

GAMs with wolves

Gerhard Jäger, Tübingen University

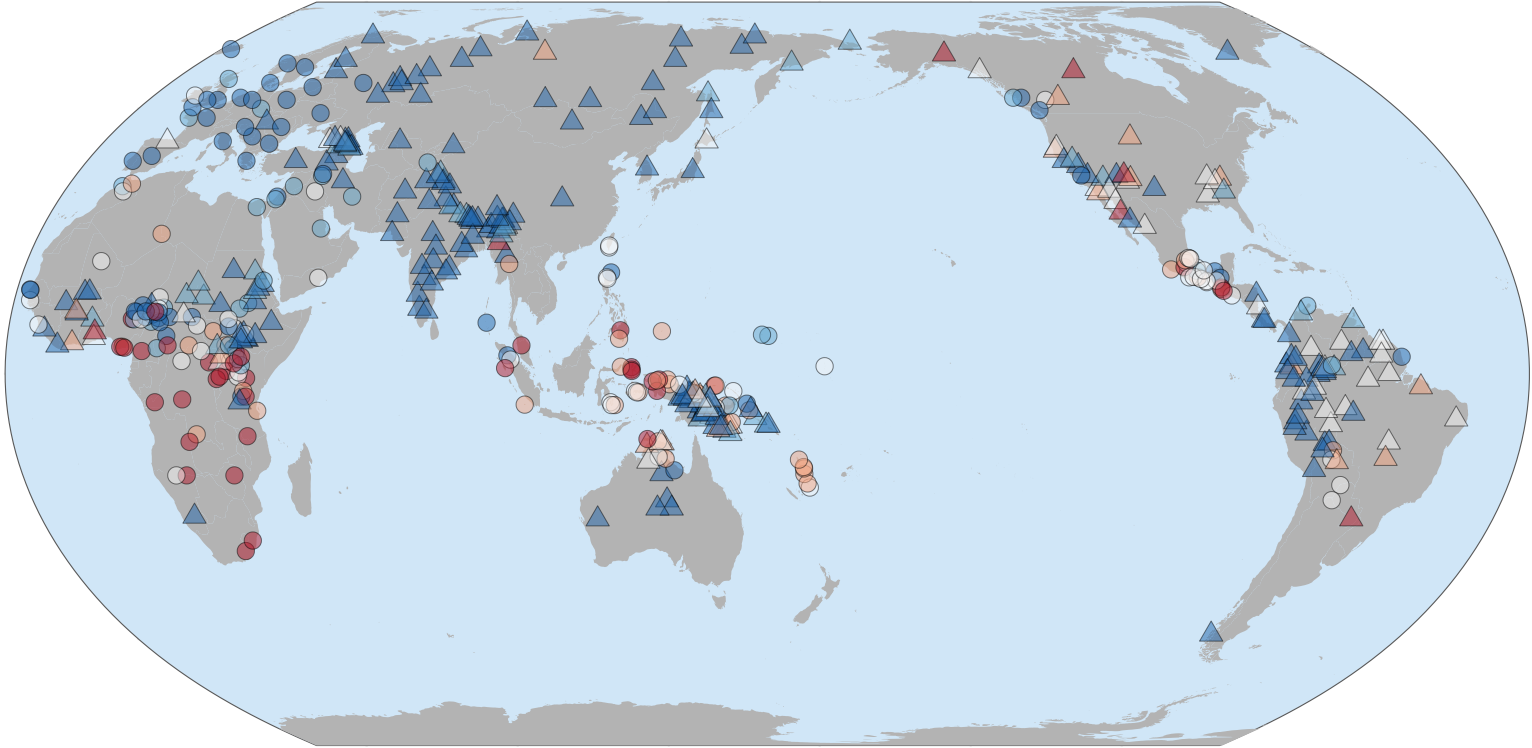
C-LESTE Workshop

Frankfurt, June 6, 2025

Non-independence of typological variables

geographically

Affixing type by adposition

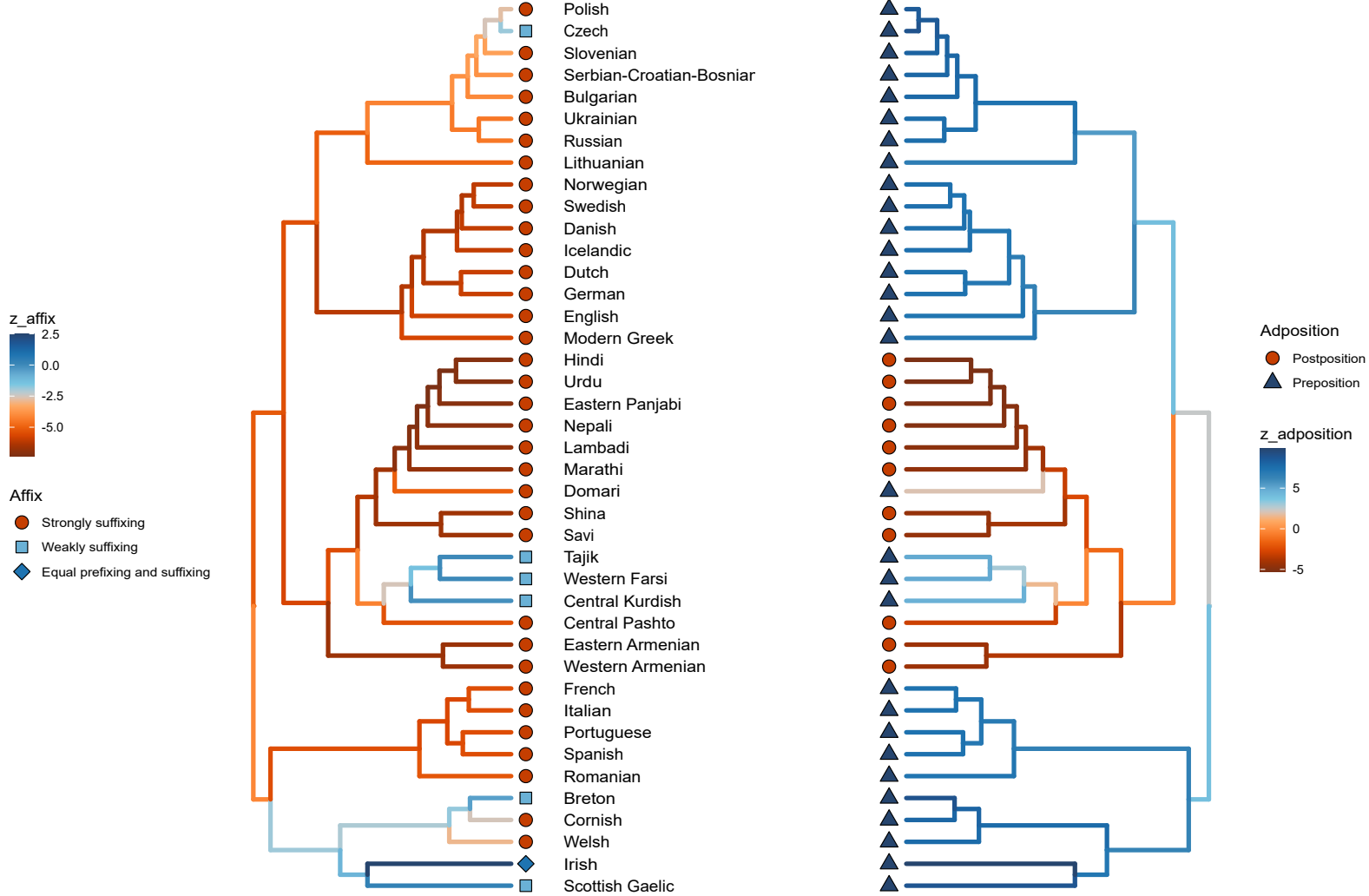


Adposition type △ Postposition ○ Preposition

Affix type ● Strongly suffixing ● Weakly suffixing ○ Equal prefixing and suffixing ● Weakly prefixing ● Strongly prefixing

(fromäger 2025 Computational Typology arxiv)

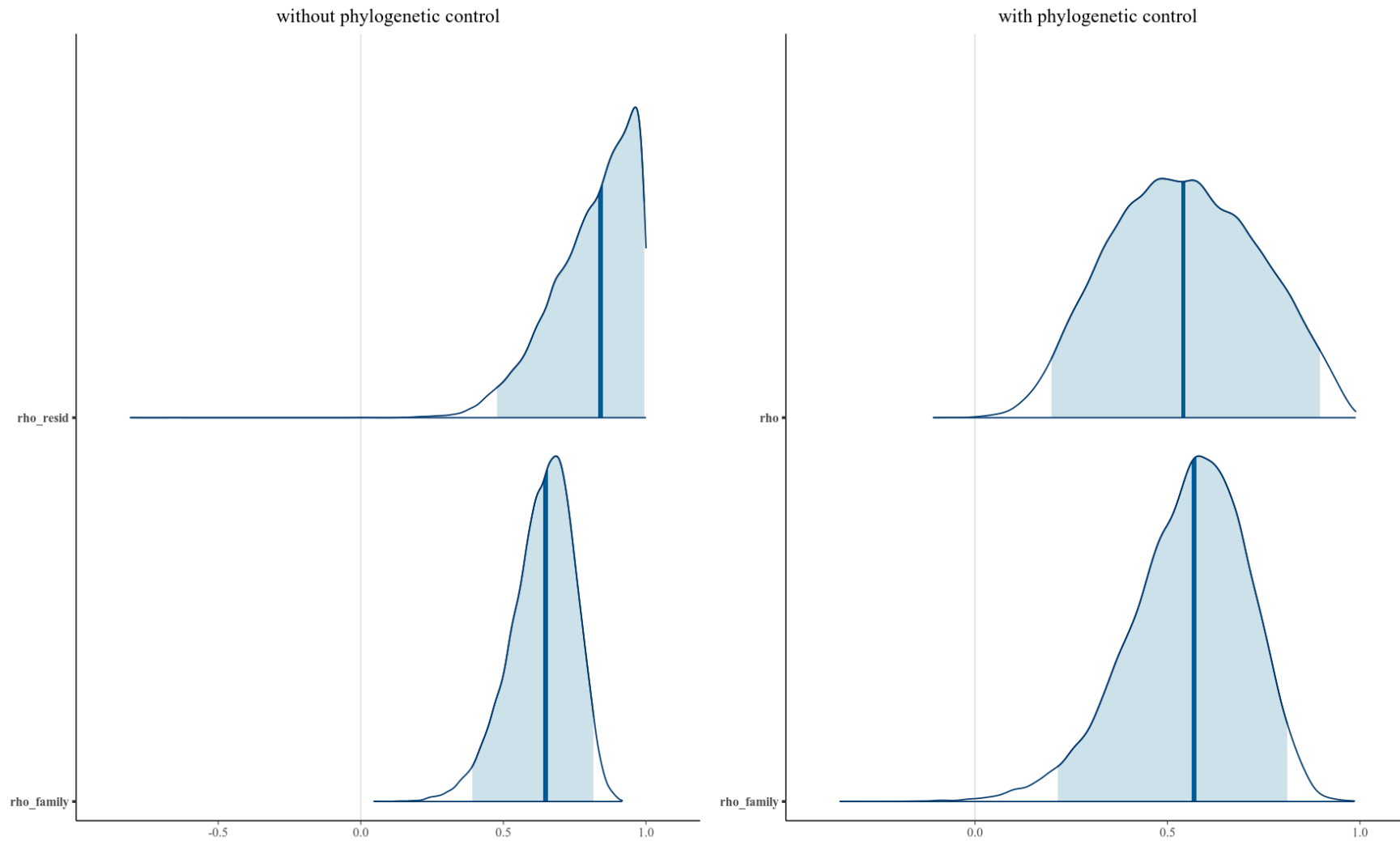
phylogenetically



(from Jäger, 2025, Computational Typology, arxiv)

non-independence

- overlooking non-independence can lead to faulty conclusions



(plots come from the study on affix-adposition correlation in Jäger 2025)

Statistical techniques to control for non-independence in typology

- **random effects** (genealogical units; macro-areas, phylogenetic covariance)
- **phylogenetic comparative method** (e.g., Pagel & Meade 2006)
- **spatial statistics**, e.g. Gaussian Processes

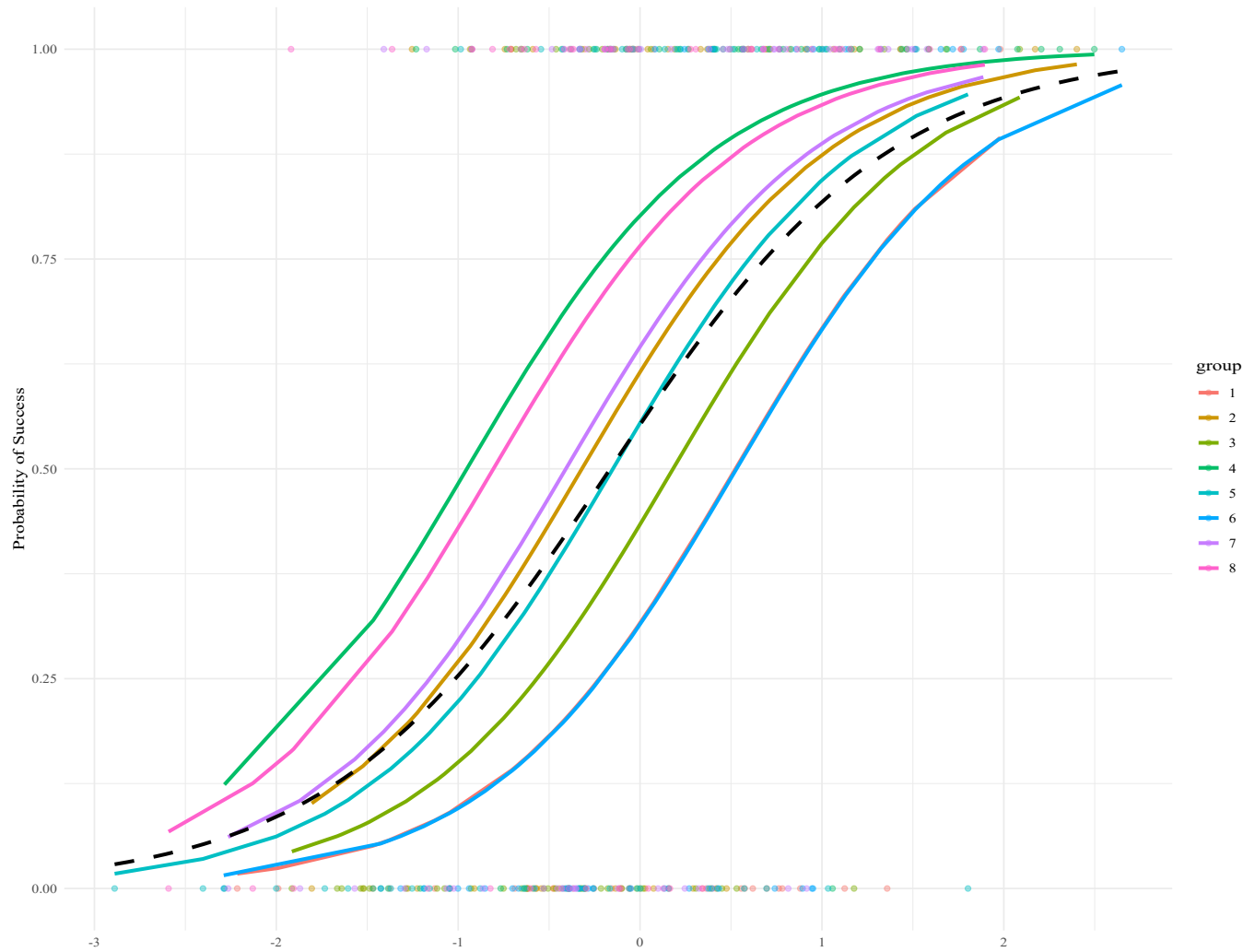
Random intercepts

- every group (family, genus, macro-area) has it's own probability distribution over observed values
- distributions come from the same statistical family

Mixed-Effects Logistic Regression with Random Intercepts

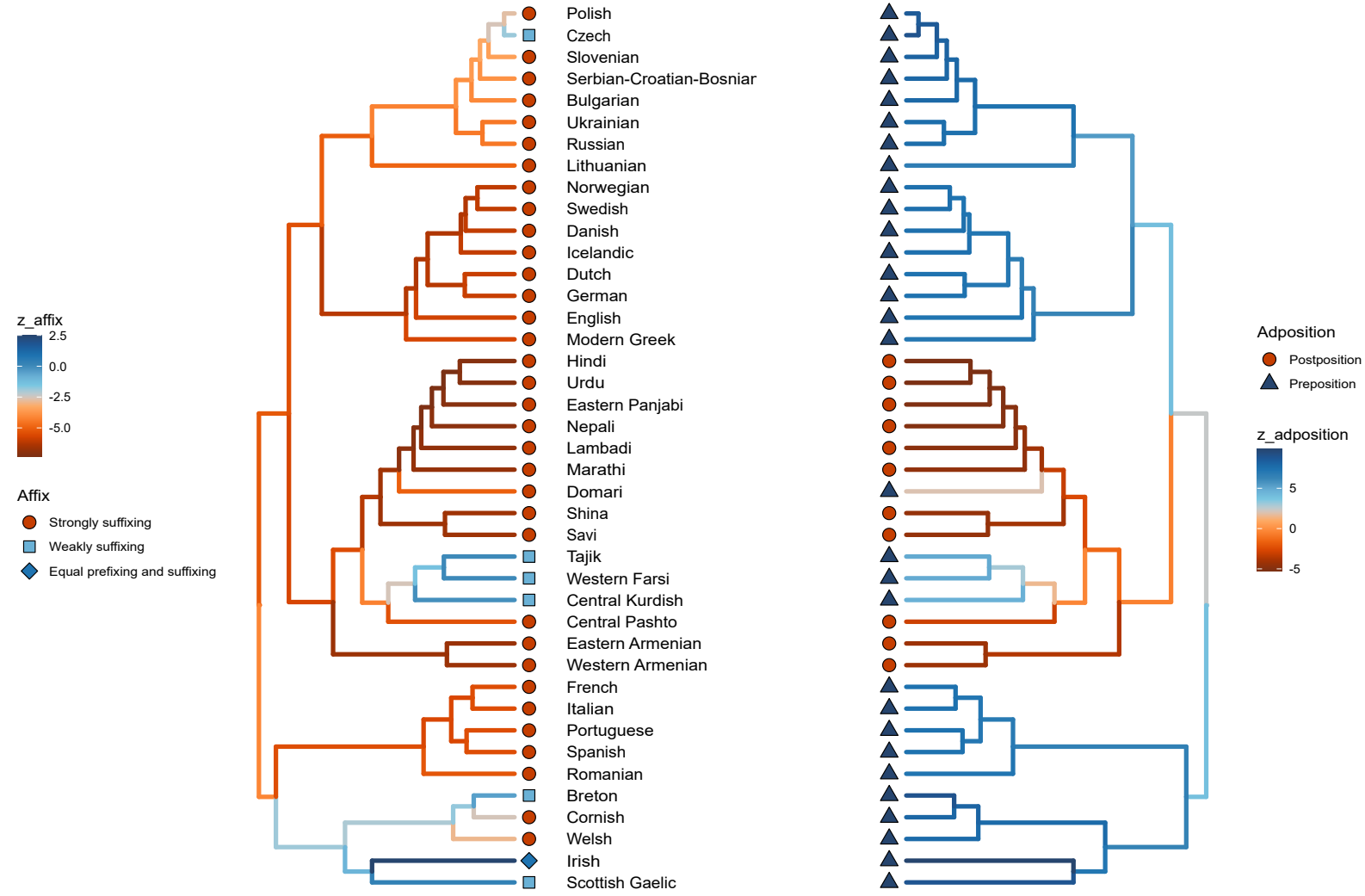
Colored lines: group-specific fits (random intercepts)

Black dashed line: global trend (fixed effects only)



Phylogenetic random effect

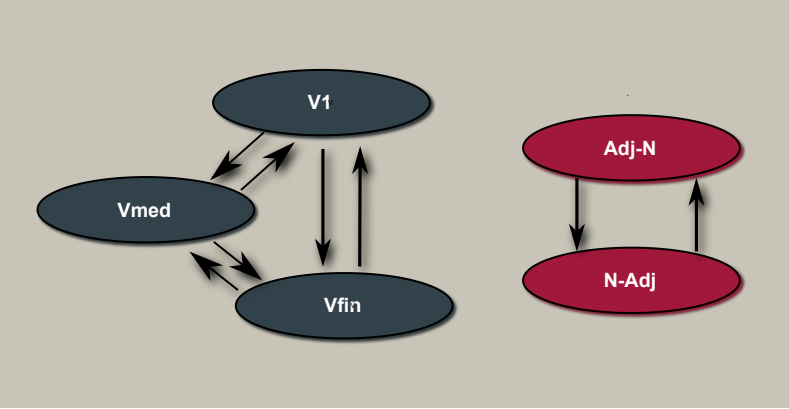
a.k.a. random walk along the phylogeny



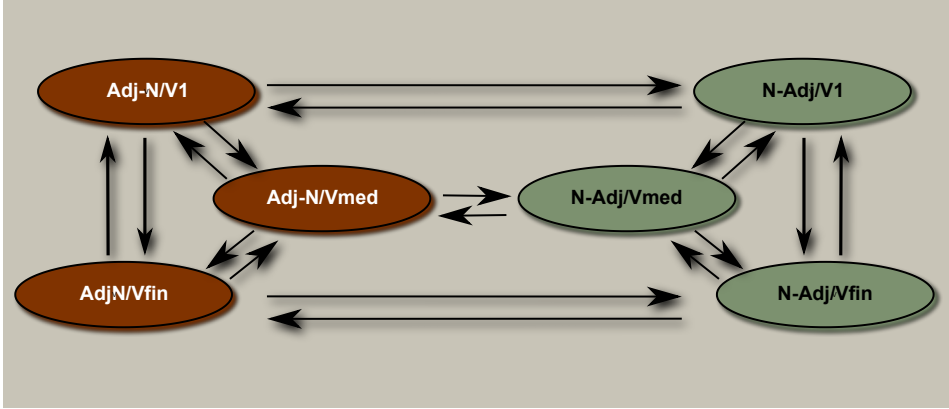
Phylogenetic comparative method

Pagel & Meade 2006

Independent model

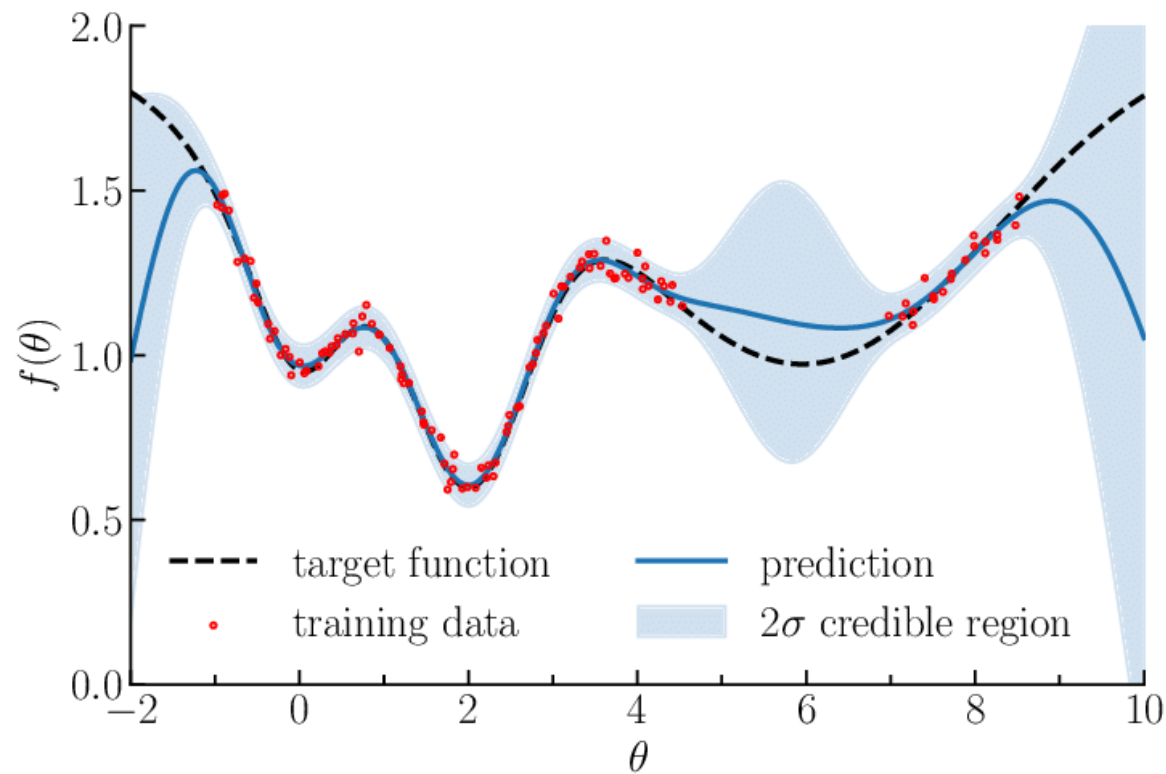


Dependent model

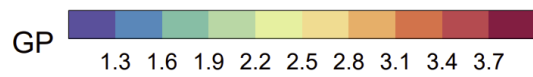
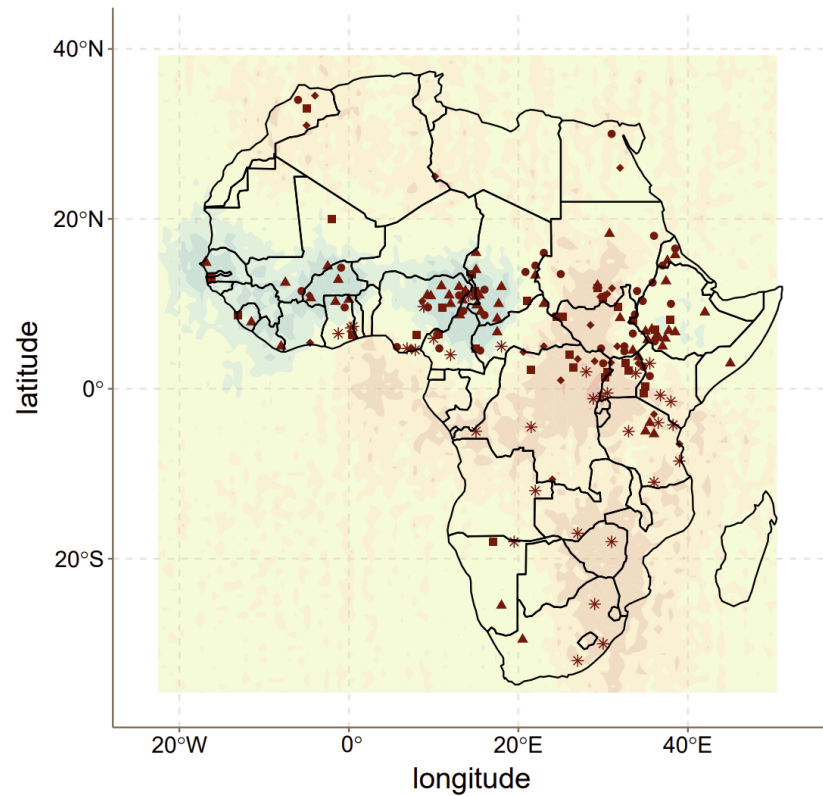


Spatial techniques

Gaussian processes



(image from <https://journals.aps.org/prd/abstract/10.1103/PhysRevD.98.063511>)

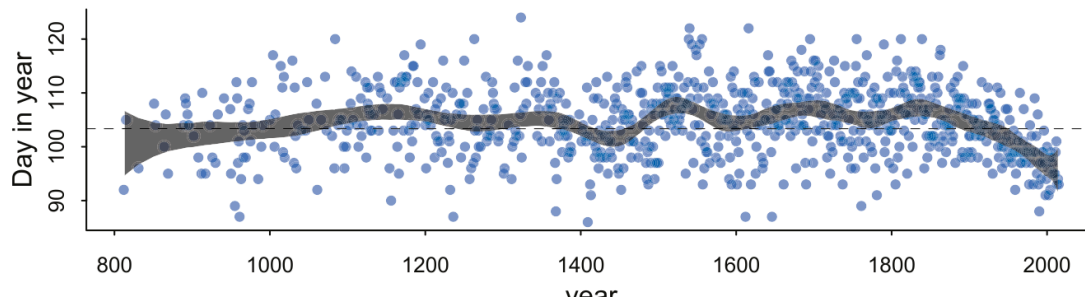
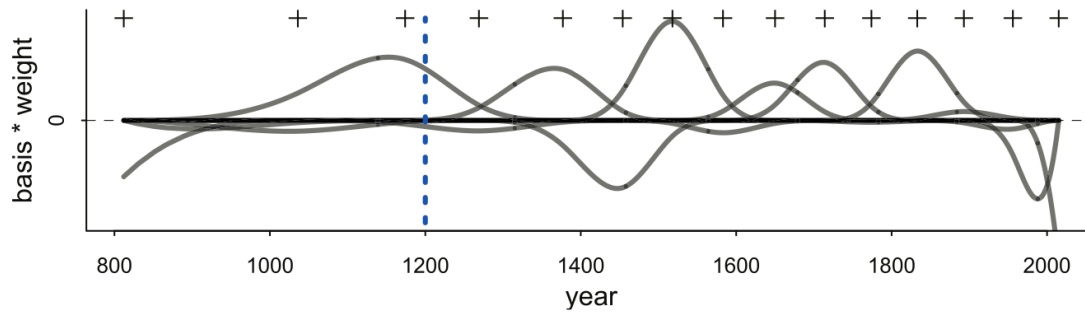
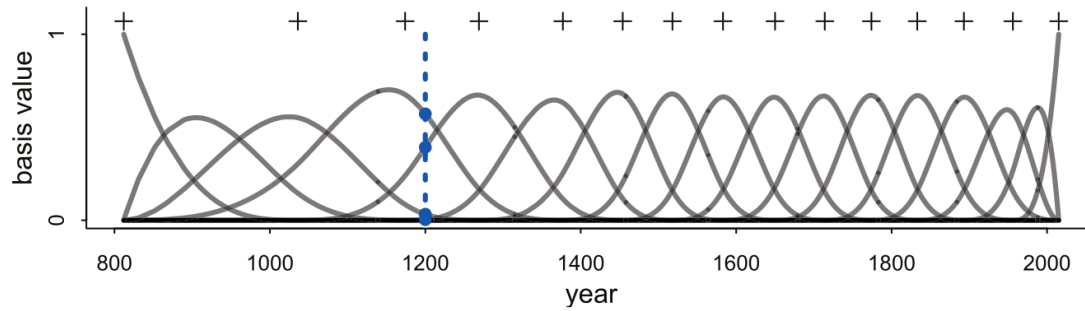


affixation ▲ strongly suffixing ● weakly suffixing ■ equal ◆ weakly prefixing * strongly prefixing

(from Guzmán Naranjo & Becker 2021)

Spatial techniques

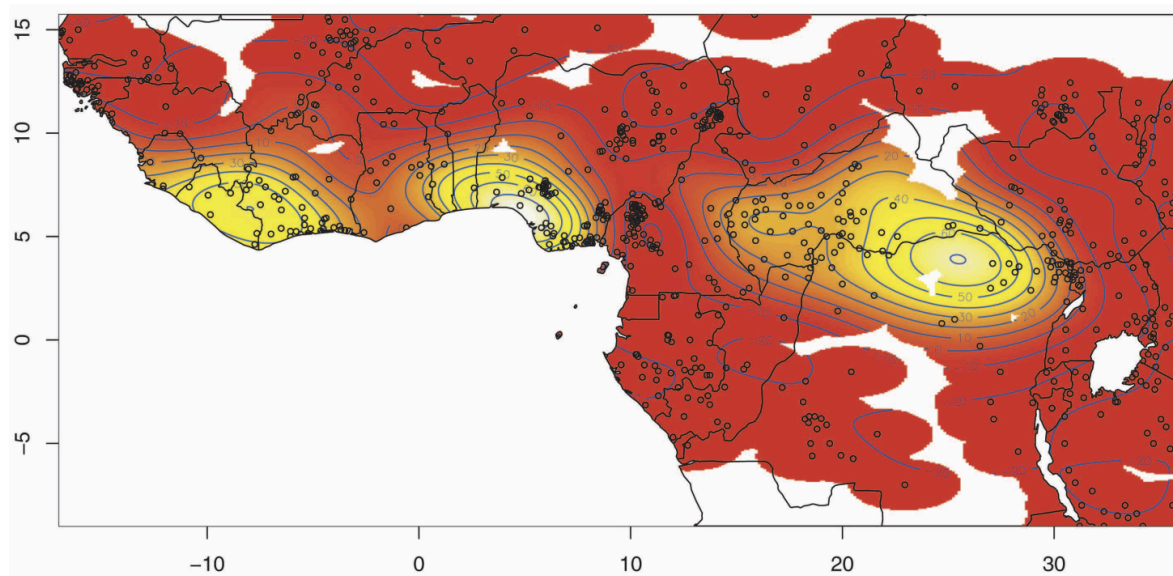
Generalized Additive Models/Splines



(from McElreath 2020)

GAMs

- note as popular yet in typology, but there are applications, e.g. Idiatov & Van der Velde 2021

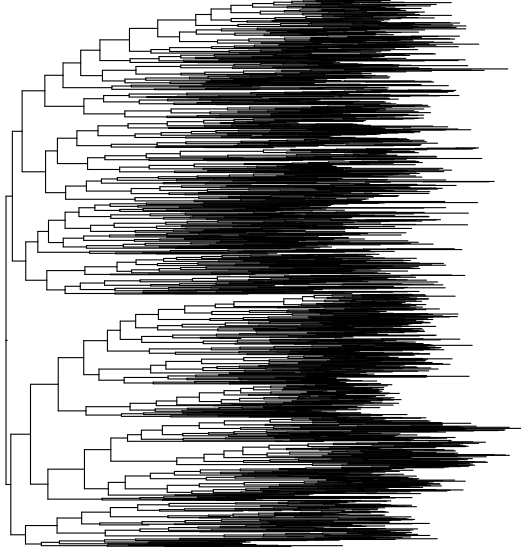


This study

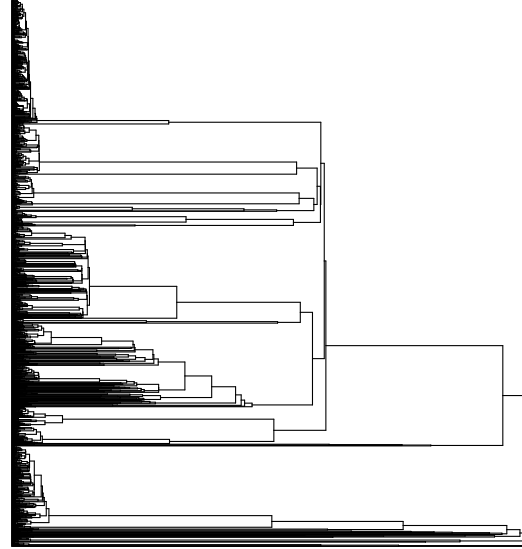
- systematic comparison of eight models:
 - two phylogenetic models with *Brownian motion* random effect for two trees
 - two phylogenetic models with *continuous time Markov chain (CTMC)* for two trees
 - hierarchical model with random intercepts for family and macro-area
 - GAM based on topographic distances, using adaptive splines, + random intercepts for family and macro-area
 - GP based on topographic distances, using adaptive splines, + random intercepts for family and macro-area
 - GAM based on Latitude/Longitude using splines on a sphere, + random intercepts for family and macro-area
- data:
 - 195 binary variables from Grambank
 - 1,500 languages
 - world tree inferred from Lexibank data + world tree from Boucaert et al. 2021
 - topographic distances from Guzmán Naranjo & Jäger 202
- goal: compare ΔAIC to assess the relative fit of the models to typological data

Phylogenetic trees

MSA Tree

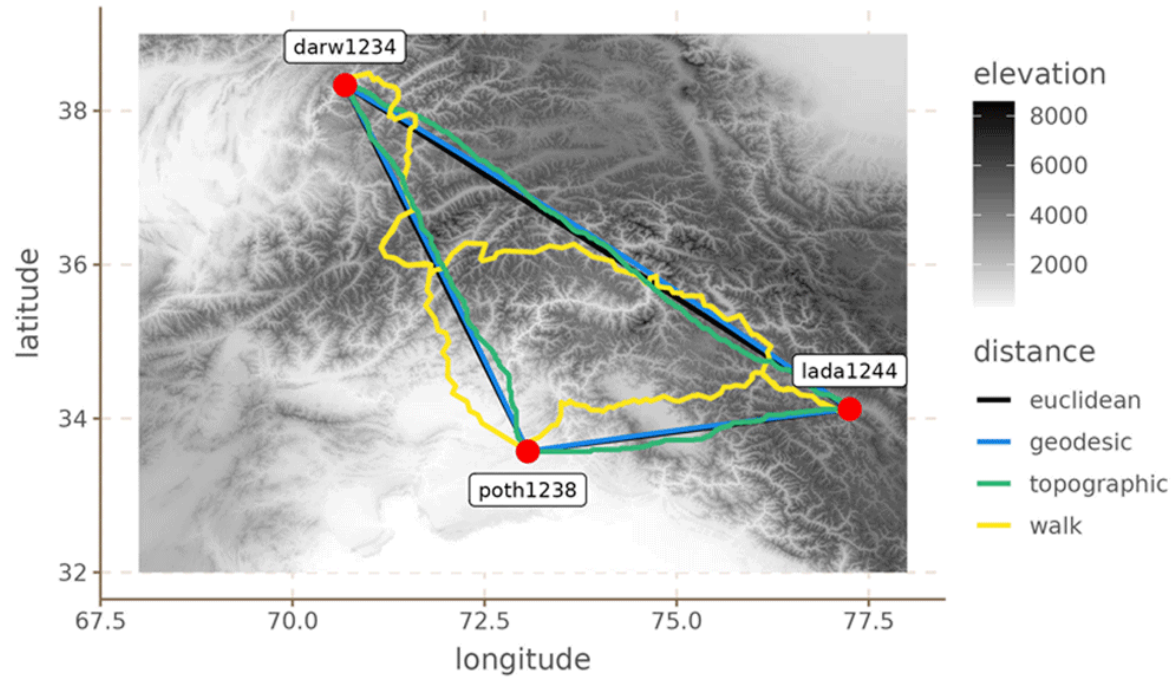


Edge Tree



Topographic distances

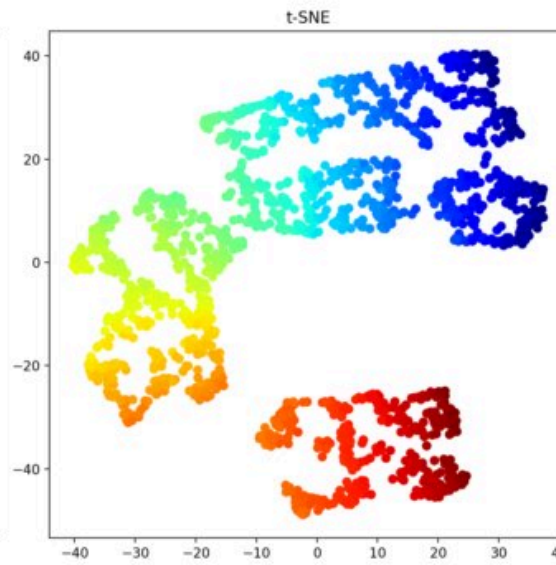
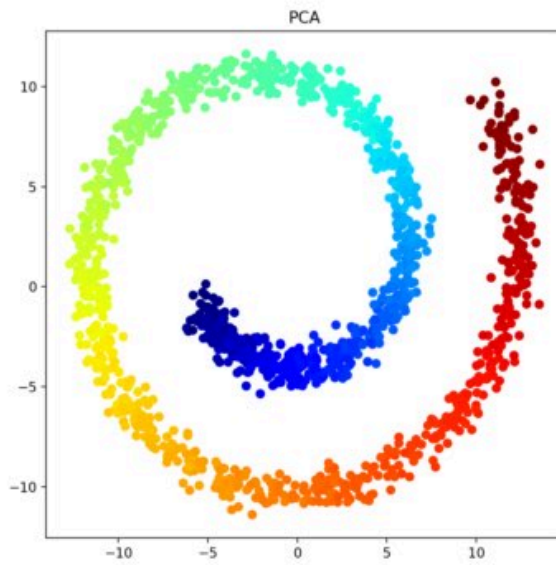
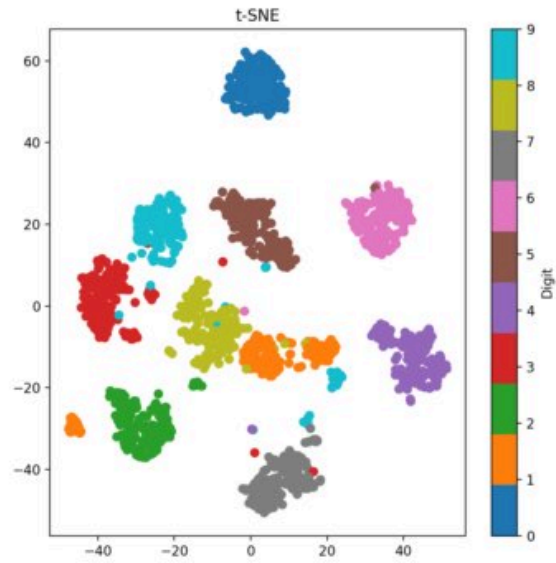
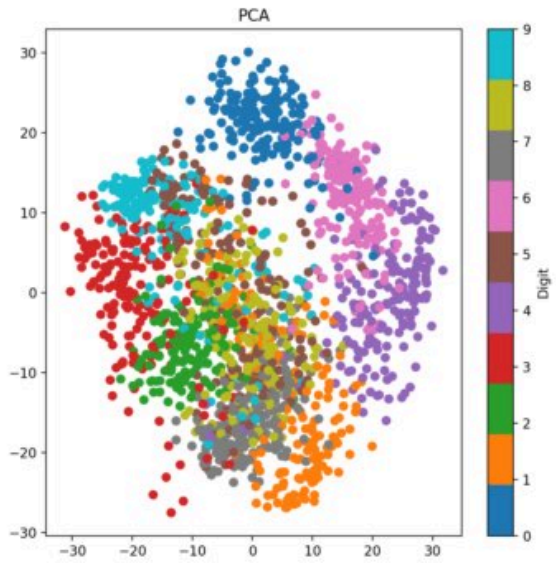
Guzmán Naranjo & Jäger 2024



- geodesic = *as the crow flies*
- topographic = *as the wolf runs*

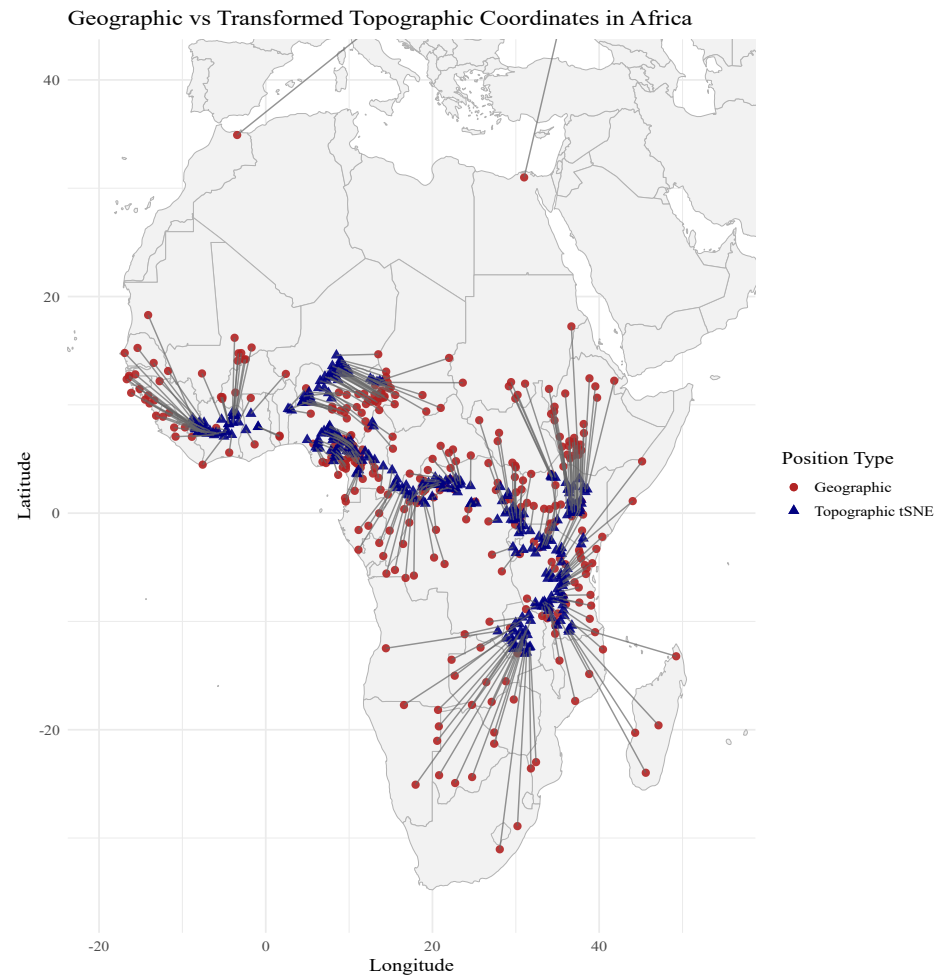
Mapping distances to two dimensions

- many methods for dimensionality reduction
- standard method: Multidimensional Scaling
- here: **t-Distributed Stochastic Neighbor Embedding (tSNE)**
 - popular in machine learning vor visualization
 - faithful mainly for small distances
 - large distances get distorted, but in an interesting way
 - neighborhood patterns are preserved



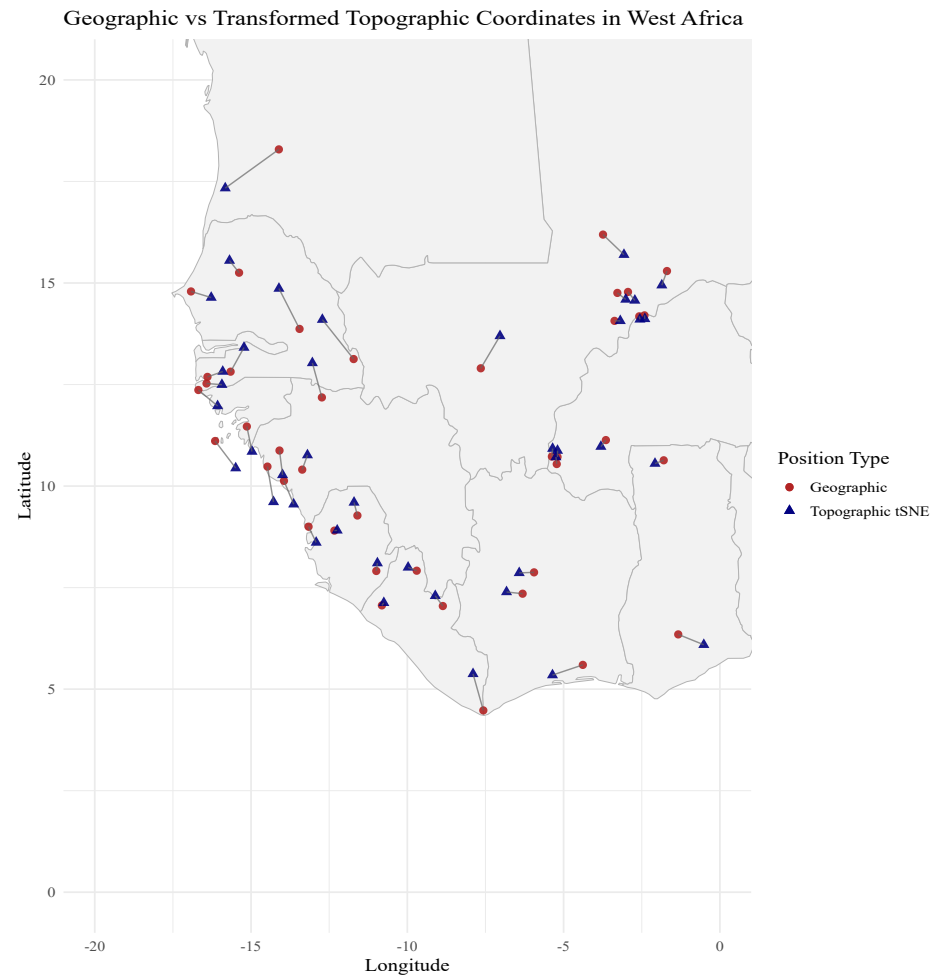
tSNE applied to topographic language distances

large scale



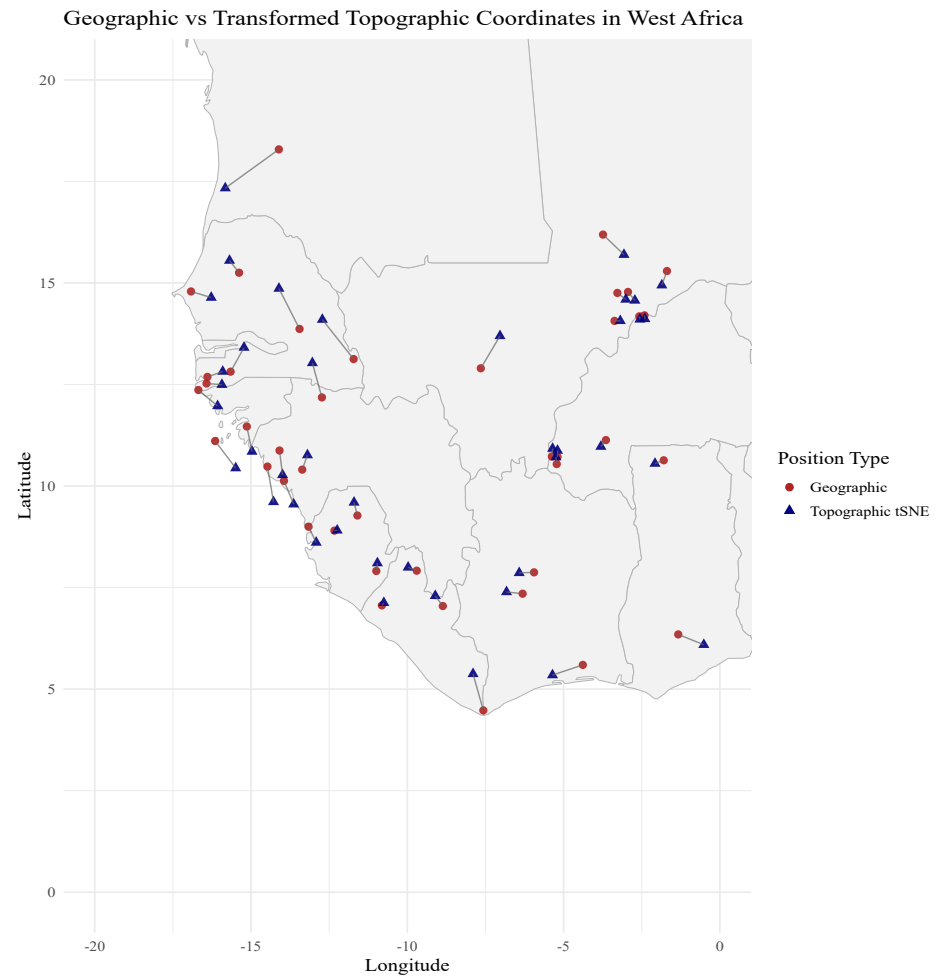
tSNE applied to topographic language distances

small scale

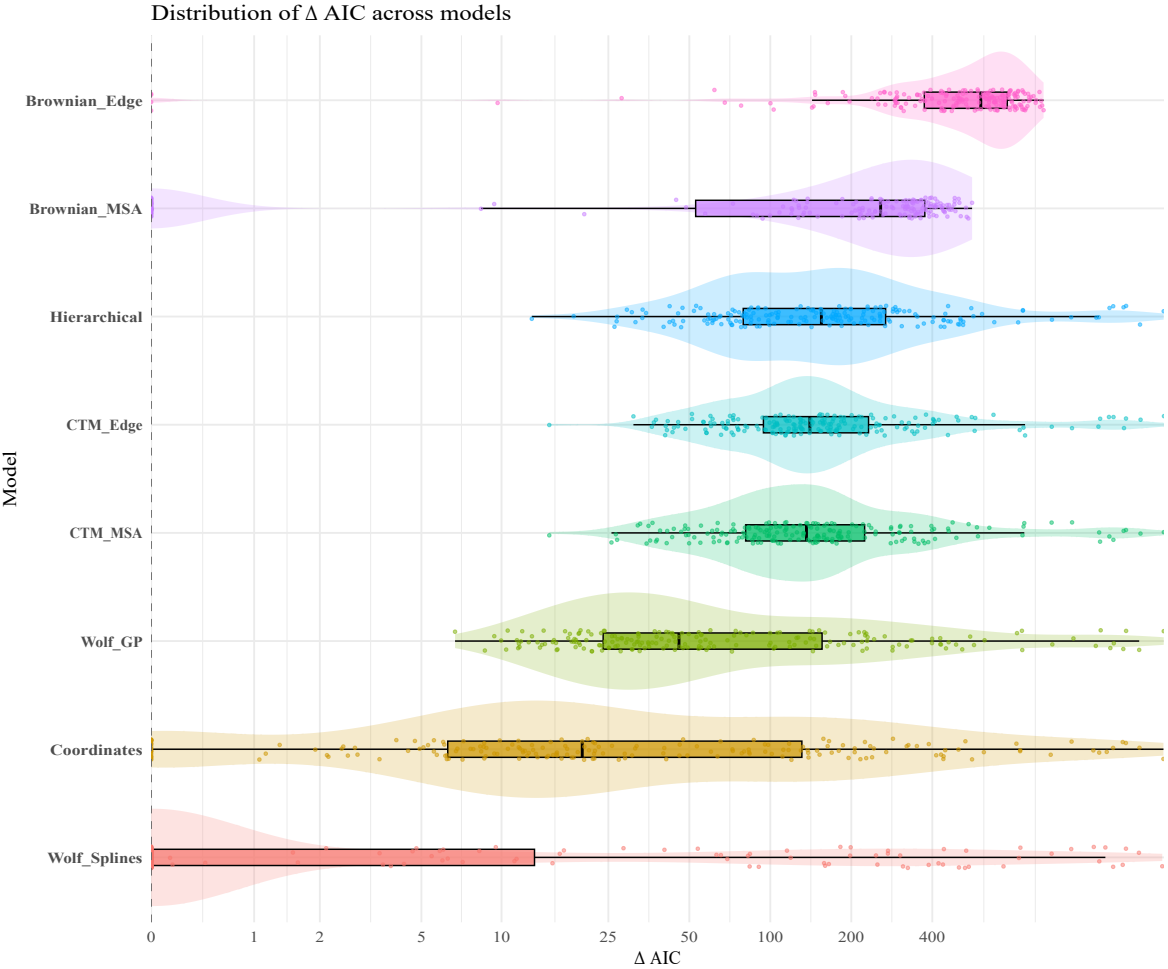


tSNE applied to topographic language distances

small scale



Results



Summary

Combination of

- topographic distances
- tSNE dimensionality reduction
- adaptive thin-plate splines
- intercepts for family and macroarea

seems to be quite powerful.

- still missing: phylogenetic methods + random effects (no off-the-shelf methods for ML inference I am aware of)
- further work
 - cross-validation
 - Bayesian implementation
 - application to correlation studies