Inclusive sampling and bias control in language typology

Matías Guzmán Naranjo & Laura Becker

25.02.2021

Sampling in typology

There are two families of sampling methods:

Probability sampling

- build a sample to draw conclusions about crosslinguistic distribution of the values of a given feature (combination)

☞ languages have to be as independent of each other as possible

Bell (e.g. 1978), Bickel (2008, 2011), Dahl (2008), Dryer (1989), Nichols (1992), and Perkins (1980, 1989)

Variety sampling

- build a sample to capture all possible values of a given feature (combination)

☞ including as many languages that are independent from each other is assumed to capture more variation

Miestamo (2005), Miestamo, Bakker, and Arppe (2016), Rijkhoff and Bakker (1998), and Rijkhoff, Bakker, et al. (1993)

Sampling in typology

- select a sample of languages so that languages (as trials) are as independent from each other as possible

☞ "Typologists know it is crucial to **control** for the **non-independences** in a dataset that stem from language **areas** and language **families** (e.g., Dryer, 1989, 1992). The **best remedy** for an areally and genealogically biased typological analysis is to **balance** the sample with respect to families and areas." (Bentz, Verkerk, et al., 2015, p. 19)

Controlling for genetic bias

- include a limited number of languages of the same genealogical / cultural grouping

☞ We have no way to know how accurate genetic control really is.

- sample genera instead of languages (including the variation within genera) (Bickel, 2008; Dryer, 1989; Sinnemaki, 2014)

☞ How do we deal with isolates, creoles, and sign languages?

Controlling for geographic bias

- include a limited number of languages from the same area

- Dryer (1989) and Hammarström and Donohue (2014): division into **6 macro areas** that are physically disconnected enough to be treated as independent units

☞ Controlling for geographic biases comes with similar issues as controlling for genetic bias.

Sampling often means reducing or restricting

- Both types of sampling methods try to include languages that are as independent from each other in order to avoid the biases mentioned.

☞ This often leads to either **reducing** the number of languages in the sample or to **restricting** the sample.

In modelling

As far as we can tell, when doing typological modelling, there are two main approaches to dealing with biases:

1. no statistical controls: bias is controlled through sampling
2. simple statistical control: family and geographic effects are controlled with (random) effects in a model (Bentz and Winter, 2013; Cysouw, 2010; Jaeger et al., 2011; Levshina, 2019)

- We believe that both are problematic.

Issues with the modelling approach

We see three main issues with the modelling approach:

- including `family` as a an effect in a model ignores the fact that there is structure within each family, and connections above it
- including `(macro)area` as an effect does not really account for variation between macro areas or across micro areas
- distance between languages is relative and depends on the population density:

☞ 100 km in Siberia are not the same as 100 km in the Amazonas

Our proposal for family bias

We want to account for the fact that language families are trees.

- we do not include any cut-off point in our model, but rather a whole phylogenetic term (PT)
- a PT includes information about all relations between the languages in the sample:
- e.g. **Spanish** is more closely related to **Catalan** than to **Italian**, but these three are closer to each other than to **German**
- this way, the model estimates effects for micro-families which must respect the phylogenetic distances

Phylogenetic term

	Hindustani	Global German	Global Dutch	Castillic Spanish	Global French	Italian Romance	Fulniô	Nyulnyulan
Hindustani	1.00	0.67	0.67	0.67	0.67	0.67	0	0
Global German	0.67	1.00	0.83	0.67	0.67	0.67	0	0
Global Dutch	0.67	**0.83**	1.00	0.67	0.67	0.67	0	0
Castilic	0.67	0.67	0.67	1.00	0.91	0.90	0	0
Global French	0.67	**0.67**	0.67	**0.91**	1.00	0.90	0	0
Italian Romance	0.67	0.67	0.67	0.90	0.90	1.00	0	0
Fulniô	0	0	0	0	0	0	1.00	0
Nyulnyulan	0	0	0	0	0	0	0	1.00

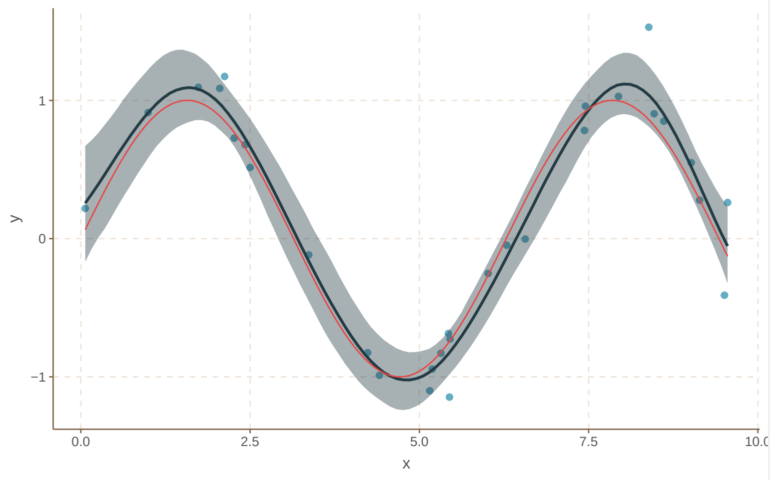
Our proposal for geographic bias

Glottolog (Hammarström, Bank, et al., 2018) has (approximate) geographic information for each language in the form of latitude and longitude.

With this information,

- we add a surface to our model which includes the latitude and longitude information of each language
- the model estimates whether there are regions in the map that are strongly associated with the response variable

Gaussian process



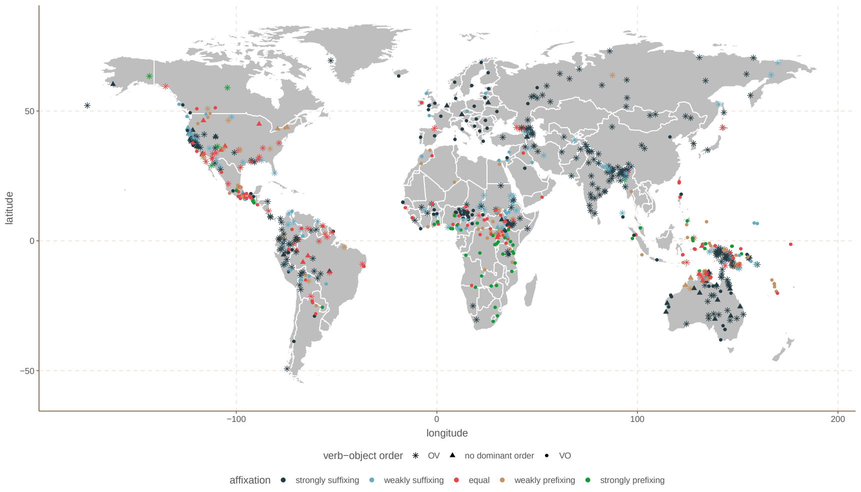
Phenomenon and dataset

We will focus on one specific example:

affix position and its association with **verb-object order**

- we use the data in WALS and Glottolog (Dryer and Haspelmath, 2013; Hammarström, Bank, et al., 2018)
- our dataset contains a total of 778 languages

Dataset


50
0
−50
latitude
−100
0
100
200
longitude
verb−object order
OV
no dominant order
VO
affixation
strongly suffixing
weakly suffixing
equal
weakly prefixing
strongly prefixing

Affixation and word order

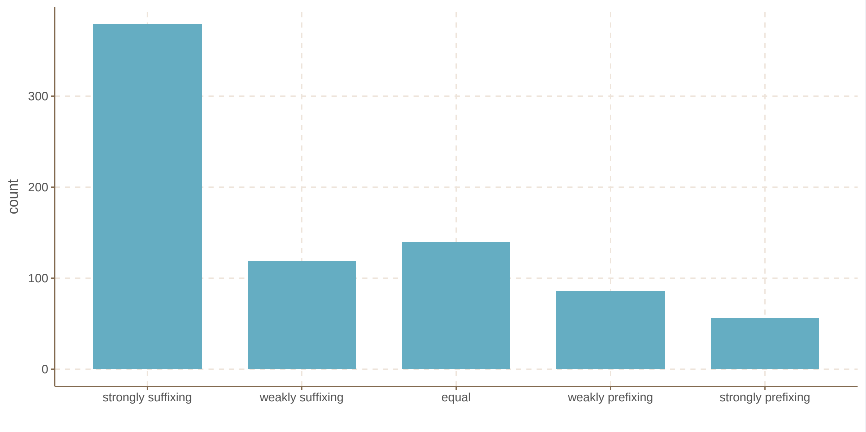
OV: strong preference for suffixation
VO: both prefixation and suffixation
(Bybee, Pagliuca, and Perkins, 1990; Cutler, Hawkins, and Gilligan, 1985; Dryer, 1992; Siewierska and Bakker, 1996)

We also know that the position also strongly depends on the type of affix. (Bybee, Pagliuca, and Perkins, 1990; Cysouw, 2009; Dryer, 1992)

There are different types of explanations:

- synchronic, cognitive motivations involving ease of processing (e.g. Hawkins and Gilligan, 1988)
- diachronic explanations based on the processes leading to (different types of) affixes (Bybee, Pagliuca, and Perkins, 1990; Himmelmann, 2014; Siewierska and Bakker, 1996)

Global distribution of affix positions



The main model

We predict affixation (as ordinal) from:

- verb-object order
- 2D gp(longitude, latitude)
- phylogeny

```
affixation ~ vo-order + gp(lat, lon) +
(1|microfamily, cov = phylogeny)
```

The hierarchical model

In addition, we fit a hierarchical model predicting affixation (as ordinal) from:

- verb-object order
- group-effect for family
- group-effect for macro area

```
affixation ~ vo-order + (1|family) + (1|macroarea)
```

The no-controls model

We also fit a model without controls, predicting affixation (as ordinal) from:

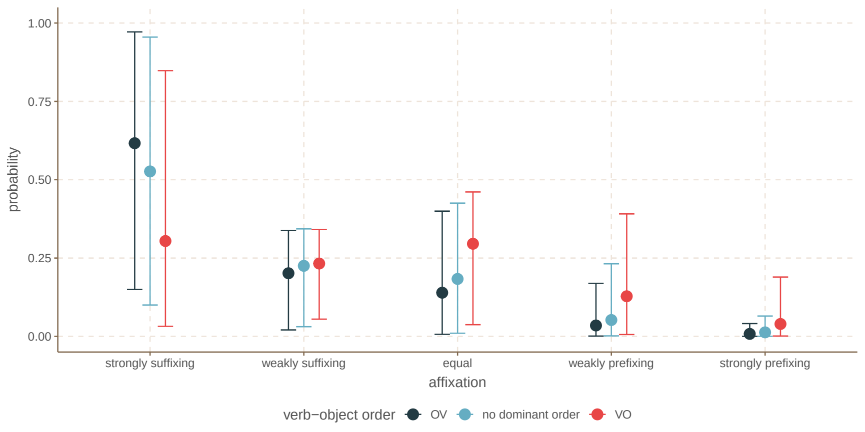
- verb-object order

`affixation ~ vo-order`

Results

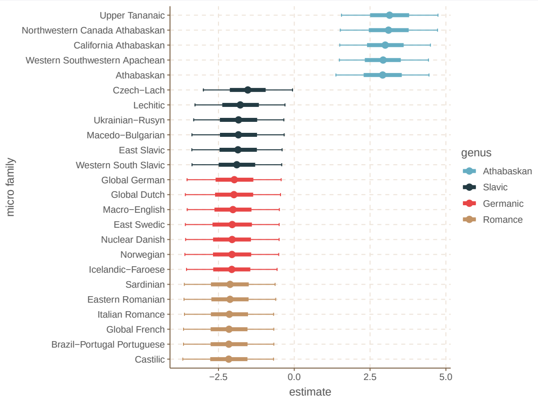
The main model

Effects of verb-object order

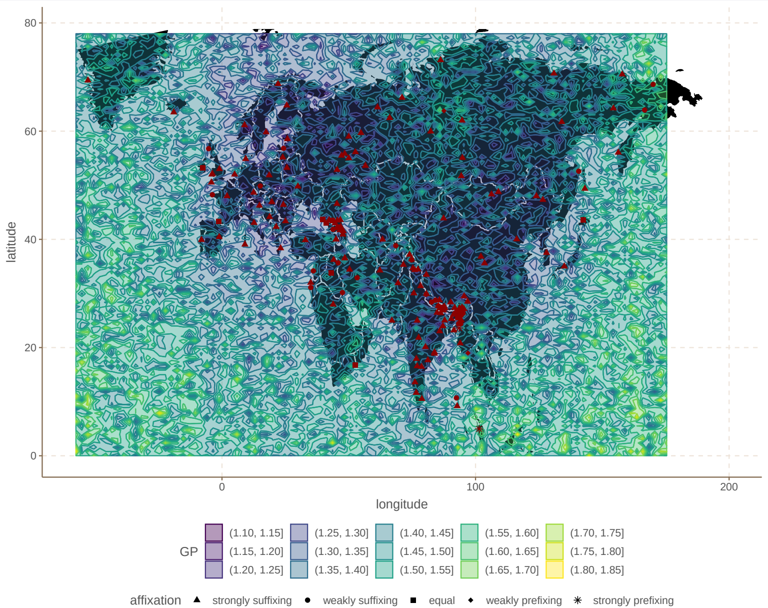


The main model

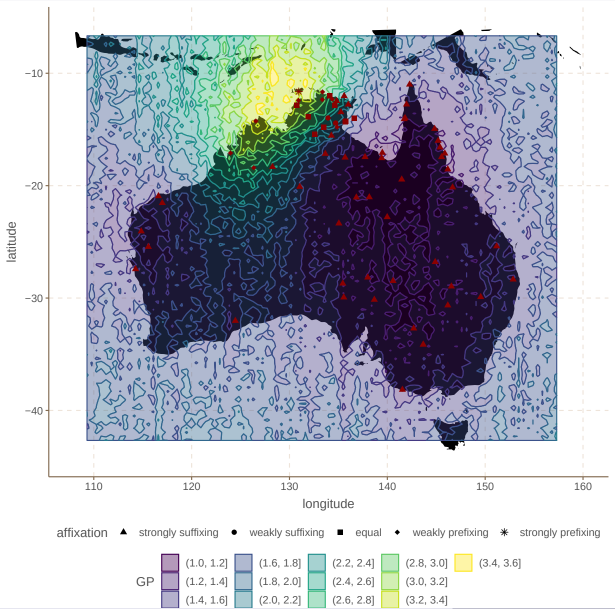
Phylogenetic term



The main model: geographic effects (Eurasia)



The main model: geographic effects (Australia)



Model performance

The main model

We carried out approximate Leave-One-Out cross-validation of the model.

	reference				
prediction	strongly suffixing	weakly suffixing	equal	weakly prefixing	strongly prefixing
strongly suffixing	**230**	24	7	1	0
weakly suffixing	124	**66**	47	16	5
equal	23	26	**70**	47	18
weakly prefixing	2	3	16	**20**	28
strongly prefixing	0	0	0	2	**6**
Accuracy			0.5		
Kappa			0.32		
rmse			0.88		

Interim results

With the main model (and the data and prior assumptions) we see that:

- suffixation is clearly much more common
- the verb-object order **very** is mildly associated with affix position:

☞ `OV` strongly prefers `strong suffixation`

`VO` allows for more `prefixation`

- but the uncertainty intervals suggest that the effect is likely due to chance
- **there are very strong geographic effects!**

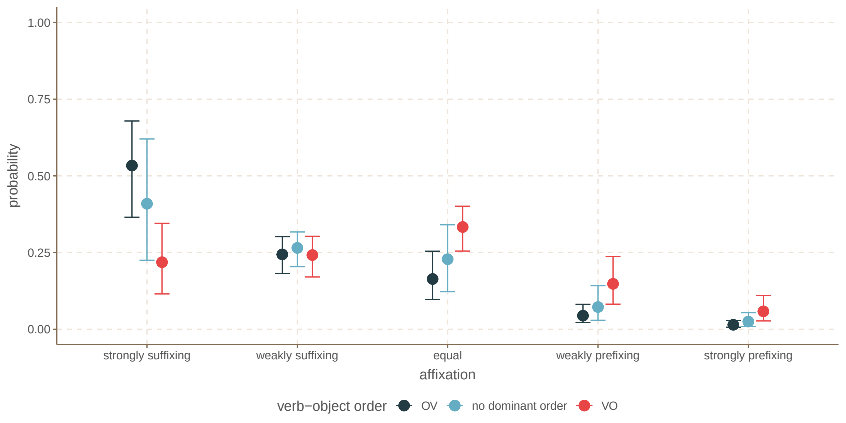
Model comparison

- main model
- hierarchical model
- no-controls model

Hierarchical model

`affixation ~ vo-order + (1|family) + (1|macroarea)`

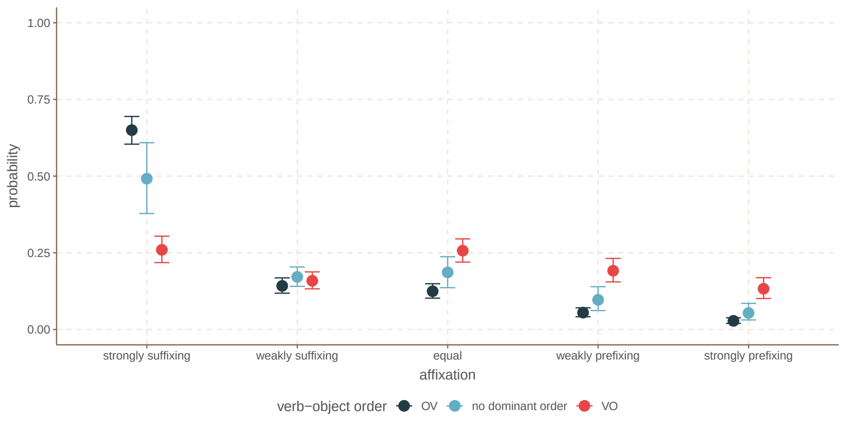
Effects of verb-object order



No-controls model

affixation ∼ vo-order

Effects of verb-object order



Additional model variants

		ELPD diff	SE diff
1	phylo + areal GP+ verb-object	0.0	0.0
2	phylo + areal GP	-10.0	5.8
3	phylo + verb-object	-15.9	6.1
4	(1|family) + areal GP + verb-object	-16.1	7.2
5	(1|family) + (1|macroarea) + verb-object	-55.7	10.5
6	(1|family) + verb-object	-55.9	10.7
7	areal GP + verb-object	-72.9	10.7
8	verb-object	-221.1	14.5

Interim results

From these comparisons we see that:

- our main model has much better performance than both the hierarchical model and the no-controls model

  (especially for predicting less common values)

- the hierarchical model and model without controls overestimate the certainty of the estimates

☞ false positives for word order effects

Oversampling

Oversampling IE

We over-sampled (added them ten more times) the following languages in the training dataset:

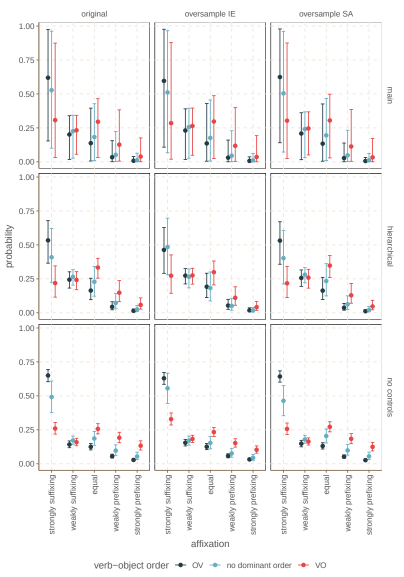
- Italian
- Swedish
- Dutch
- Danish
- Czech
- Slovenian
- Irish
- Welsh
- Tajik
- Central Kurdish

If our method works as we claim, the over-sampled model should not be heavily biased towards IE features.

Oversampling SA

We added all data points ($\sim$100) in South America twice (with a small jitter to their latitude and longitude).

Oversampling



Interim results

With regards to oversampling of IE languages we see that:

- it has a *very small* effect on the the estimates of our model and the hierarchical model

☞ as long as we use some statistical controls, moderate oversampling does not seem problematic

Concluding remarks

Concluding remarks

We have shown how we can control for:

- family bias $\rightarrow$ control through a phylogenetic term
- areal bias $\rightarrow$ control through a 2-dimensional GP

☞ Crucially, our method does not require us to limit our sample.

Concluding remarks

Is **systematic** sampling still required?

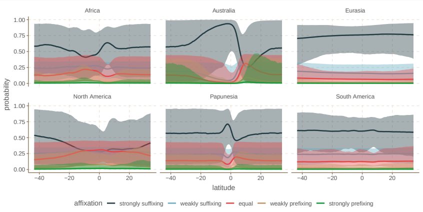
- While some care is still needed, we do not believe our sampling methods need to exclude languages.

☞ We should try to include as much data as we can, and control for bias using statistics.

Thank you!

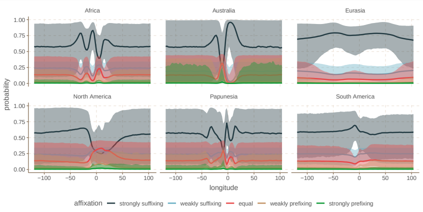
Our model

Geographic effects: **latitude**

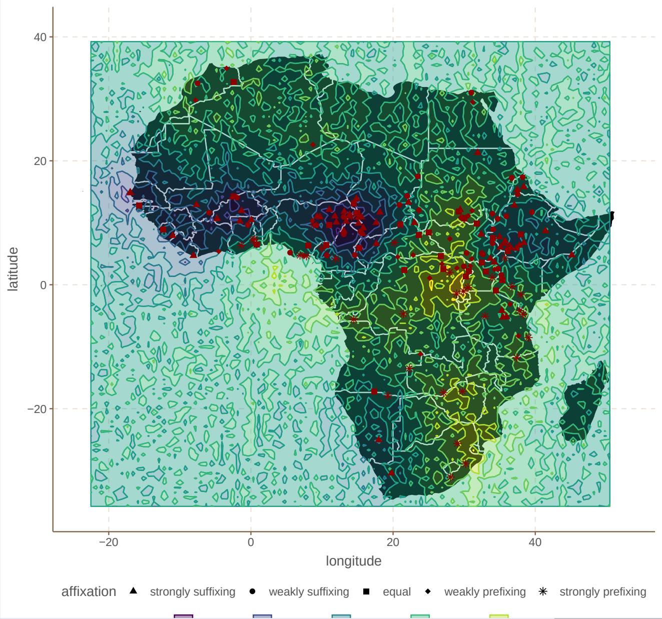


Our model

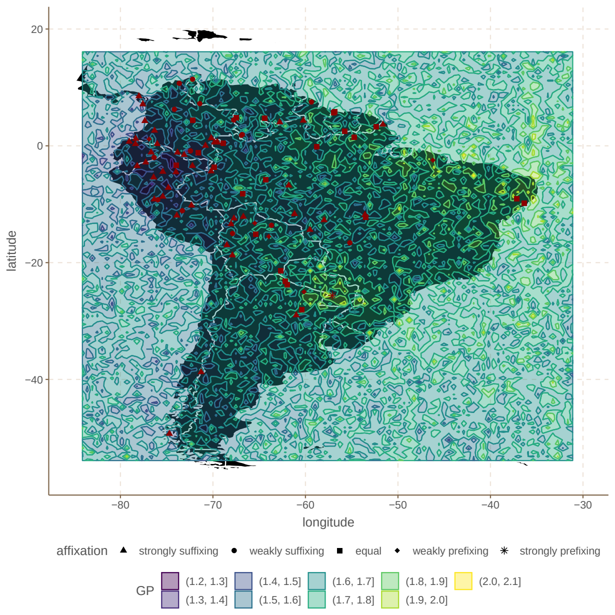
Geographic effects: **longitude**



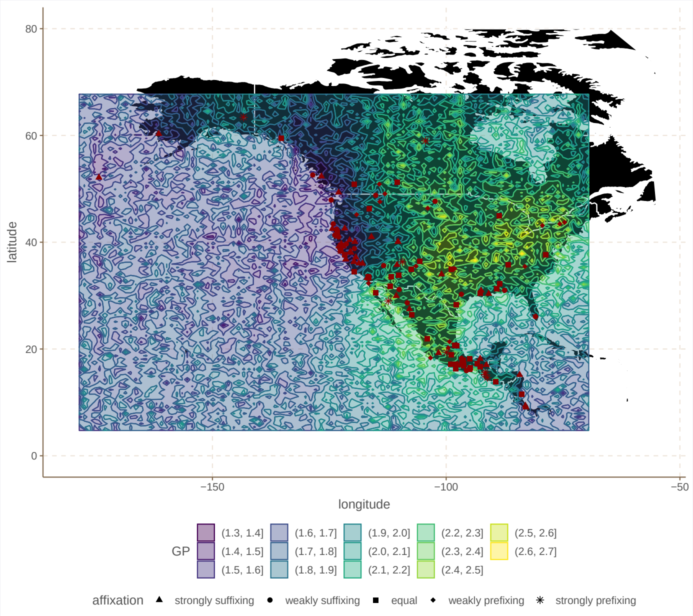
Our model: geographic effects (Africa)



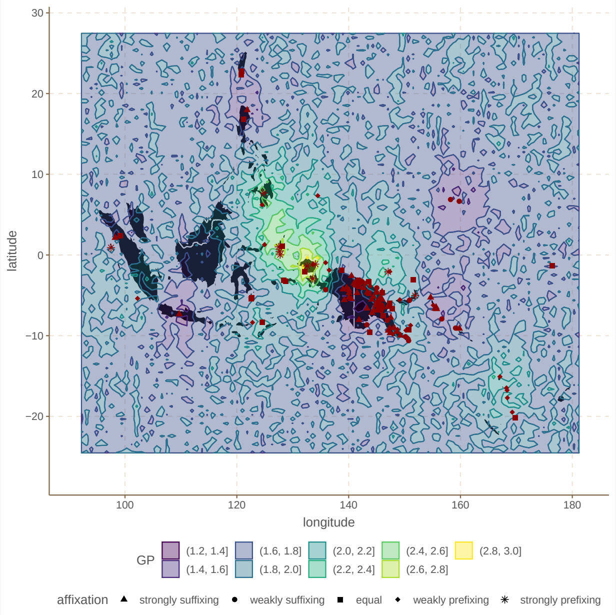
Our model: geographic effects (South America)



Our model: geographic effects (North America)



Our model: geographic effects (Papunesia)



Hierarchical model

`affixation ~ vo-order + (1|family) + (1|macroarea)`

	reference				
prediction	strongly suffixing	weakly suffixing	equal	weakly prefixing	strongly prefixing
strongly suffixing	**247**	36	6	0	2
weakly suffixing	93	**40**	56	35	4
equal	28	37	**64**	36	12
weakly prefixing	11	6	14	**15**	38
strongly prefixing	0	0	0	0	**0**
Accuracy	0.47				
Kappa	0.26				
rmse	0.97				

No controls

affixation ∼ vo-order

	reference				
prediction	strongly suffixing	weakly suffixing	equal	weakly prefixing	strongly prefixing
strongly suffixing	**0**	0	0	0	0
weakly suffixing	286	**77**	64	29	5
equal	93	42	**76**	57	51
weakly prefixing	0	0	0	**0**	0
strongly prefixing	0	0	0	0	**0**
Accuracy	0.2				
Kappa	0.04				
rmse	1.2				