Phoneme inventory size in Polynesian languages

Matías Guzmán Naranjo, Laura Becker & Miriam Schiele

25.08.2022

Phoneme inventory sizes: some hypotheses

Some proposals in the literature include:

- **Population size** (Trudgill 2004; Hay and Bauer 2007; Nettle 2012; Atkinson 2011; Wichmann, Rama, and Holman 2011; Fenk-Oczlon and Pilz 2021)
- **Serial founder effect** or **Distance** from Africa (Atkinson 2011; Wichmann, Rama, and Holman 2011)
- **Contact** with other languages & **isolation** (Trudgill 2004; Trudgill 2011; Nichols 1992; Haudricourt 1961; Elbert 1965; Rivierre 1994; Ozanne-Rivierre 1994)

As far as we are aware, however, these have not been tested with realistic models.

Polynesian languages

Why focus on Polynesian languages?

- "Andy Pawley [...] paints a very nice picture of Polynesians setting off in their canoes, throwing consonants overboard as they go. But is there anything which linguists can actually say about this?" (Trudgill 2004: 312)

→ Some have done so (Trudgill 2004; Hajek 2004; Elbert 1965; Rivierre 1994; Ozanne-Rivierre 1994)

- Relatively little variation in the phoneme inventories
- Large variation in population sizes
- Long distance spread
- Small number of languages, which allows us to be more precise

While in theory this question can be asked globally, more languages means less detail in the models.

Data: languages

We annotated a total of:

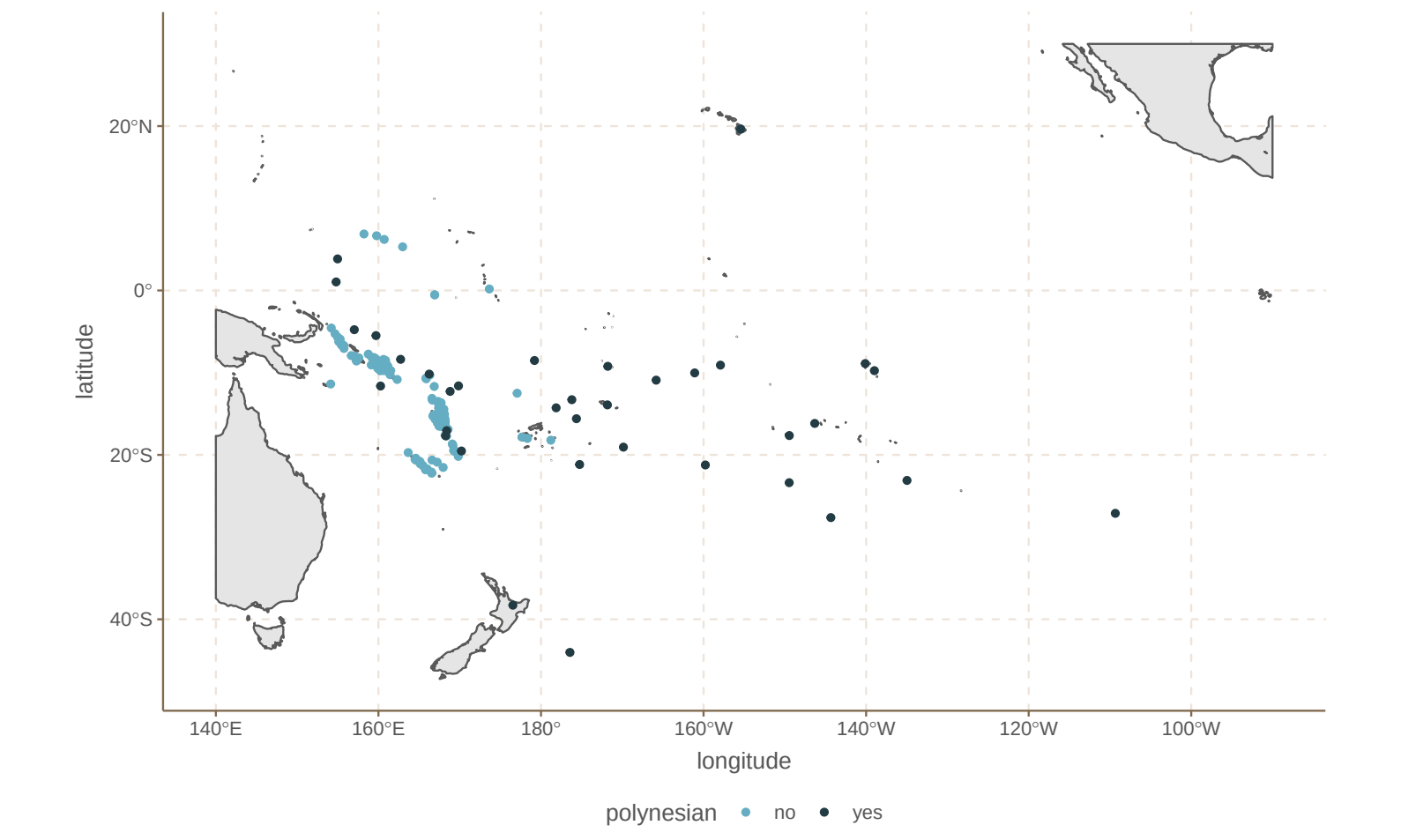
- 35 Polynesian languages
- 124 Non-Polynesian languages of the region

Additionally, we extracted:

- island in which it is spoken
- location (lon lat)
- population size

we did not use ready-made datasets.

Data: languages



Data: counting phonemes

Counting the number of phonemes in a language is not straightforward:

- inconsistent accounts across grammars?
- long and short vowels?
- long and short consonants?

We include length

Models

We use three techniques to control for inheritance and contact:

- Phylogenetic regression
- Gaussian Process
- Data imputation (with a GP)

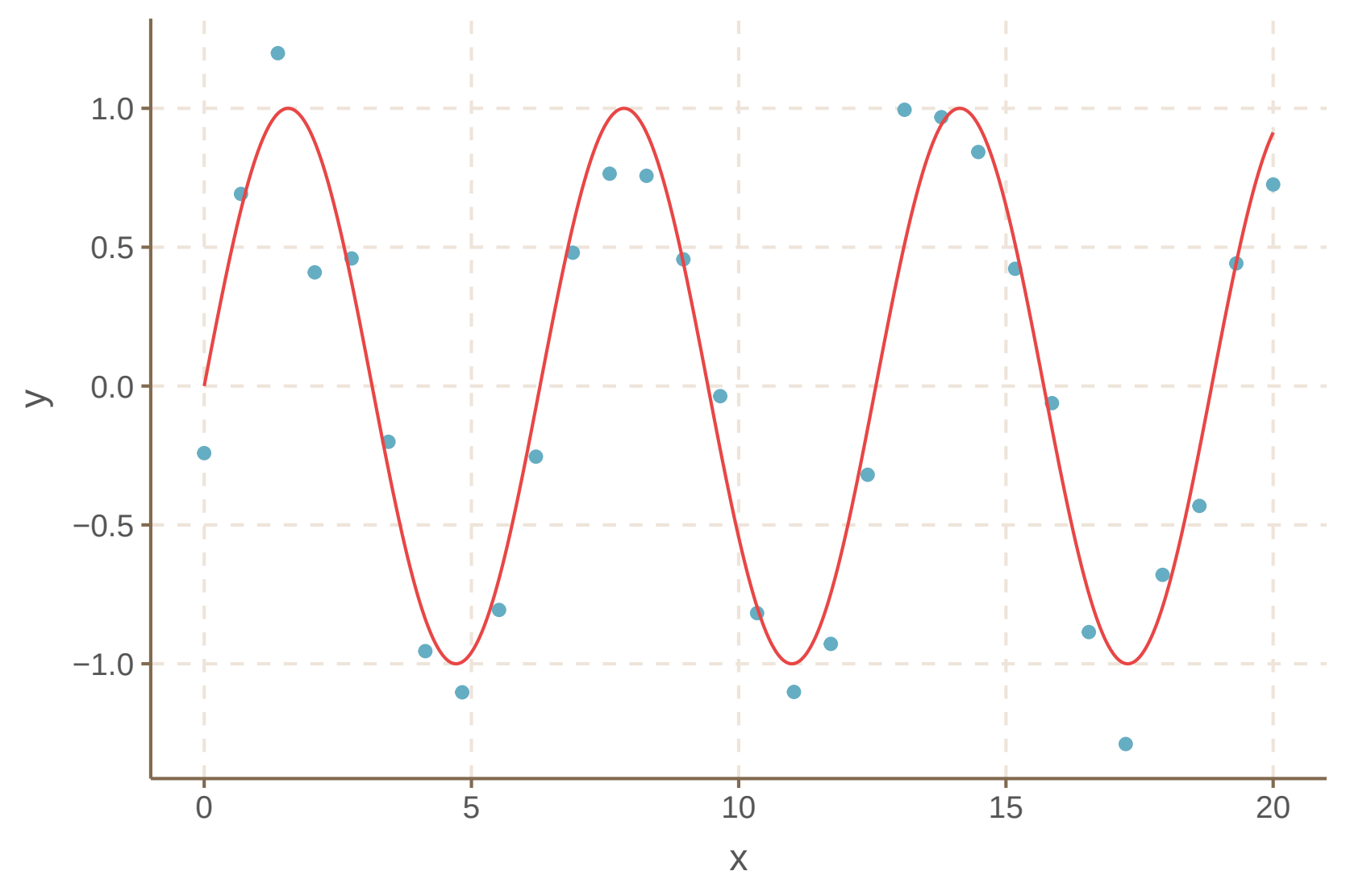
Phylogenetic regression

Phylogenetic regression is a way of controlling for genetic bias.

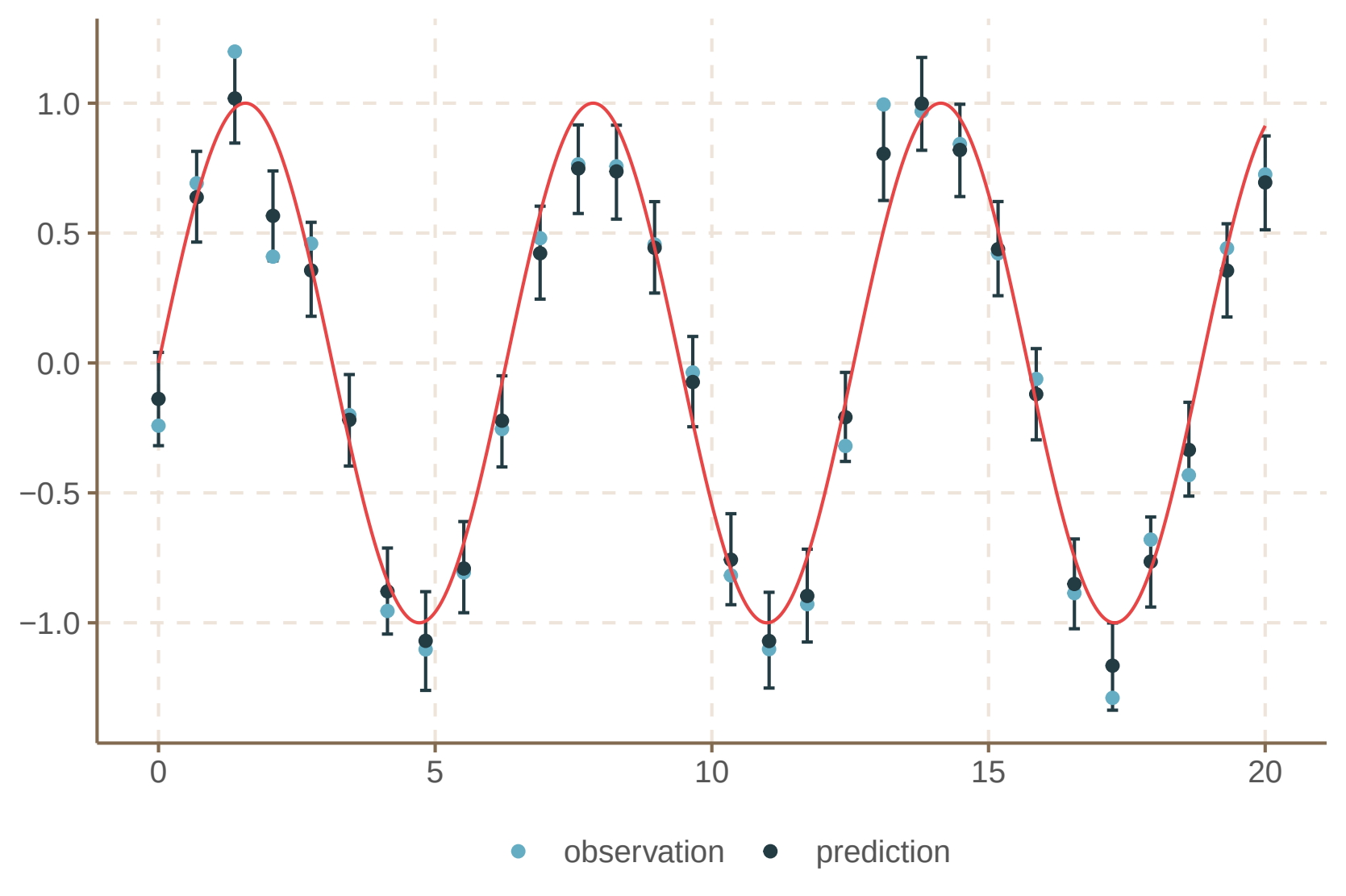
Unlike adding effects for family or genus, we add individual intercepts for each language.

However, we constraint these intercepts to co-vary according to a phylogenetic tree (Glottolog in our case).

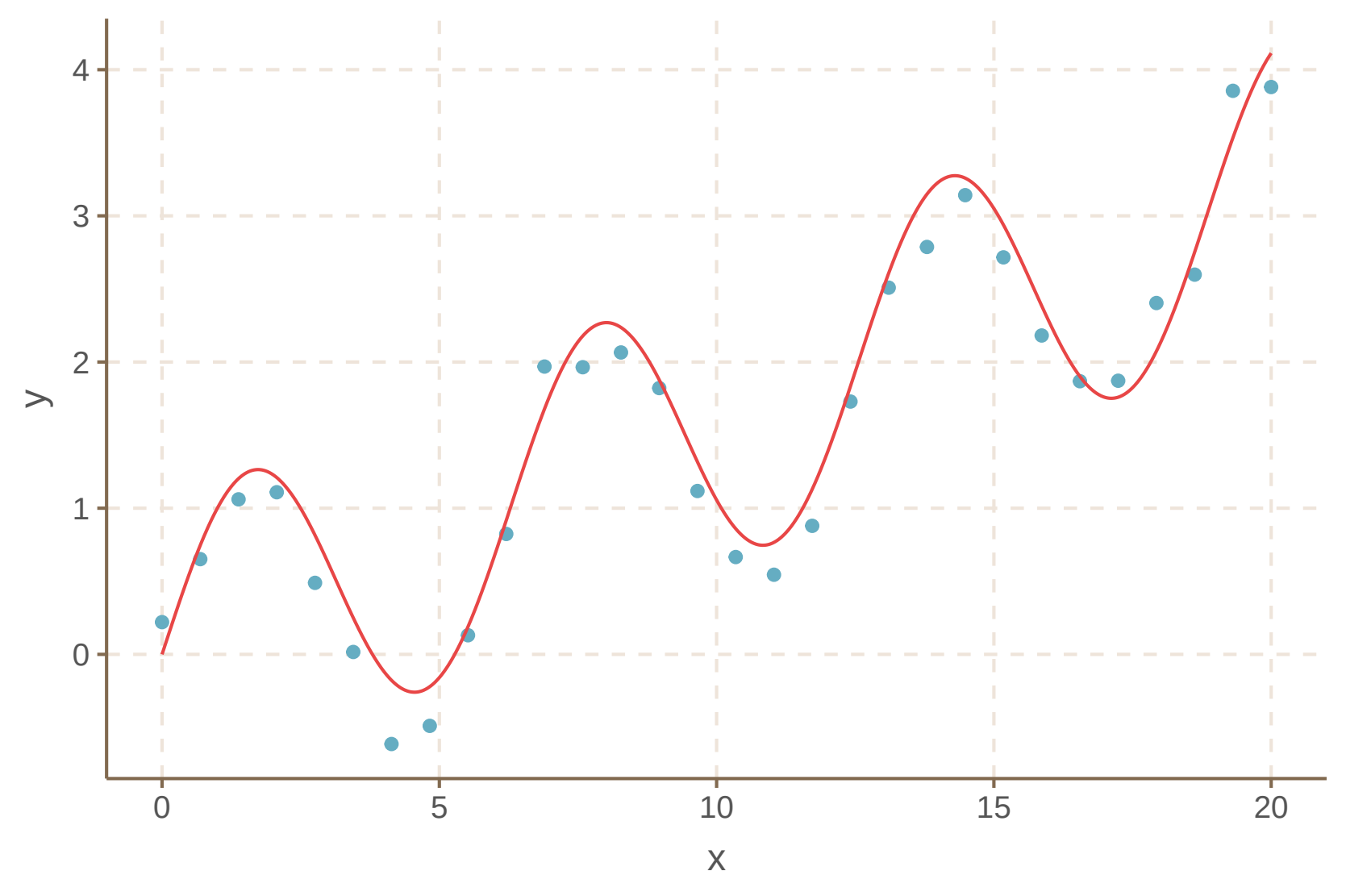
Gaussian process - stationary



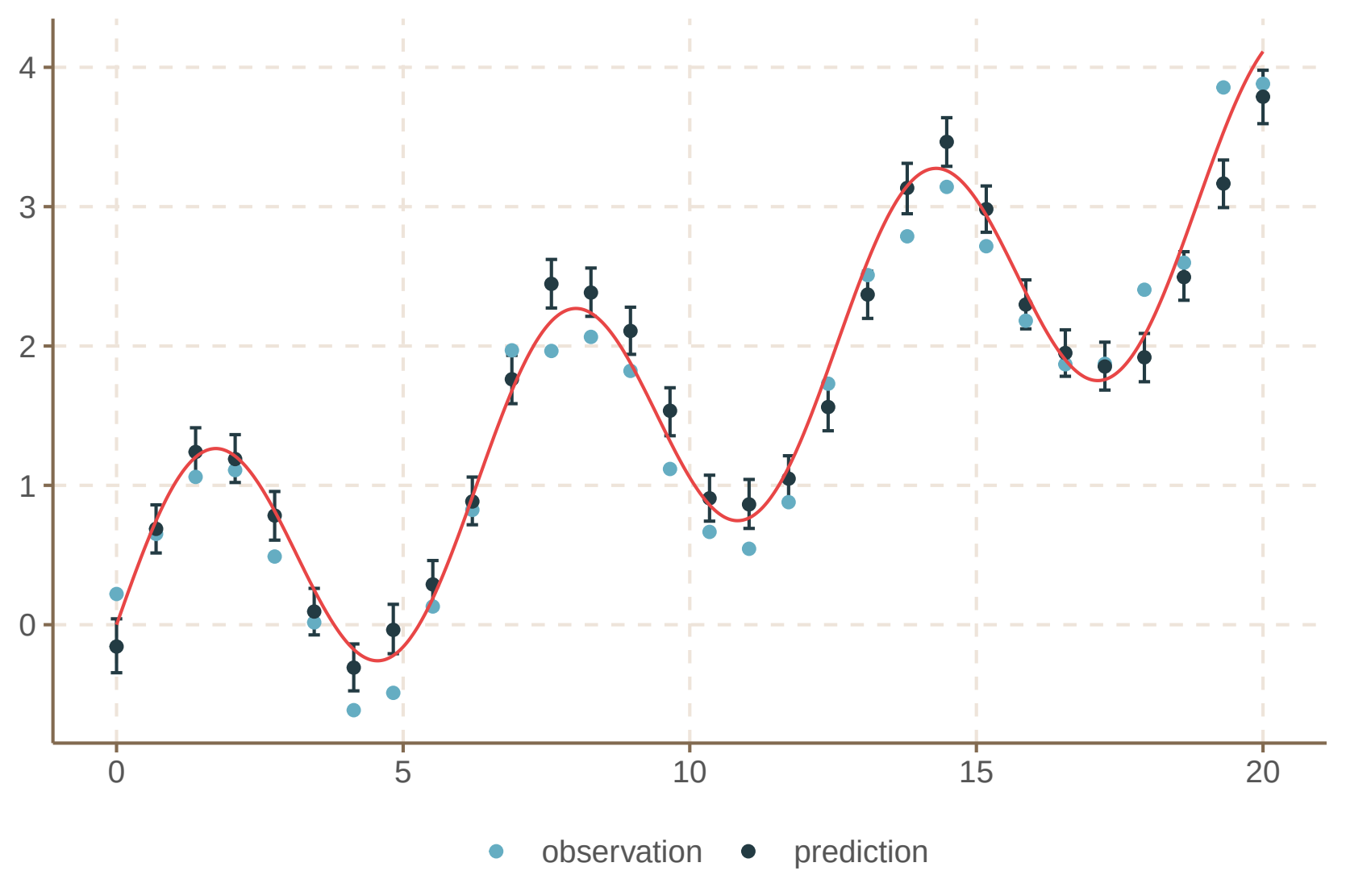
Gaussian process - stationary



Gaussian process - non-stationary



Gaussian process - non-stationary



Two types of contact

There are two types of spatial patterns.

- The PIS of my neighbour impacts my PIS
- The presence of phoneme /f/ in my neighbour, increases the chance of /f/ in my inventory, thus increasing the likelihood of a larger PIS in my language

The first case tends to be easy to model (Guzmán Naranjo and Becker 2021), the second case is harder, but a better representation of how contact works.

Contact features

We chose the following features as possible contact features which could affect PIS in Polynesian languages (others could be added):

Language has :

- vowel length
- consonant length
- /v/
- /f/
- /s/
- /l/
- /ɾ/
- /ʔ/
- /r/

While these phonemes were present in PP, they either disappeared in several languages, or changed to different ones.

Modelling indirect contact

To model indirect contact we did the following:

For all Non-Polynesian languages in our sample, we:

- Build a model with a GP predicting one of the features in question (e.g. has /f/)
- Predict the expected value of the Polynesian languages using the model fitted to Non-Polynesian languages
- We use these predicted values as predictors in the main model

This should capture the indirect effect of Non-Polynesian languages on Polynesian language PIS

Models: linear predictors

We have thus the main following predictors/controls:

- log population size
- neighborhood size (how many near neighbors a language has)
- indirect contact features (as per above)

Plus distance from potential Urheimat:

- distance from Samoa
- distance from Tonga

(Kirch and Green 2001: 100)

Models: specification

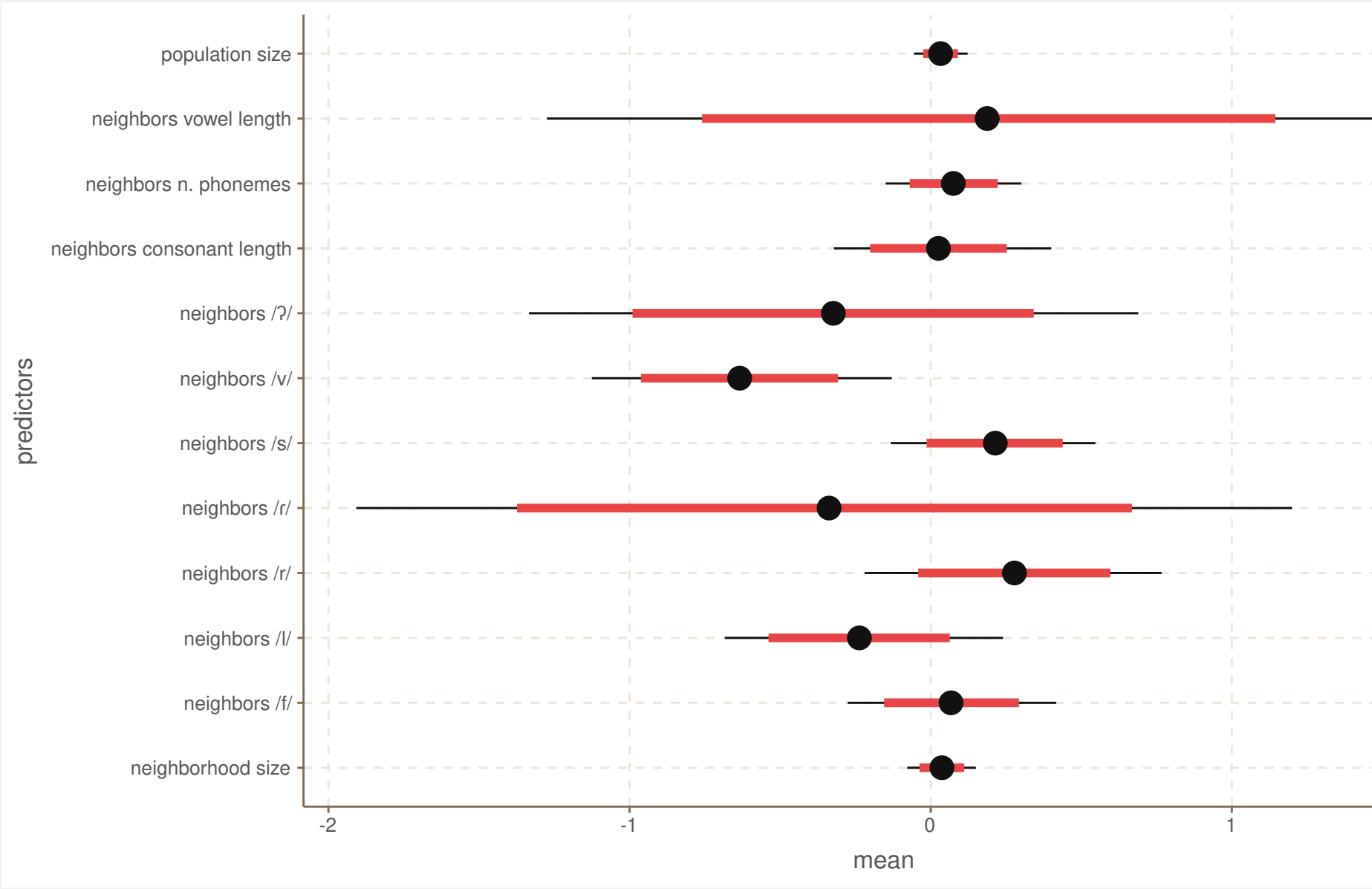
We tested the following models:

- stationary model + no linear predictors
- stationary model + linear predictors + distance from the Urheimat
- non-stationary model + no linear predictors
- non-stationary model + linear predictors

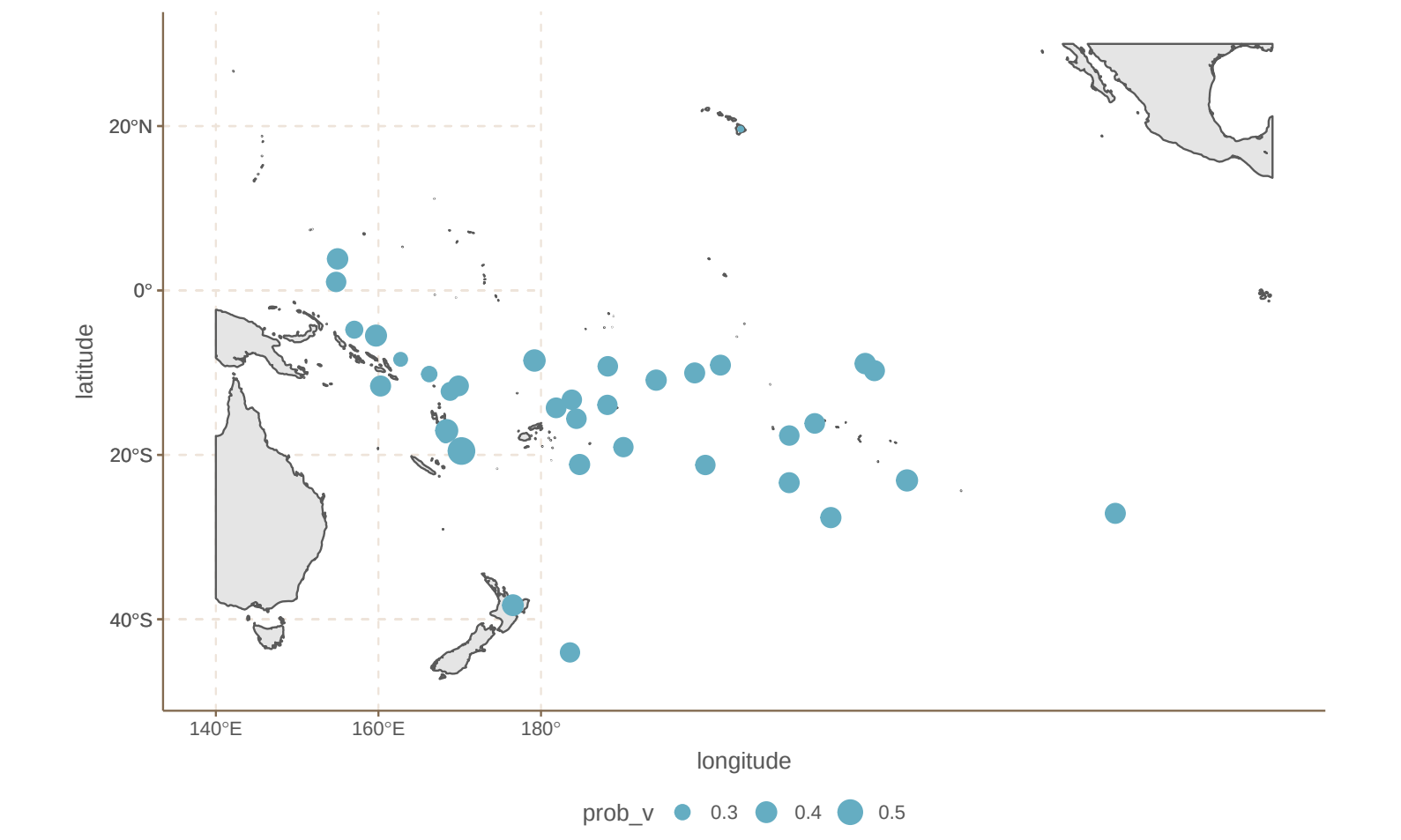
We used a Poisson likelihood for all models.

Coefficients

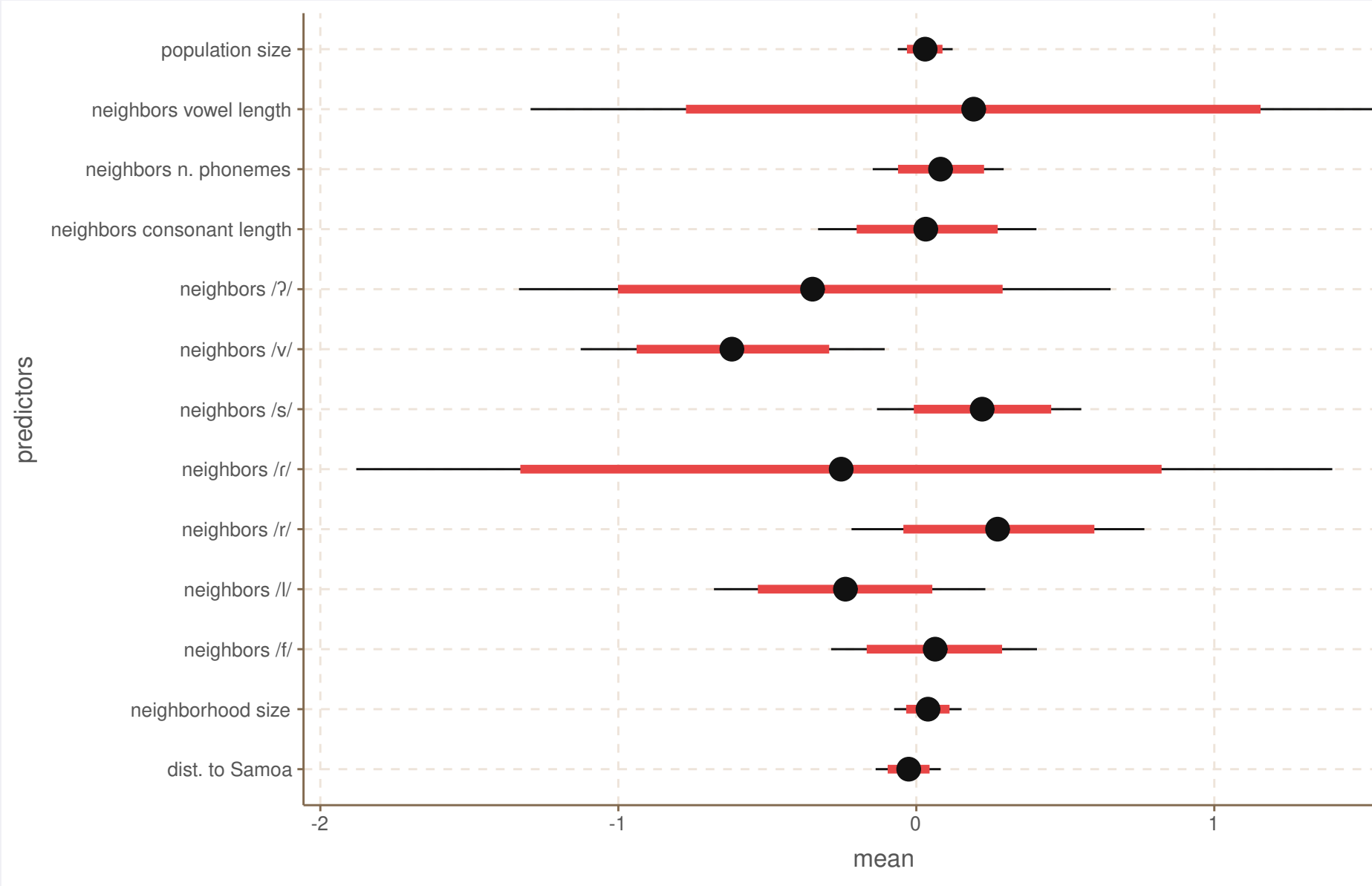
Stationary + linear predictors



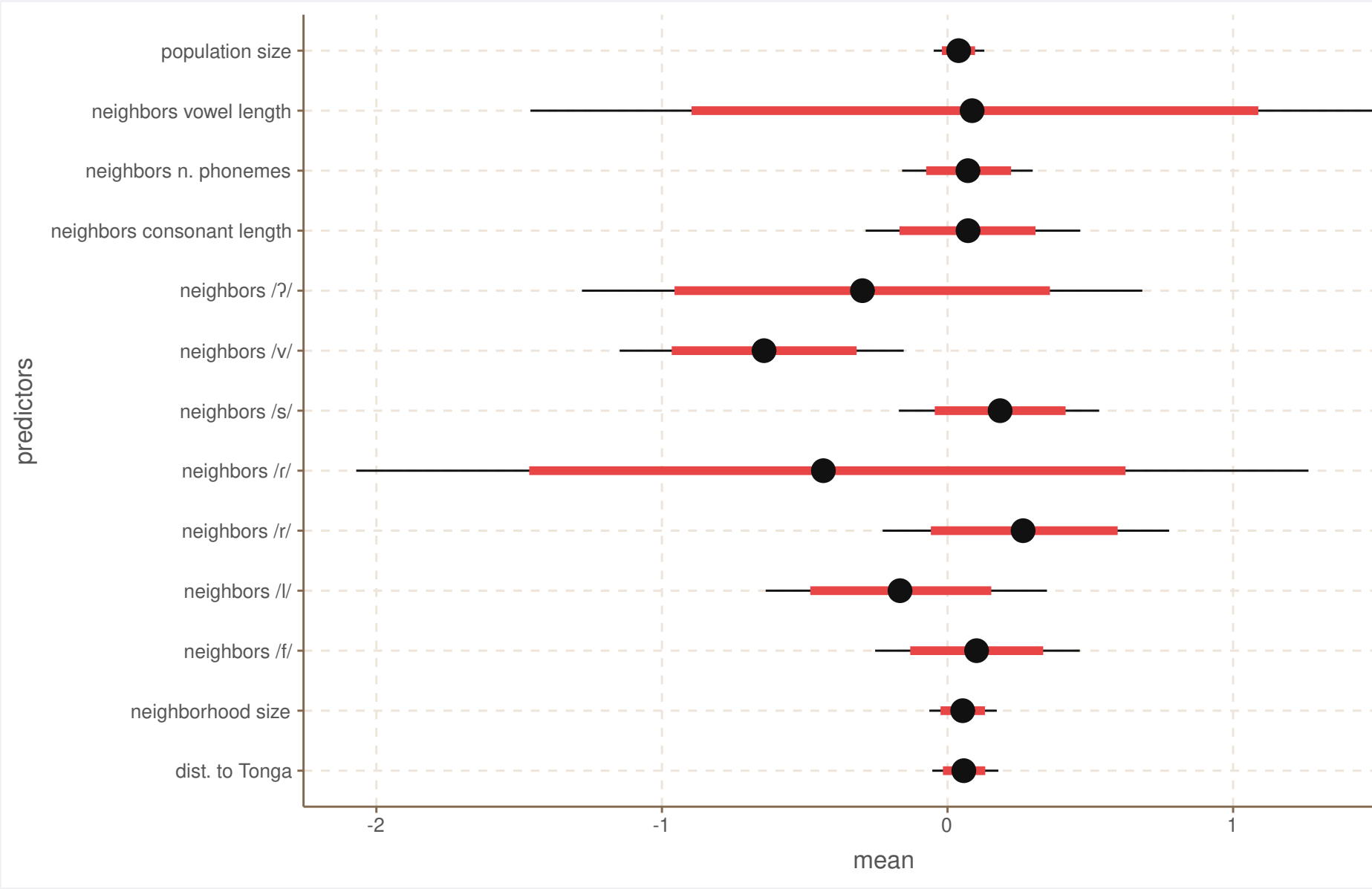
Prob /v/



Stationary + linear predictors + distance to Samoa

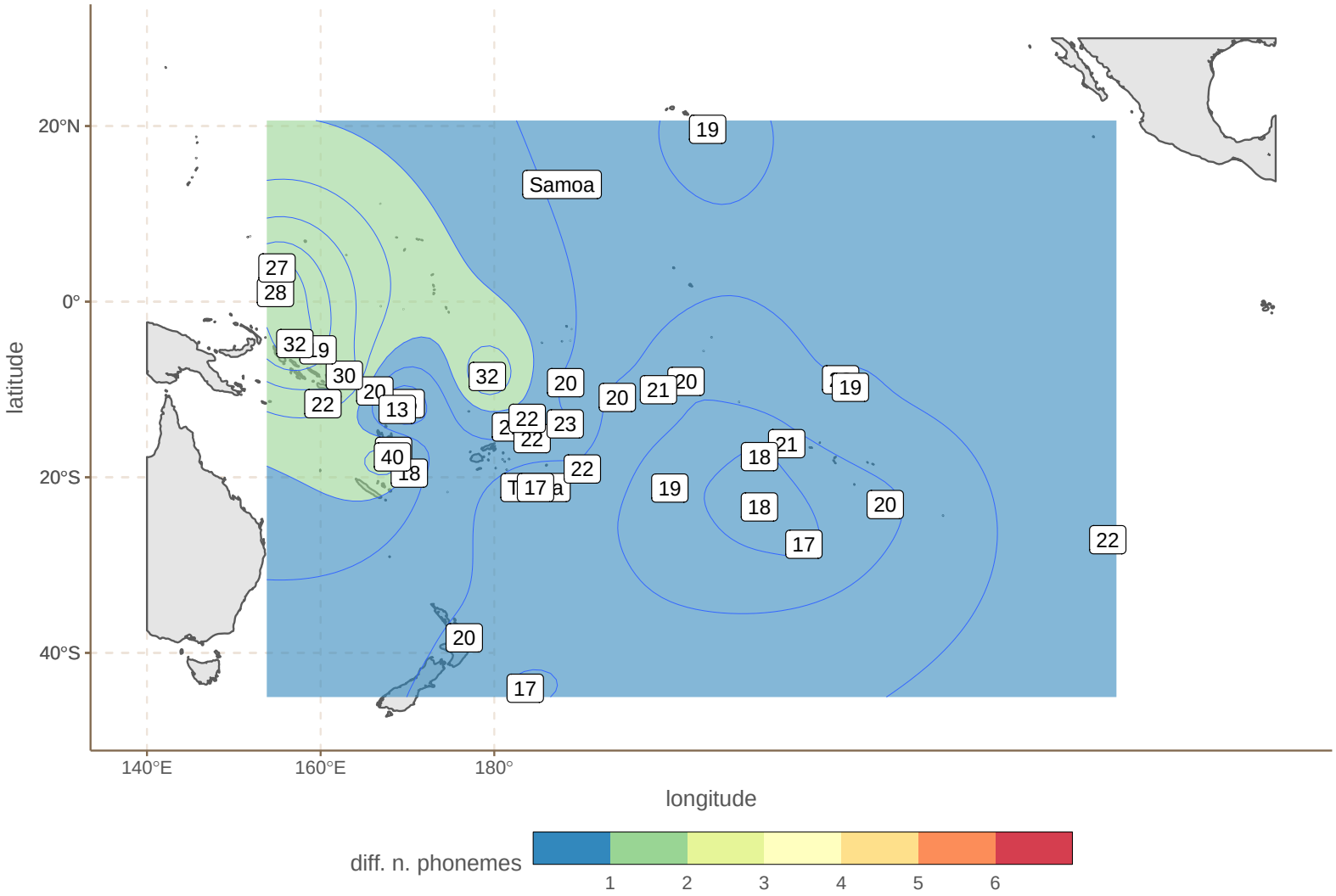


Stationary + linear predictors + distance to Tonga

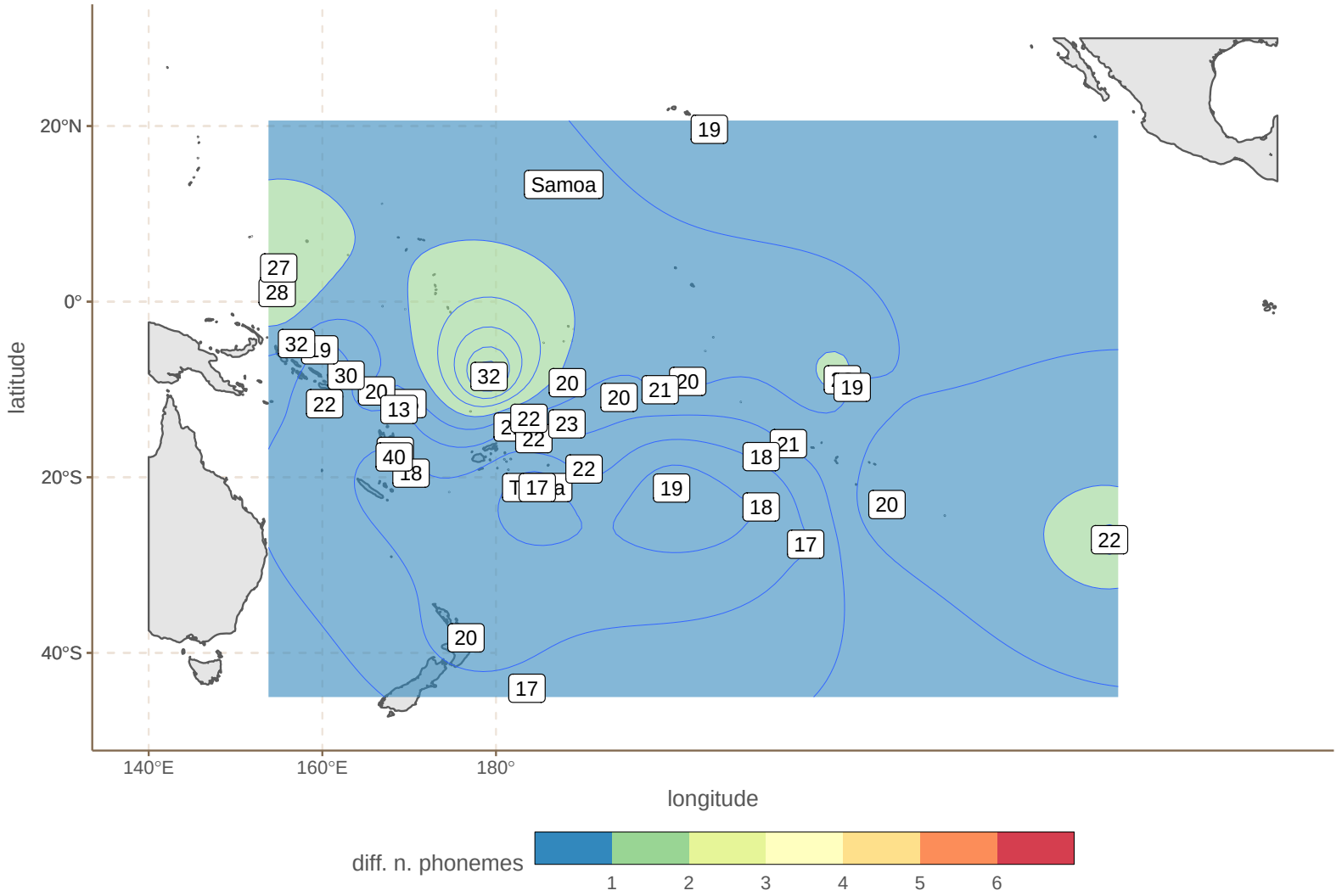


Spatial effects

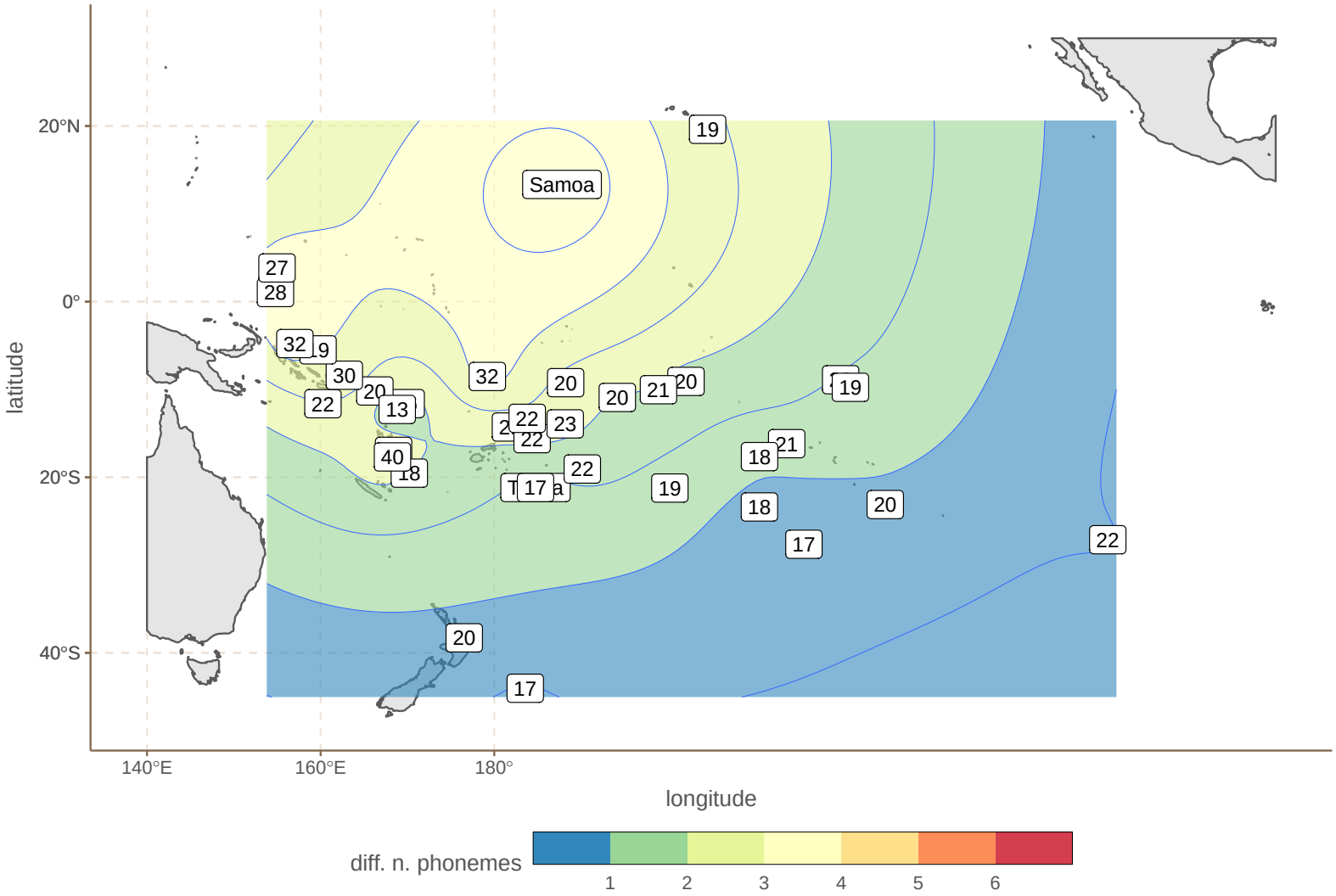
Spatial effects - stationary



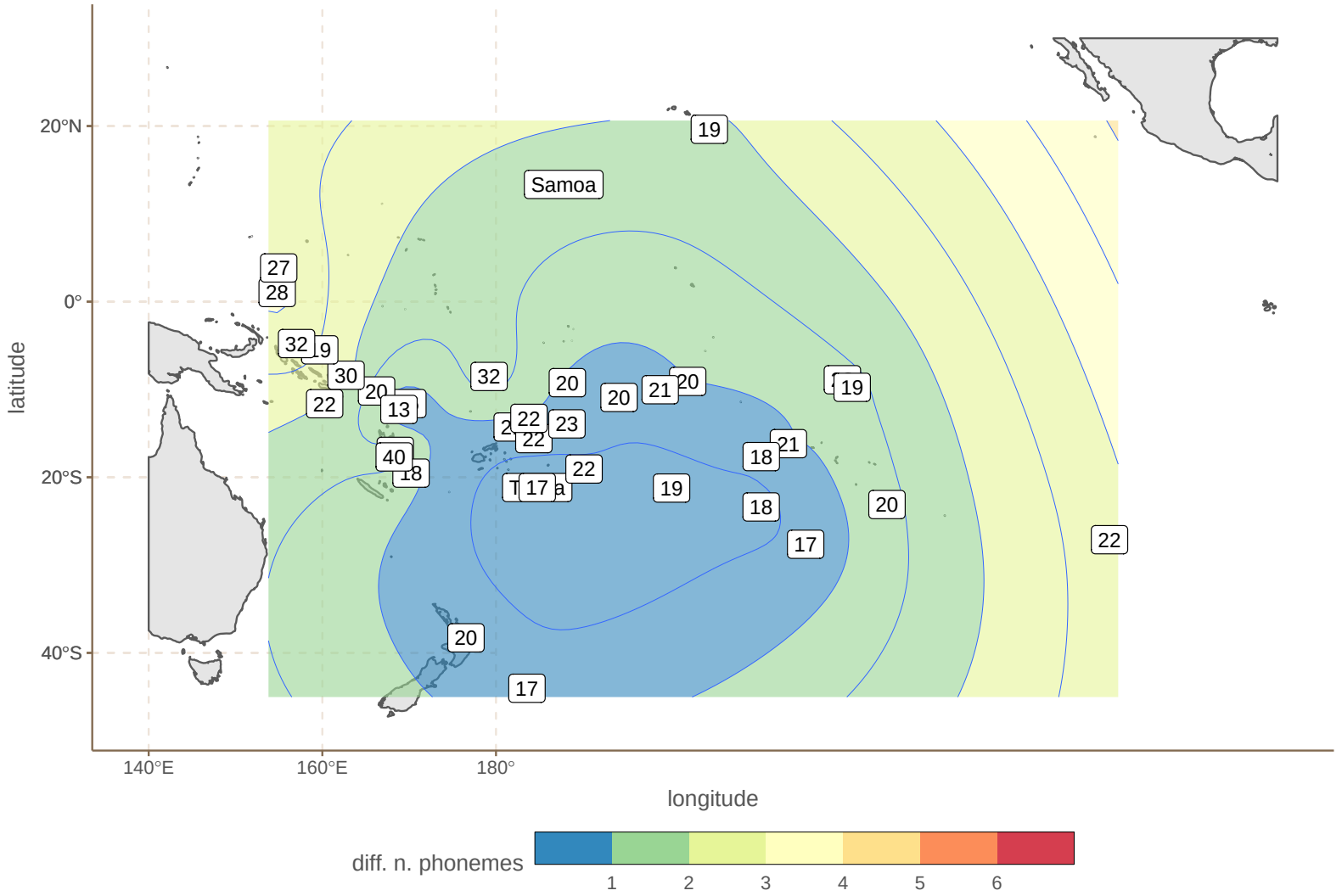
Spatial effects - stationary + linear predictors



Spatial effects - non-stationary Samoa



Spatial effects - non-stationary Tonga



Leave-one-out cross-validation

RMSE

However, the simpler models perform better in predictive terms:

model	RMSE
stationary	5.47
non-stationary tonga	5.50
non-stationary samoa	5.52
non-stationary tonga + all preds	9
non-stationary samoa + all preds	9.48
stationary + all preds	10
stationary + all preds + dist to tonga	10.12
stationary + all preds + dist to samoa	10.3

We also see no evidence for the non-stationary term.

Discussion

We have shown that:

- Population does not seem to have a direct impact
- Distance from the Urheimat does not have an impact

→ observed effects (Atkinson 2011) for the distance to Africa are likely mediated by population sizes (Donohue and Nichols 2011; Wichmann, Rama, and Holman 2011)

- There is likely some contact-induced effect
- There might be some indirect contact effects, but it is far from straight-forward

In terms of methods

We proposed two new techniques:

- A ways of controlling for indirect contact effects by using data imputation
- A way of modelling directional spatial relations using non-stationary GPs

We would welcome further testing of these approaches on new data...

To do...

We still have some things to do:

- Single model (neighbourhood size+imputation+main model sampled simultaneously)
- Known contact relations (which we annotated but haven't included)
- Ethnographic features (type of economy, social structure, etc.)
- Likely path taken form the Urheimat instead of how-the-crow-flies distances

Thank you!