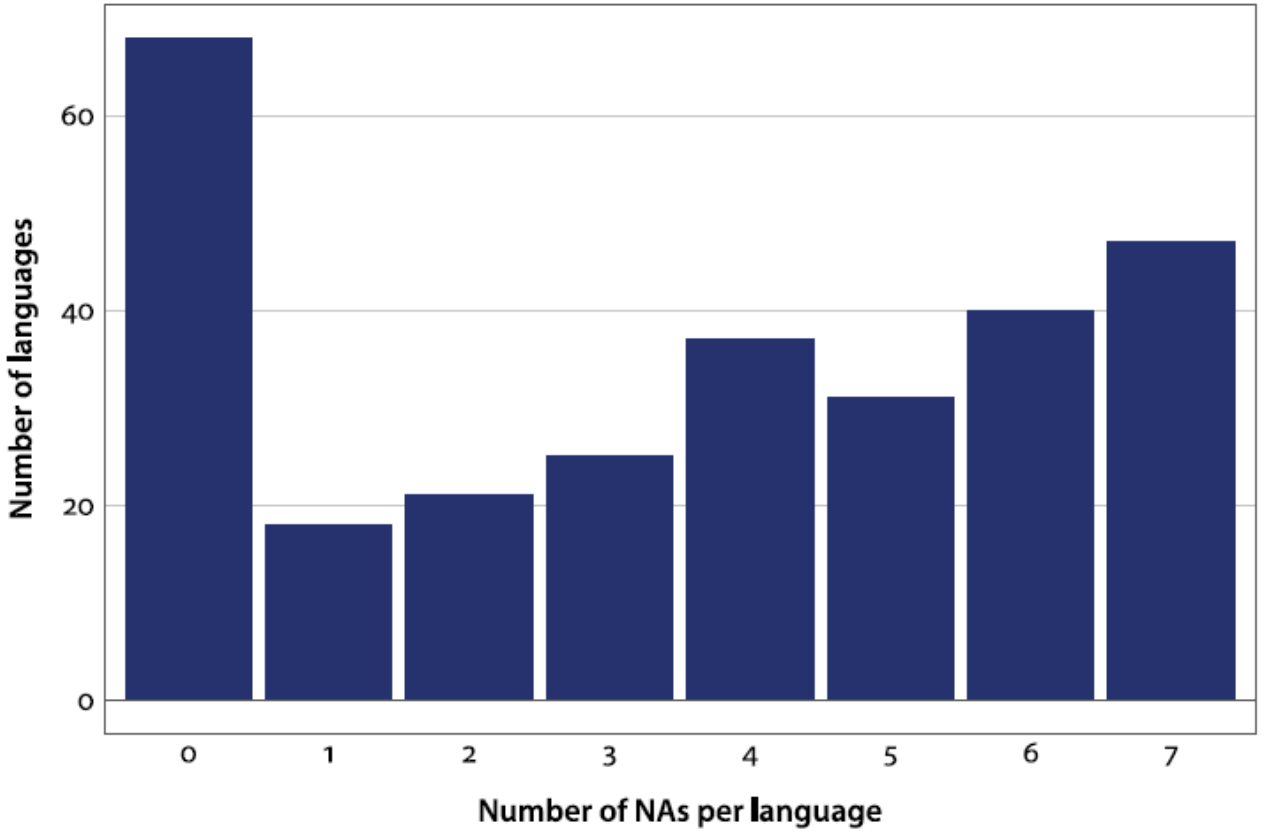
```
            genDistPP
NoGenDist       : 314
3pSgOnly        :  75
3pSgAndNonSg    :  42
3pPlus1pAndOr2p:  18
3pSgAndPlOnly   :   5
(Other)         :   5
NA's            :1493
```

Methodology: Language samples and missing values



Missing values

Data set	Proportion of NA's
10NA	0.02
20NA	0.07
30NA	0.12



Methodology: Statistical models

Clustering

- Phylogenetic trees
  - Neighbor-net
  - Neighbor-joining algorithm
- Hierarchical cluster analysis

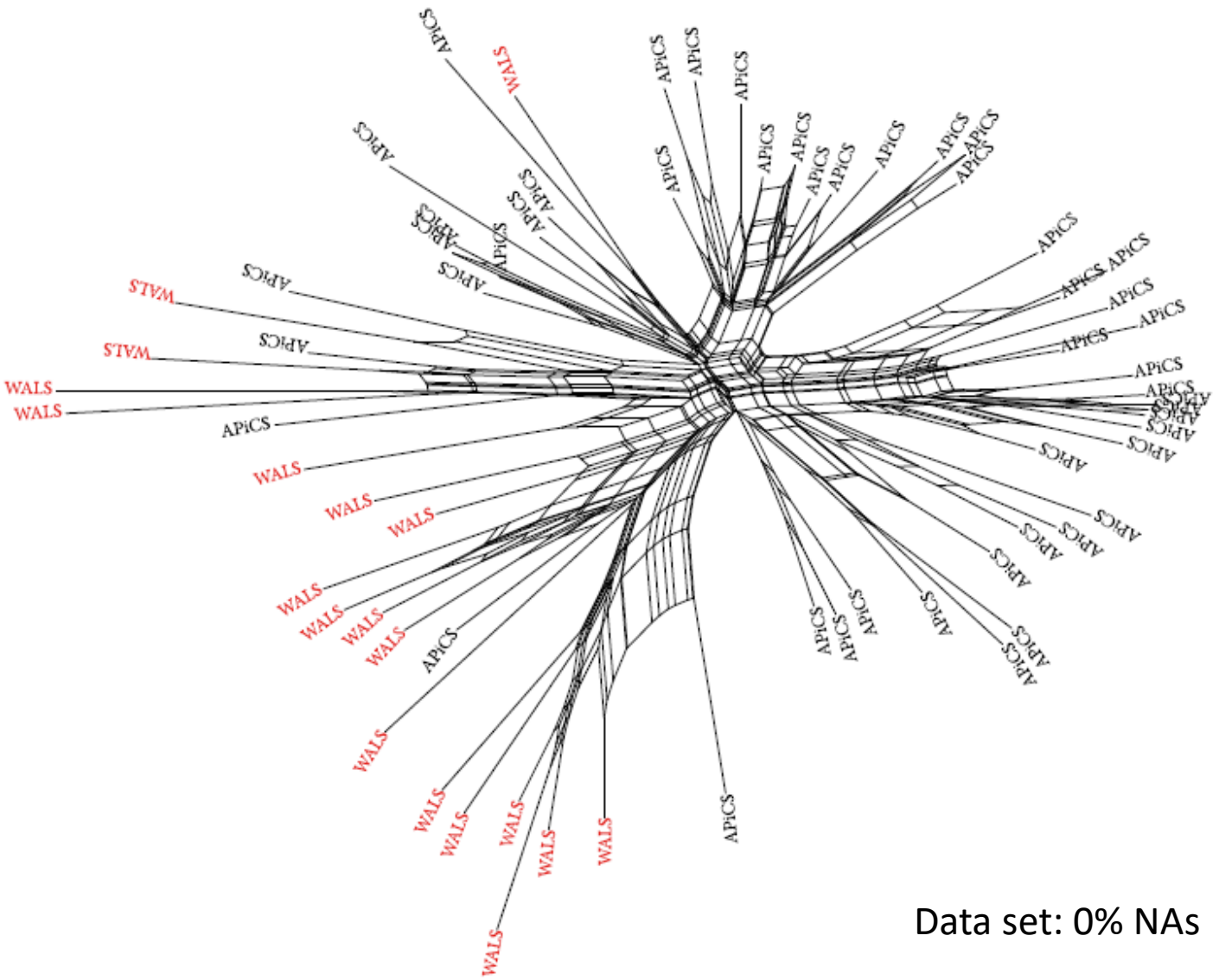
Predicting

- Classification trees
- Random forests

Stats: `R`; packages: `ape, caret, cluster, mltools, phangorn, party, partykit`

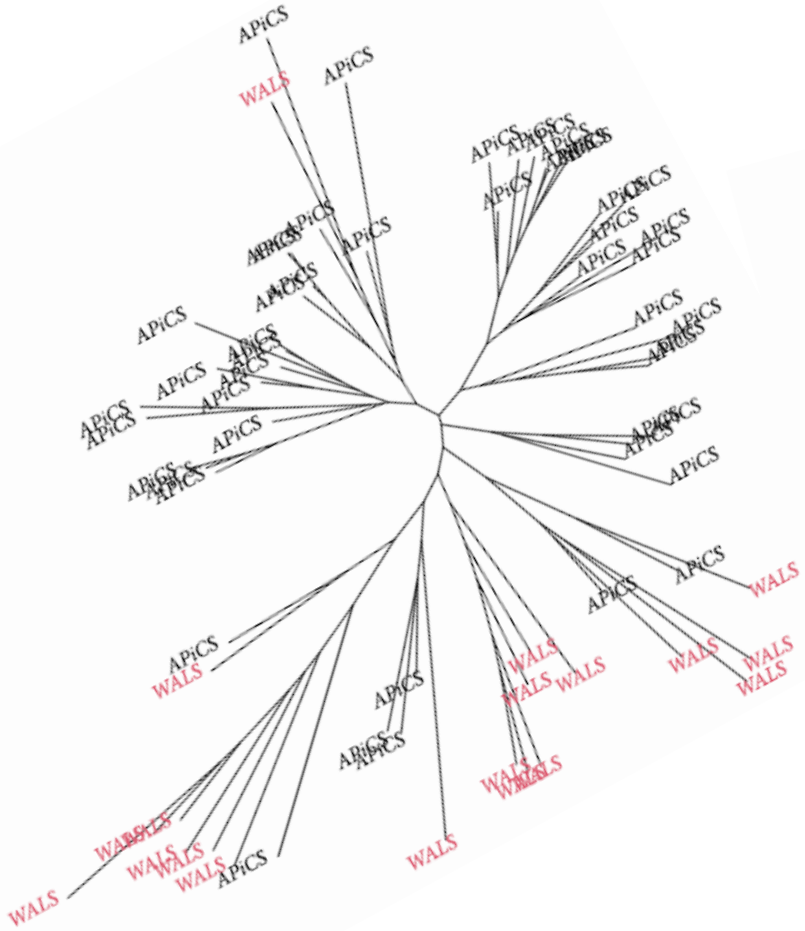
Results

Phylogenetic trees: Neighbor-net

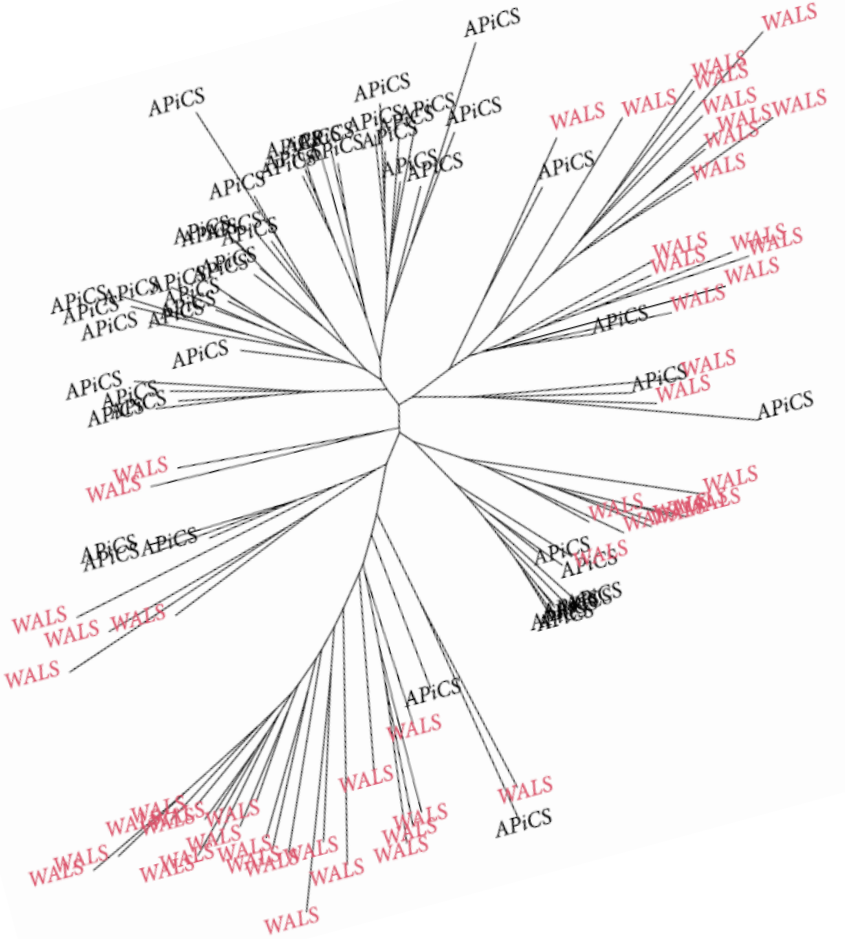


Data set: 0% NAs

Phylogenetic trees: Neighbor-joining algorithm

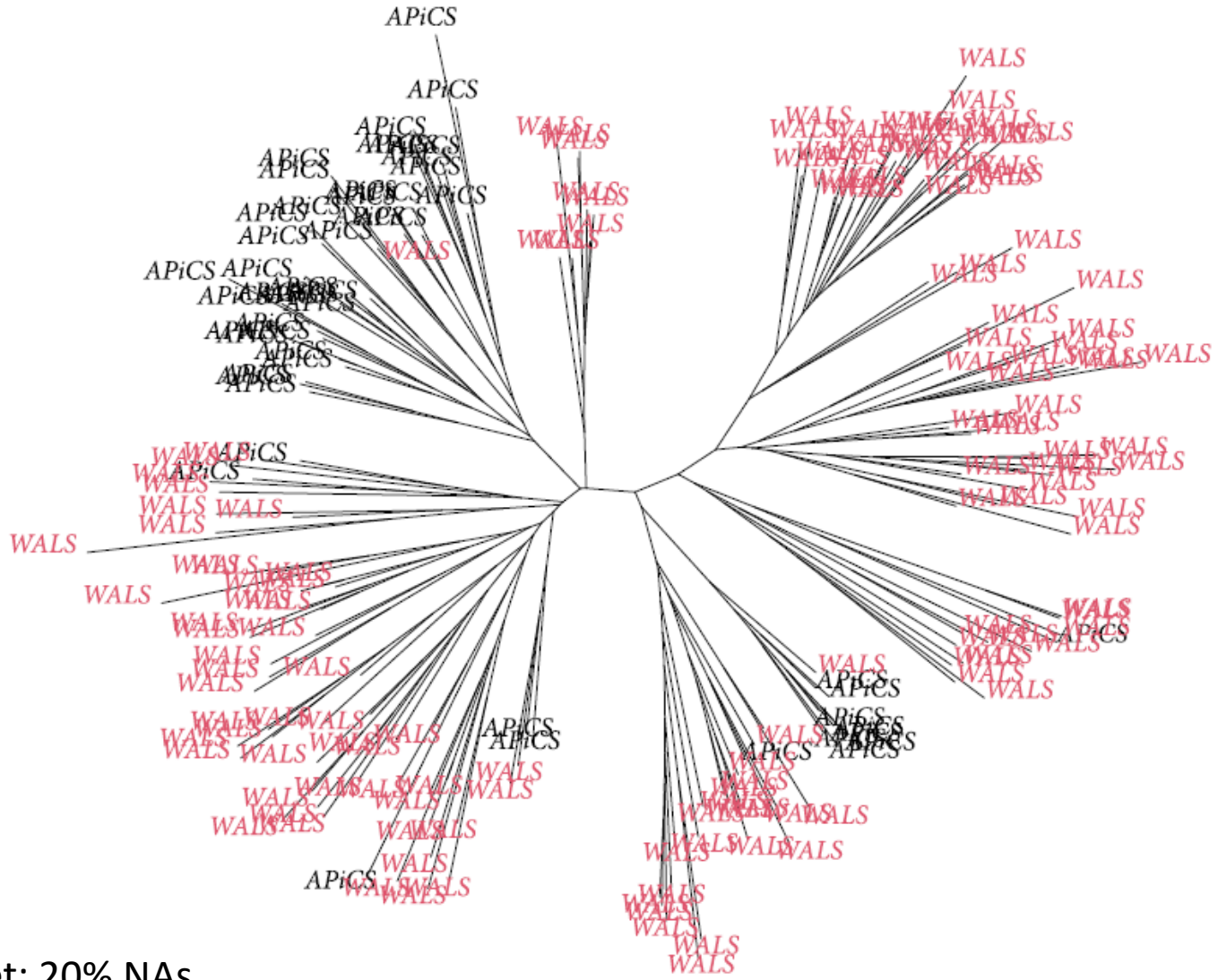


Data set: 0% NAs



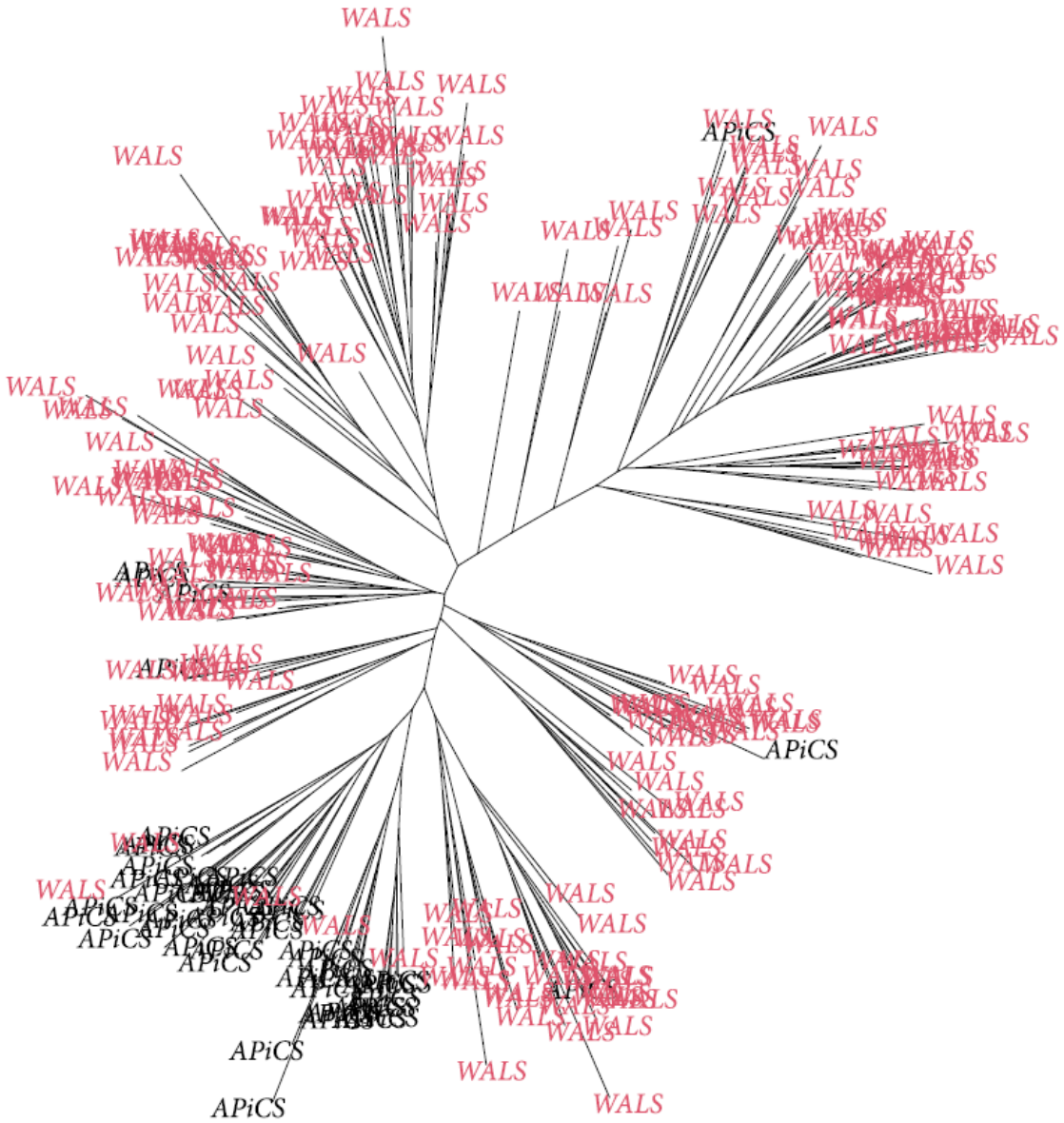
Data set: 10% NAs

Phylogenetic trees: Neighbor-joining algorithm



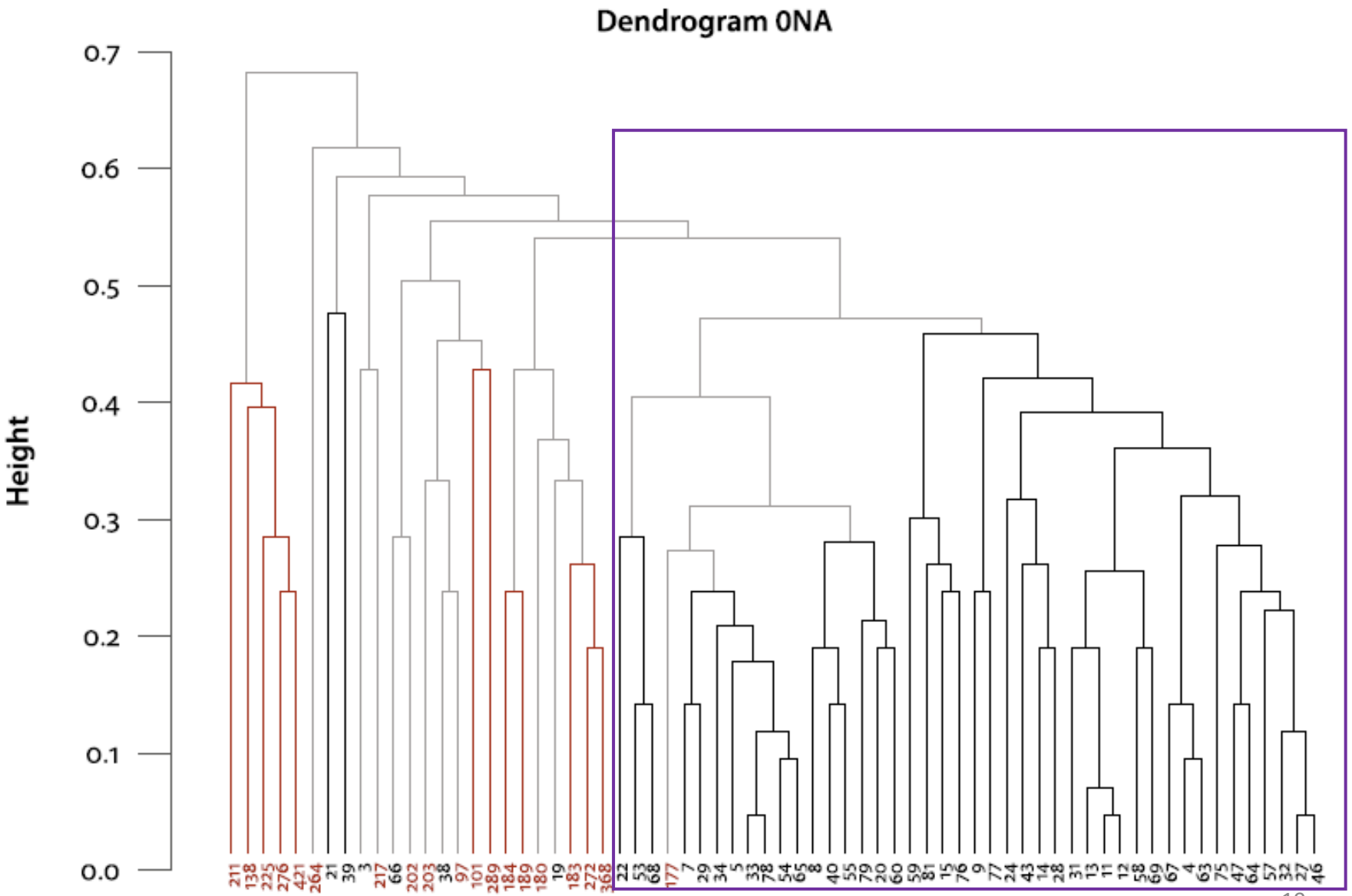
Data set: 20% NAs

Phylogenetic trees: Neighbor-joining algorithm

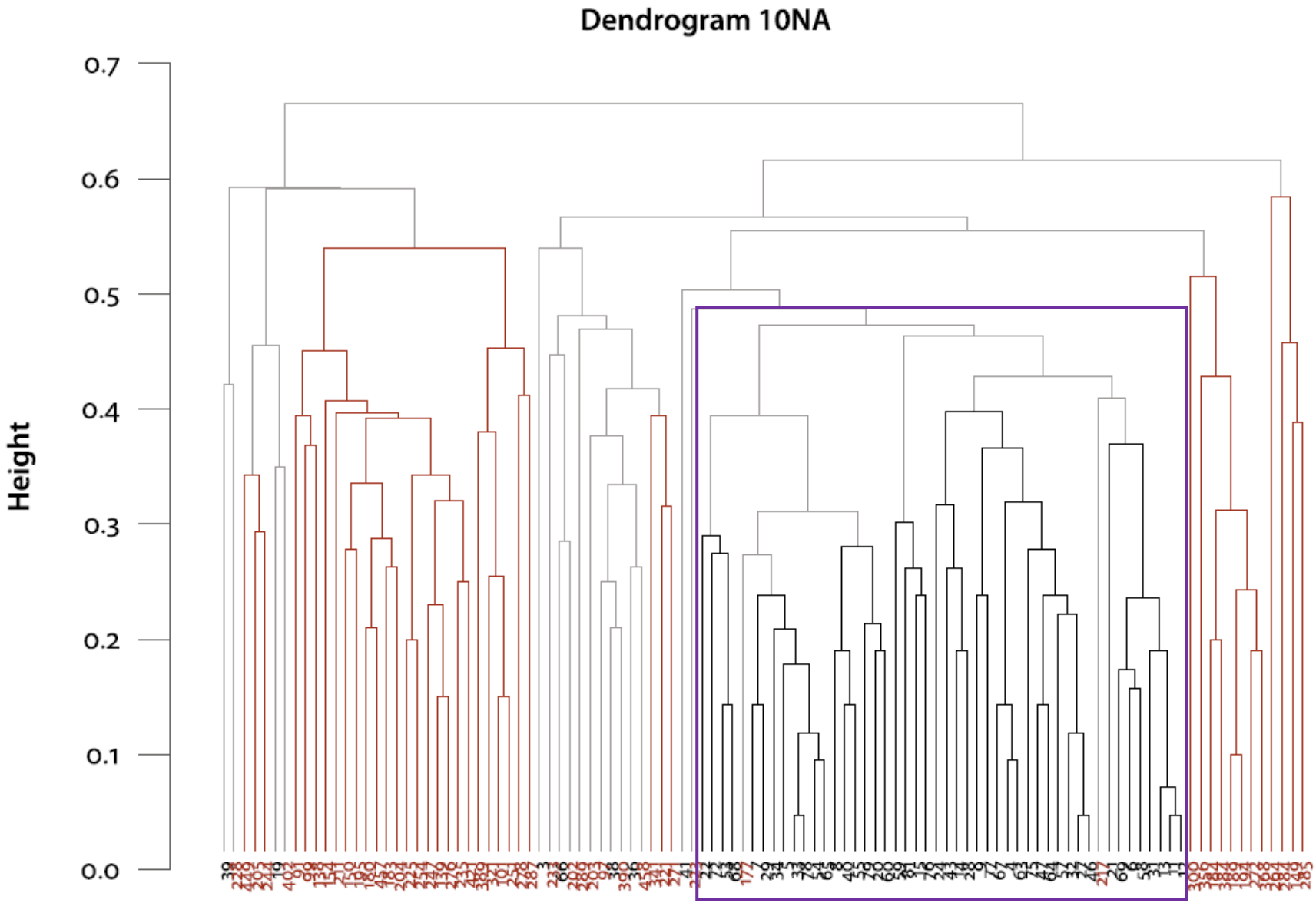


Data set: 30% NAs

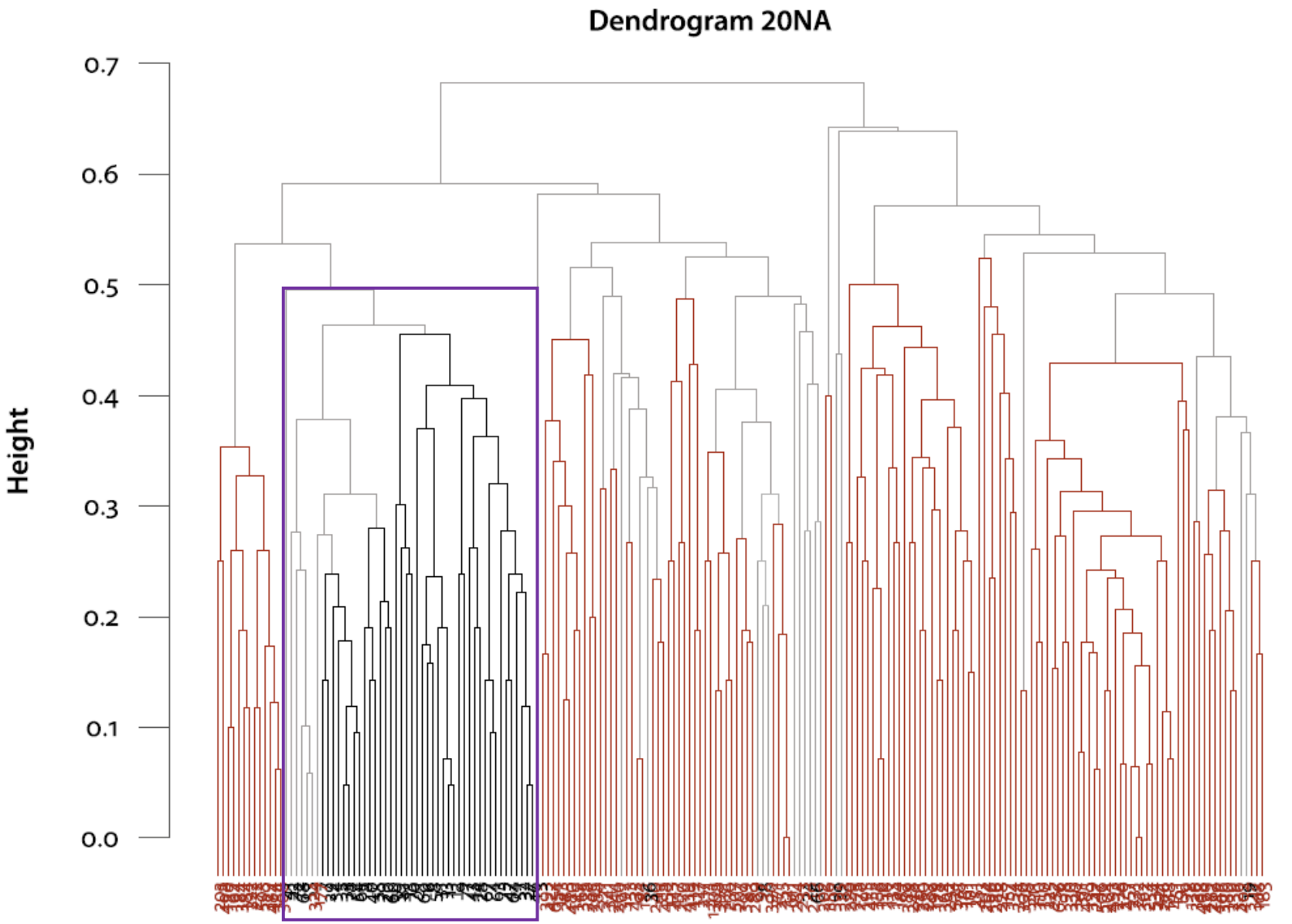
Hierarchical cluster analysis



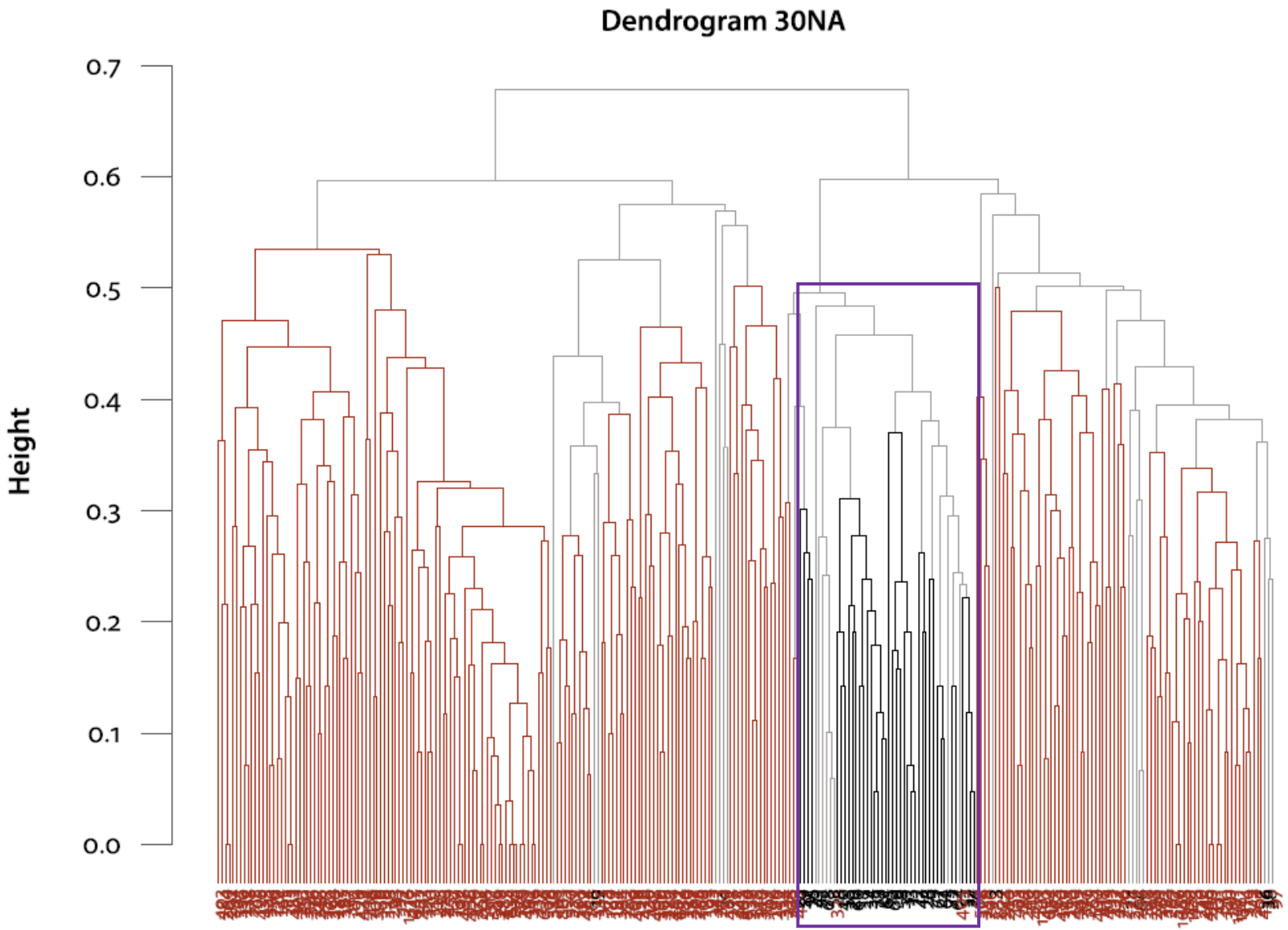
Hierarchical cluster analysis



Hierarchical cluster analysis



Hierarchical cluster analysis



Clustering: Interim summary

- Creoles do cluster to a large extent.
- Non-creoles do cluster to a large extent.
- The two kinds of languages are not completely separated in different clusters.
- There is a clear **tendency** for creoles and non-creoles to differ.

Predicting: Trees

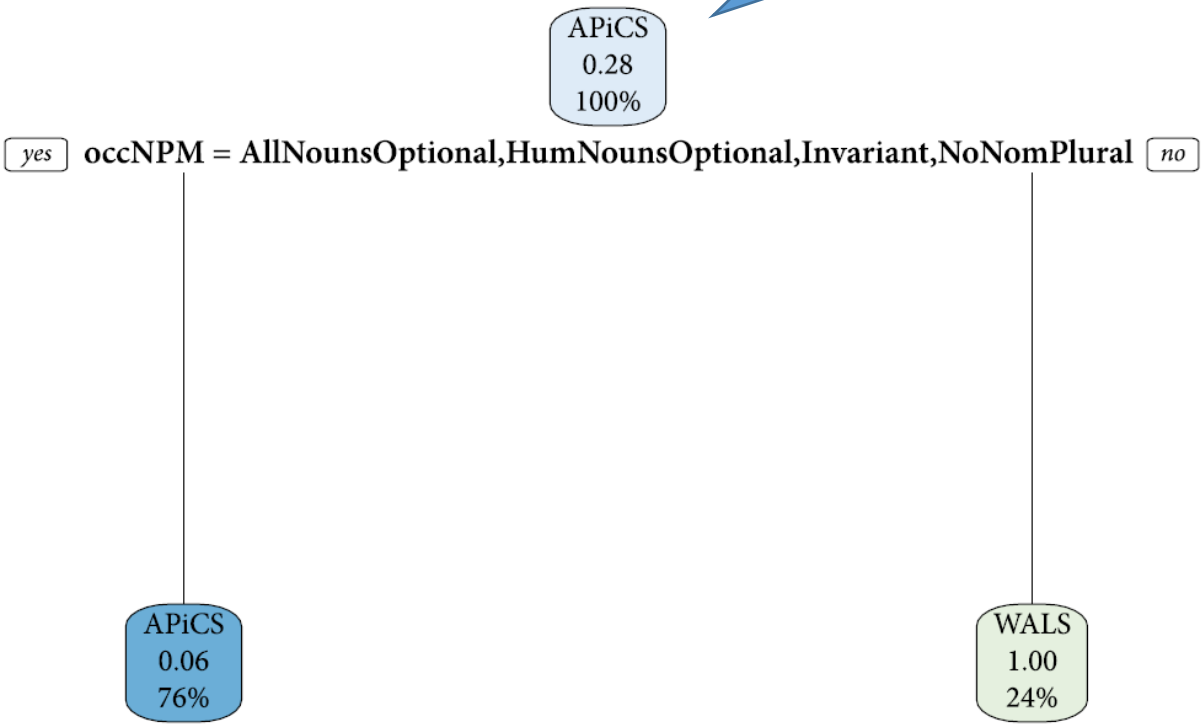
- Classification and regression trees predict an outcome (class of language) on the basis of a set of predictors and *their specific constellations* (features).
- We use
  - Recursive partitioning and regression trees.
  - Conditional inference trees with random forests.

Idea: Find subsets of languages that share specific feature constellations and behave uniformly w.r.t. the outcome (i.e. the class of language)

Predicting:
Recursive partitioning trees

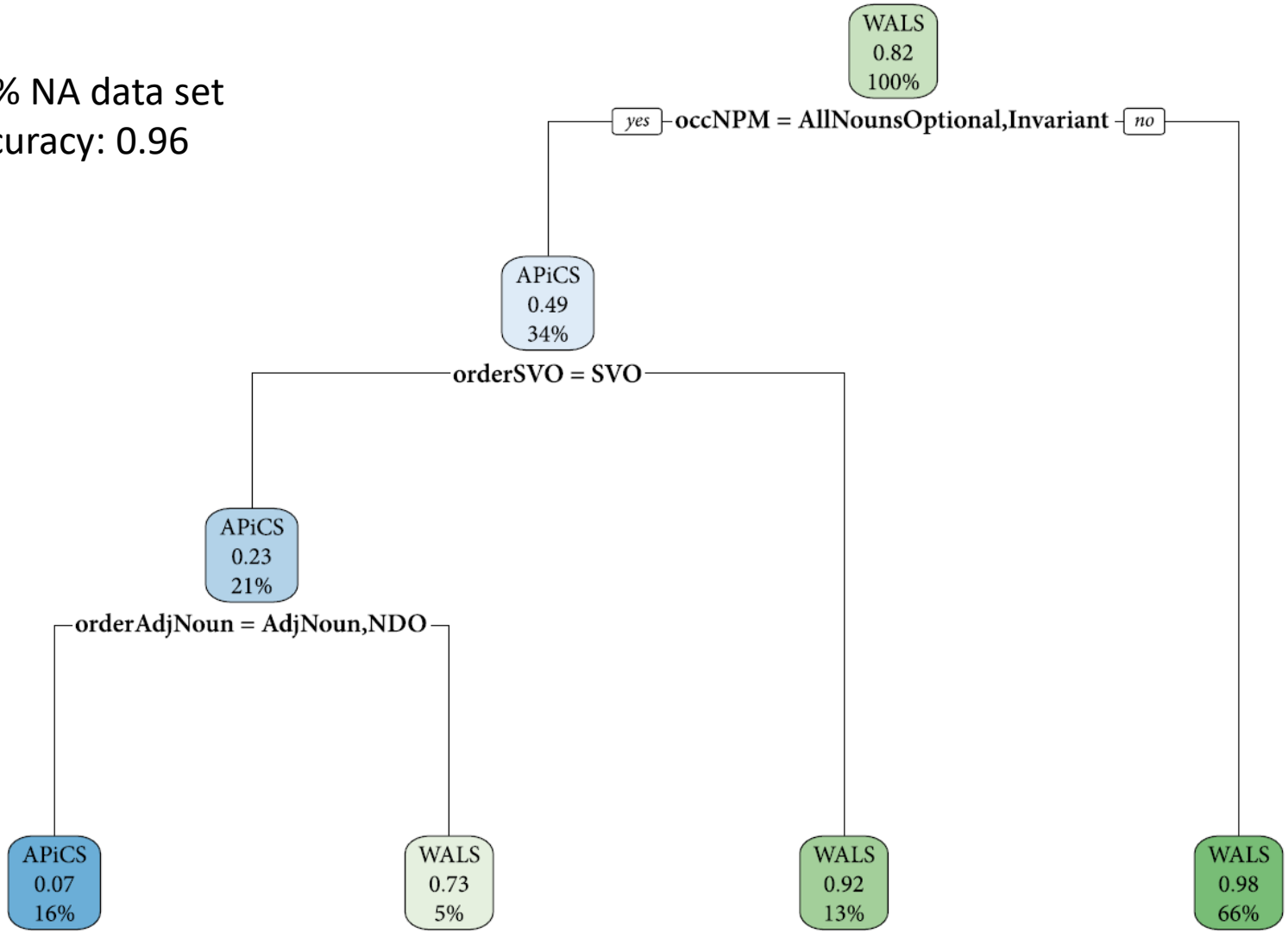
- the predicted class
- the predicted probability of WALS
- the percentage of observations in the node.

0% NA data set
accuracy: 0.96



Predicting: Recursive partitioning trees

30% NA data set
accuracy: 0.96



Predicting: Random forests

A ‘random forest’

- ... combines multiple decision trees.
- Each tree is devised on a random subset of the data and predicts an outcome.
- The prediction of the random forest as a whole is determined by aggregating the predictions of its trees.
- 1000 trees for each data set, each tree was built using 0.632 of the data set.

Data set	Feature	F1
0NA	occNPM adpositions orderAdjNoun	0.97
10NA	occNPM adpositions orderAdjNoun	0.96
20NA	adpositions occNPM orderSVO	0.94
30NA	occNPM adpositions orderSVO	0.90

Summary and discussion

- Different data sets yield very similar results.
- Different statistical models yield similar results.
- Clustering techniques show a tendency towards two classes.
- Predictive models reach satisfactory accuracies in classifying languages as either creoles or non-creoles, based on very few features.
- There are gradient, but clear and predictable, differences between creoles and non-creoles.
- Certain features are highly predictive of language class (creole):

  nominal plural marking (no marking)

  adpositions (prepositions)

  order of adjective and noun (ADJ before N)

  order of verb, object, subject (SVO)

- These facts call for explanations.

Thank you very much
for your attention!