Assessing the effect of geographical isolation on morphological complexity

Christian Bentz,[1,2] Gerhard Jäger,[1,2] & Johanna Nichols[3,4,5]

December 13, 2017

[1]DFG Center for Advanced Studies, University of Tübingen
[2]Department of General Linguistics, University of Tübingen
[3]Department of Slavic Languages, University of California, Berkeley
[4]Linguistic Convergence Laboratory, Higher School of Economics, Moscow
[5]University of Helsinki

Acknowledgements

WORDS BONES GENES TOOLS
Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past

EVOLAEMP
LANGUAGE EVOLUTION: THE EMPIRICAL TURN

Tanja Samardžić

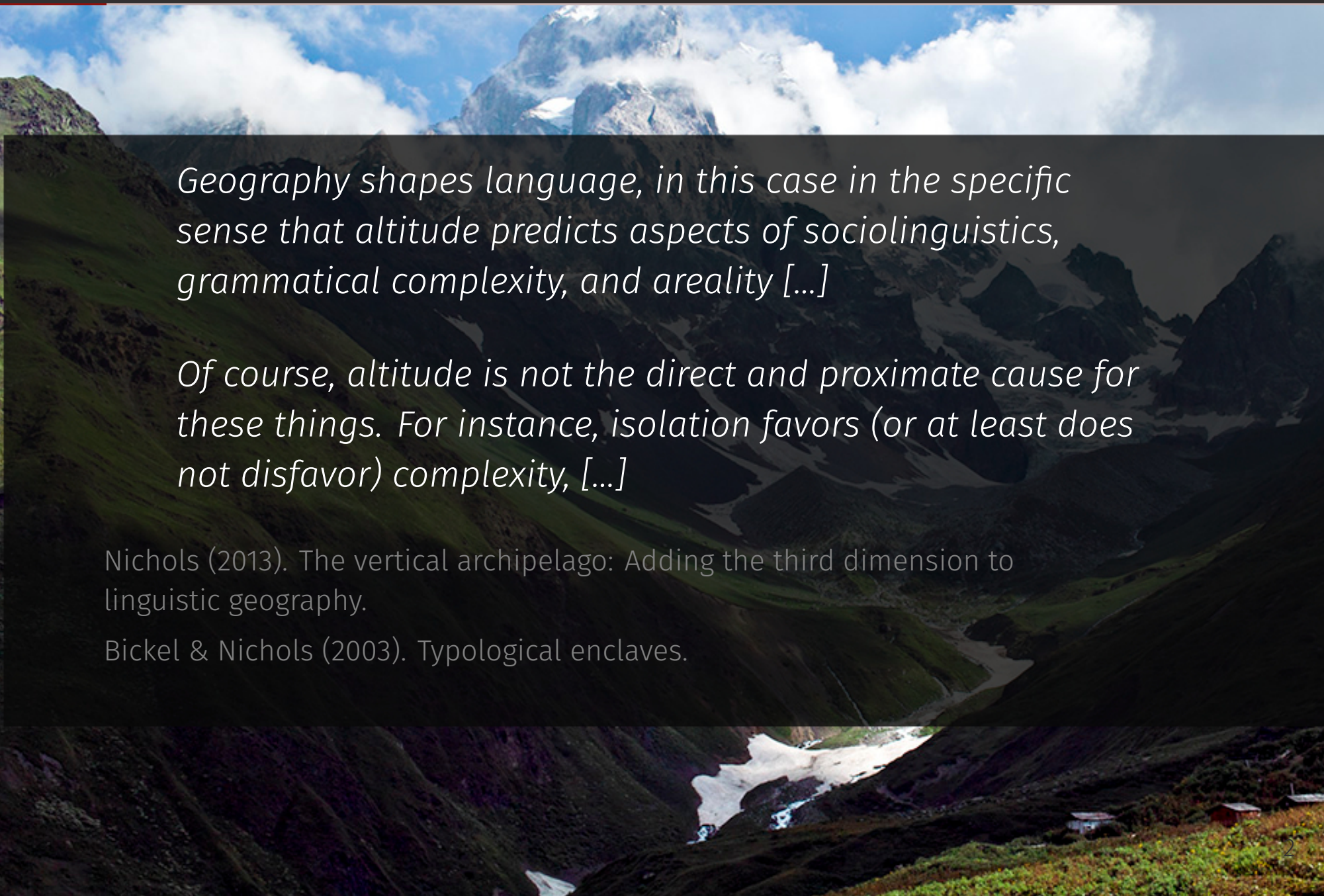
Dimitrios Alikaniotis

Tatyana Ruzsics

Ramon Ferrer-i-Cancho

Introduction

Introduction: "Refuge" areas



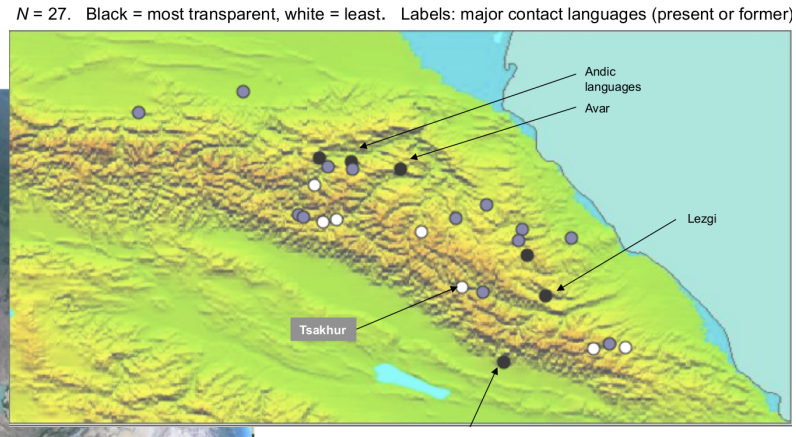
Geography shapes language, in this case in the specific sense that altitude predicts aspects of sociolinguistics, grammatical complexity, and areality [...]

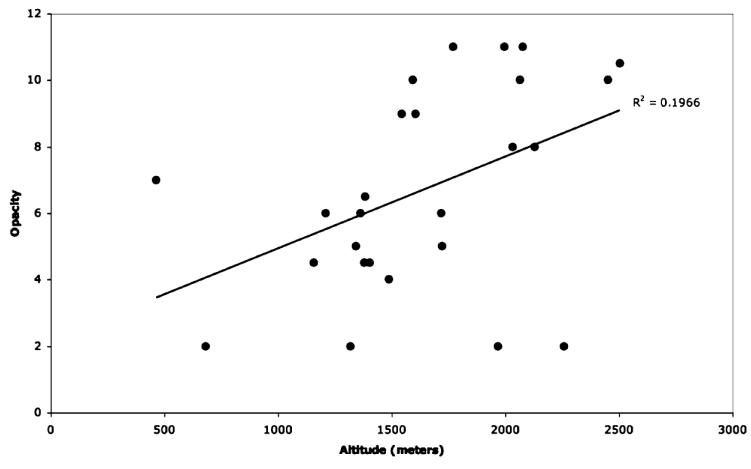
Of course, altitude is not the direct and proximate cause for these things. For instance, isolation favors (or at least does not disfavor) complexity, [...]

Nichols (2013). The vertical archipelago: Adding the third dimension to linguistic geography.

Bickel & Nichols (2003). Typological enclaves.

Introduction: Altitude





Nichols (2016). Complex edges, transparent frontiers. Grammatical complexity and language spreads.

Morphological Complexity Measures

Morphological Complexity Measures

- Inventory complexity prepared by Johanna Nichols (Nichols Compl.)
- Inventory complexity based on WALS (WALS Compl.)
- Unigram word entropy in parallel texts (Unigram H)

Haspelmath & Dryer (eds.) (2013). The World Atlas of Language Structures Online.

Inventory Complexity (Nichols)

Sum of values for 8 morphological features

Feature	Value
Inflectional synthesis	3 to 14
Noun plural	0 = absent, 1 = present
Noun dual	0 = absent, 1 = present
Numeral classifiers	0 = absent, 1 = present
Possessive classification	1 = none, 2 = few, 3 = many
Gender	0 = absent, 1 = present
Auto-gender (on nouns)	0 = absent, 1 = present
Inclusive/exclusive distinction	0 = absent, 1 = present

Note: only languages are included for which all features are available

Inventory Complexity (WALS)

Mean value of up to 28 morphological features

Feature	Value
Number of case markers	1 to 8 (no case to 10 or more)
Future tense	0 = absent, 1 = present
Number of genders	1 to 5 (none to five or more)
Coding of Nominal Plurality	0 = absent, 1 = present
...	...

Note: only languages are included for which more than 10 features are available

Bentz, Ruzsics, Koplenig & Samardžić (2016). A comparison between morphological complexity measures: typological data vs. language corpora.

Corpus-based measure

(1) Kayardild

Ngada maarra junku-ru-thu,
I only right-FACTITIVE-POTENTIAL,
thaku-ru-nangku
left-FACTITIVE-NEGATIVE.POTENTIAL

"I will only make things correct, I will not twist or distort anything"

Evans (2010). Dying words: endangered languages and what they have to tell us. p. xxi.

Corpus-based measure

As you can see, the Kayardild version compresses two English sentences into just four words.

Evans (2010). Dying words: endangered languages and what they have to tell us. p. xxi.

Corpus-based measure

(2) Hawaiian (haw)

A ua olelo aku o Ioane ia ia [...]
Then PERF say to SUBJ Johan he.DAT [...]

"Then Johan said to him [...]"

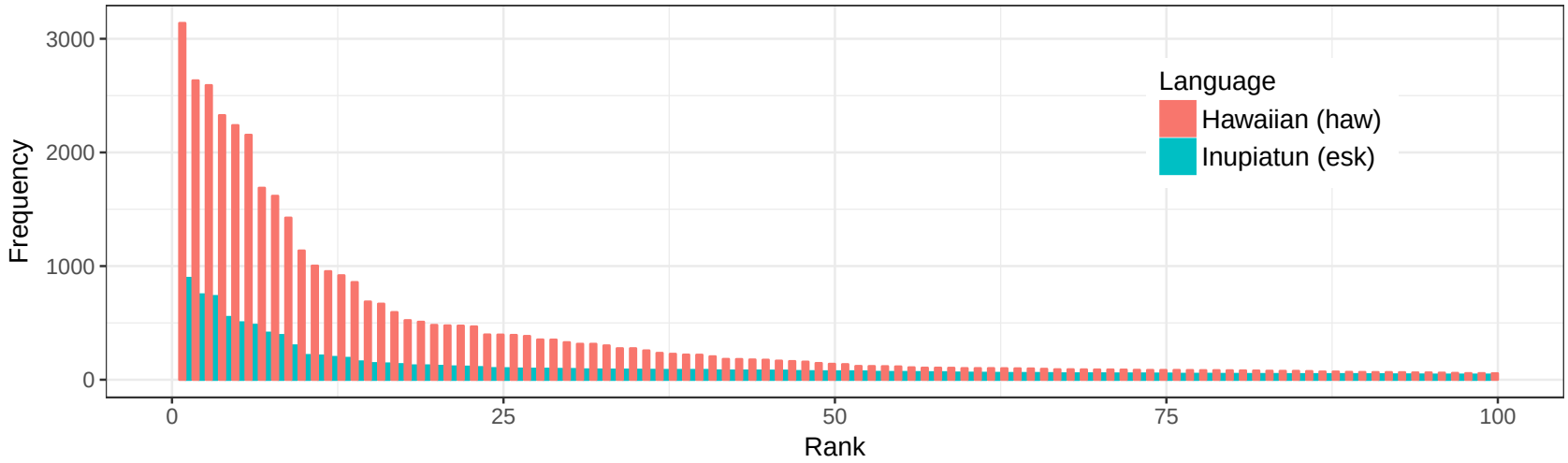
(3) Iñupiatun (esk)

Aglaan Jesus-ŋum itna-ġ-ni-ġai [...]
But Jesus-ERG this-say-report-3S.to.3PL

"But Jesus said to them [...]"

Mayer & Cysouw (2014). Creating a massively parallel bible corpus.

Unigram Word Entropy



$$H(T^{haw}) \sim 7 \text{ bits/unigram}$$
$$H(T^{esk}) \sim 13 \text{ bits/unigram}$$

Shannon (1948)

Bentz, Alikaniotis, Cysouw & Ferrer-i-Cancho (2017). The entropy of words — learnability and expressivity across more than 1000 languages.

Data

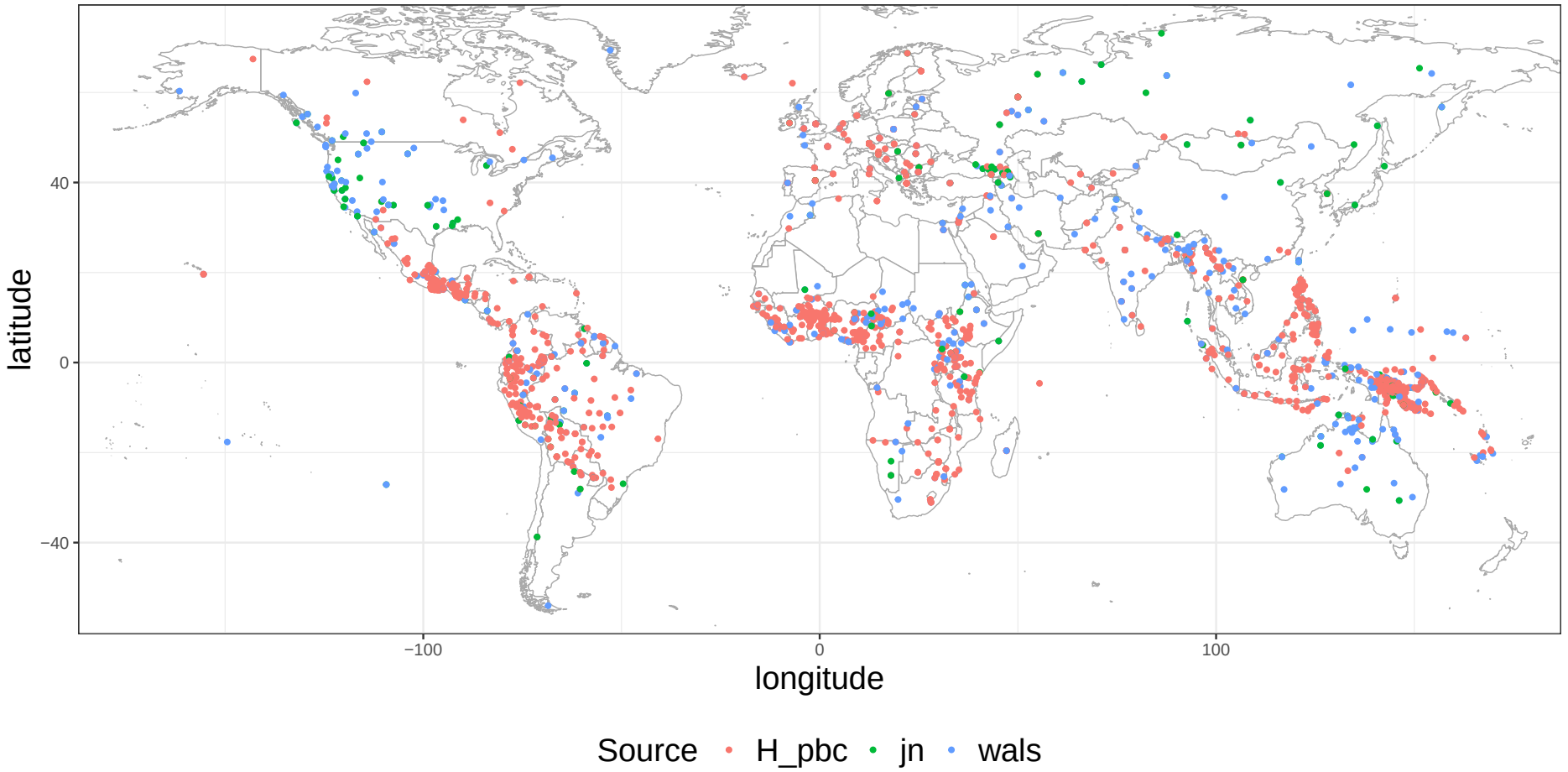
Data

Table 1: Three data samples used for statistical analyses.

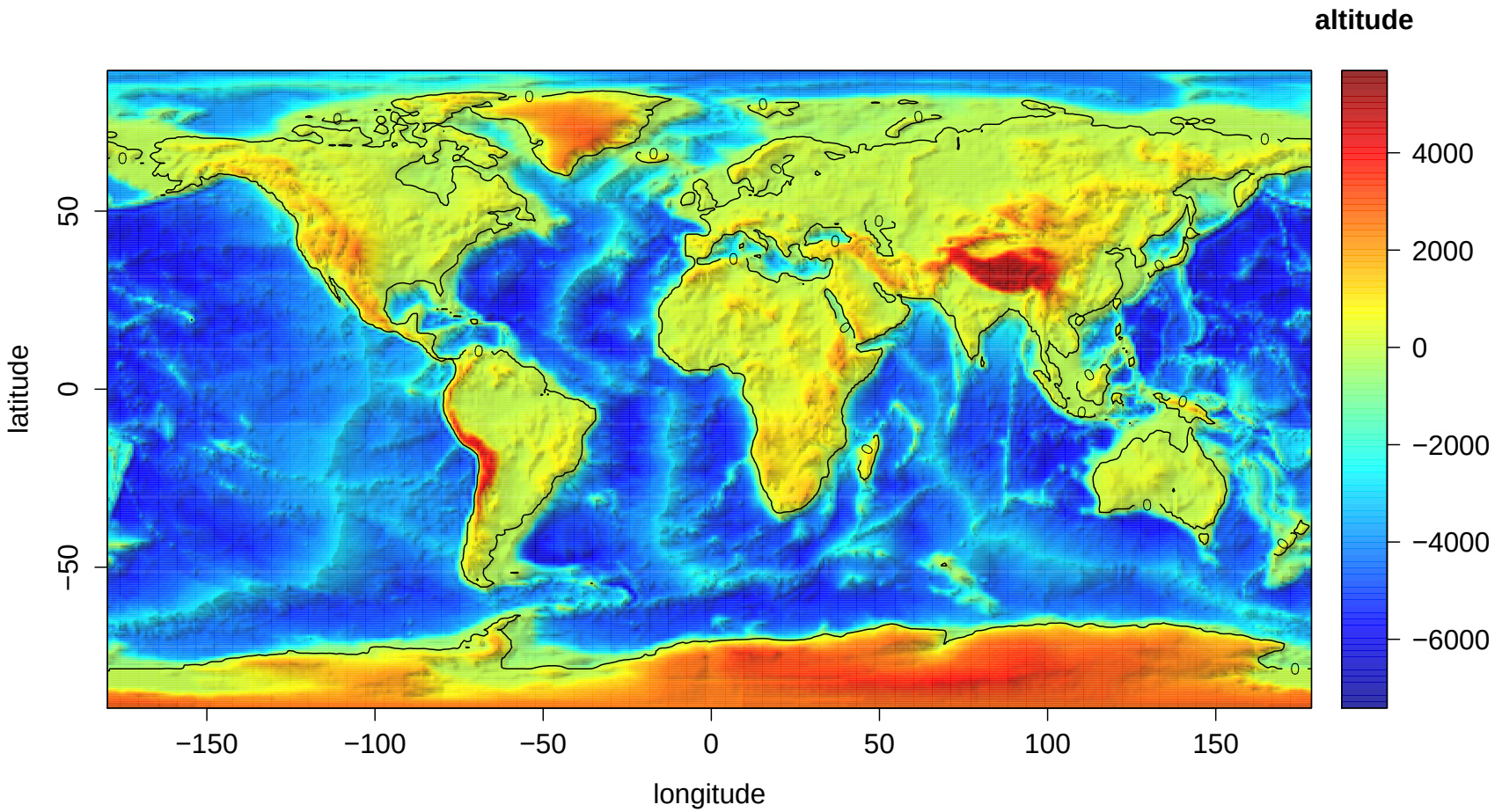
Sample	Data points	Lang. (ISO)	Families[1]	Areas
Unigram H (PBC)	1432	1071	100	6
Nichols (JN)	183	183	84	6
WALS	468	465	112	6
Total	2083	1380	152	6

[1] Families and macro areas from Glottolog 2.7 (Hammarström et al. 2016)

Data: Global Distribution



Altitude



Google Maps API

R package *plot3D*

Statistical Models

Results: Linear Regression

$$y_i^{compl} = \beta_0 + \beta_1 x_i^{alt} + \epsilon_i$$

Results: Linear Regression

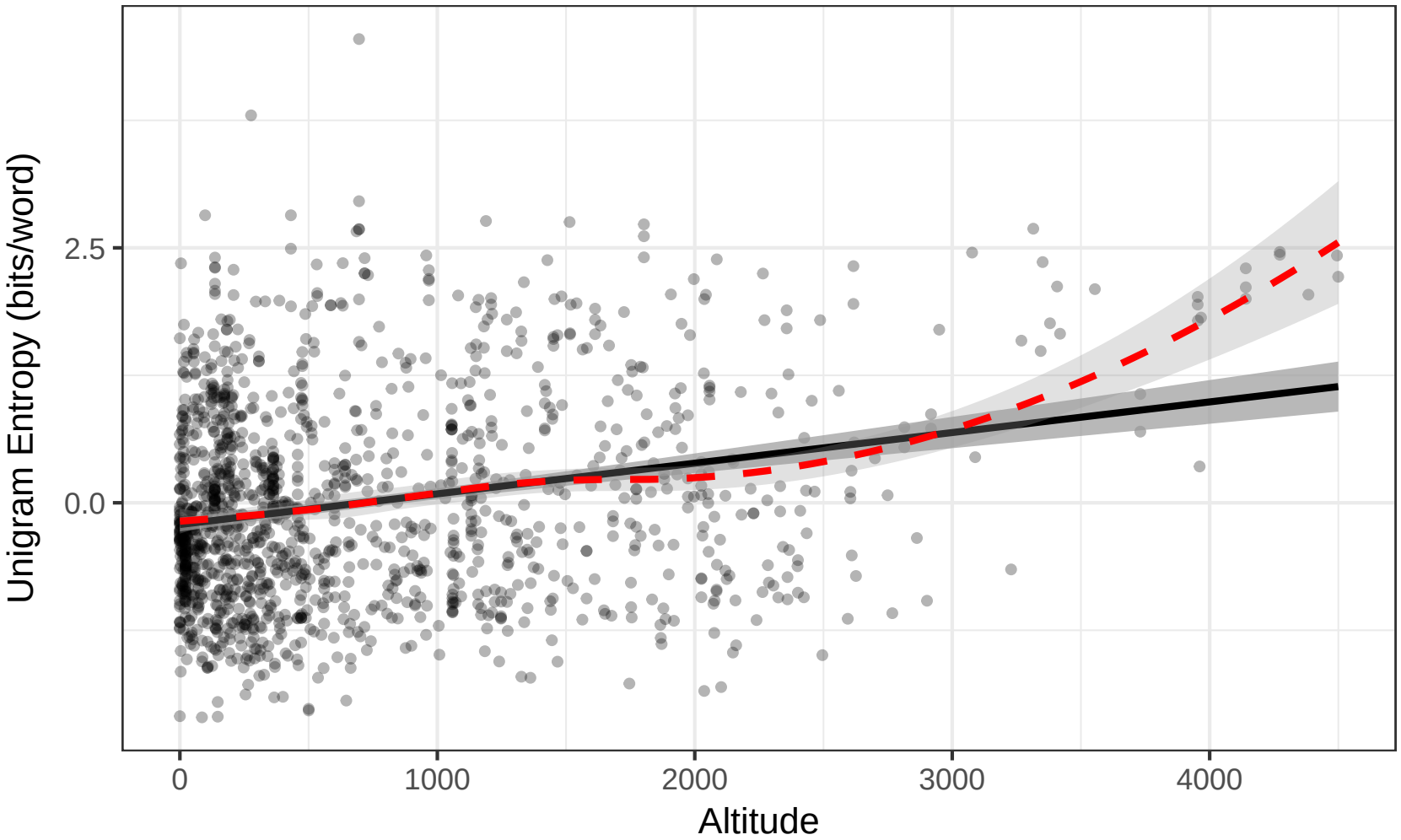
$$y_i^{compl} = \beta_0 + \beta_1 x_i{}^{alt} + \epsilon_i$$

Table 2: Results of linear regressions.

Dependent	Predictor	coef. (β_1)	SE	t-value	R^2
Unigram H	altitude	0.0003	0.00003	9.31***	0.06
Nichols Compl.	altitude	0.00008	0.00007	1.22	0.22
WALS Compl.	altitude	0.0001	0.00003	3.34***	0.02

*** p<0.001; ** p<0.01; *p<0.05

Results: Linear Regression



Results: Linear Mixed-Effects Regression

$$y_i^{compl} = \beta_0 + \beta_{0f} + \beta_{0a} + (\beta_1 + \beta_{1f} + \beta_{1a}) x_i^{alt} + \epsilon_i$$

Results: Linear Mixed-Effects Regression

$$y_i^{compl} = \beta_0 + \beta_{0f} + \beta_{0a} + (\beta_1 + \beta_{1f} + \beta_{1a}){x_i}^{alt} + \epsilon_i$$

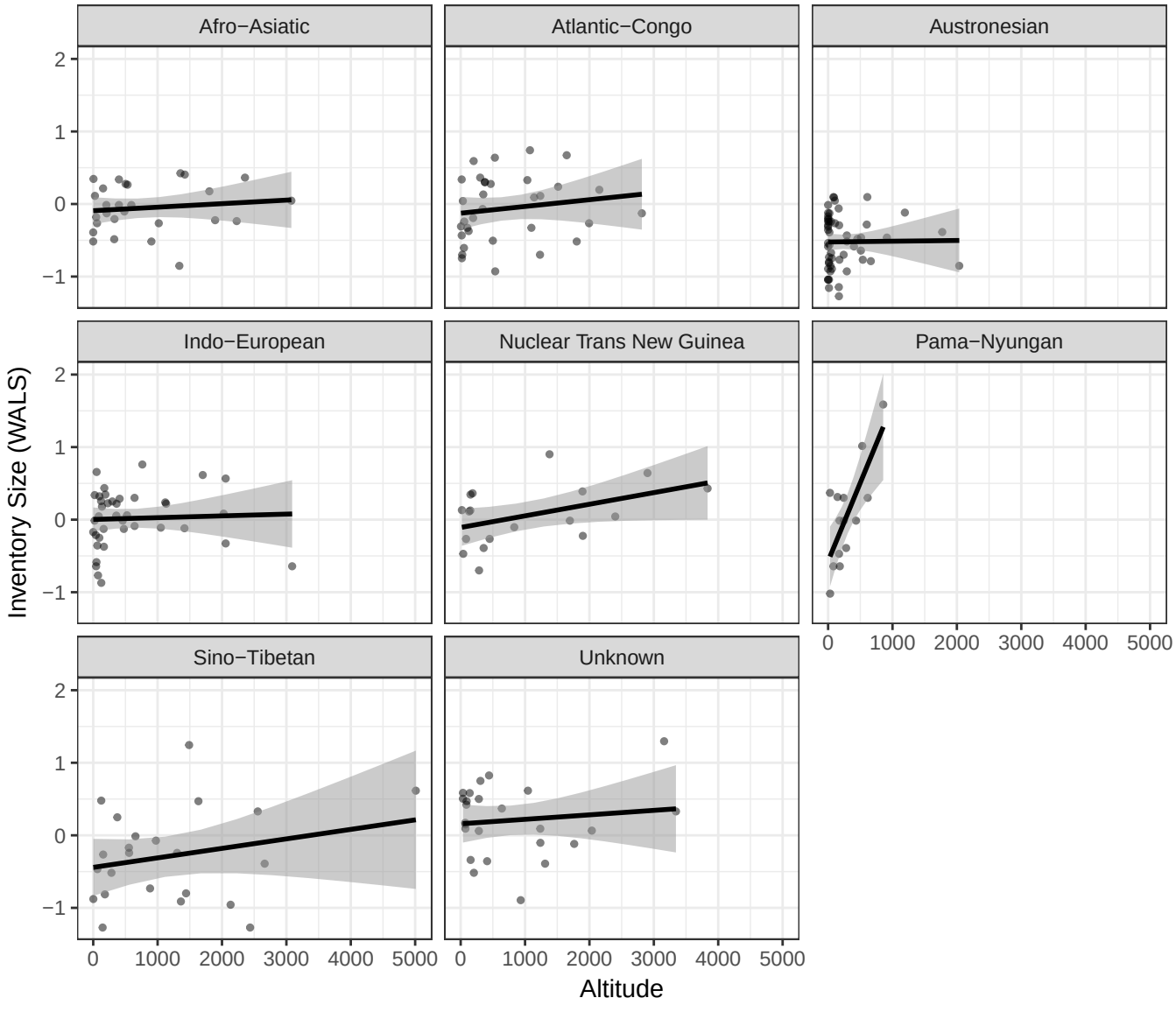
Table 3: Results of stepwise linear mixed-effects regression.

Dependent	Fixed	Random		coef. (β_1)	t-value	R^2
		interc.	slope			
Unigram H	altitude	f, a	--	0.0003	8.84***	0.64
Nichols Compl.	altitude	f, a	--	0.00005	0.72	0.2
WALS Compl.	altitude	f	--	0.00008	2.87**	0.24

*** p<0.001; ** p<0.01; *p<0.05

† f: family, a: area

Results: Linear Mixed-Effects Regression



Results: Phylogenetic Generalized Least Squares

Table 4: Results of PGLS.

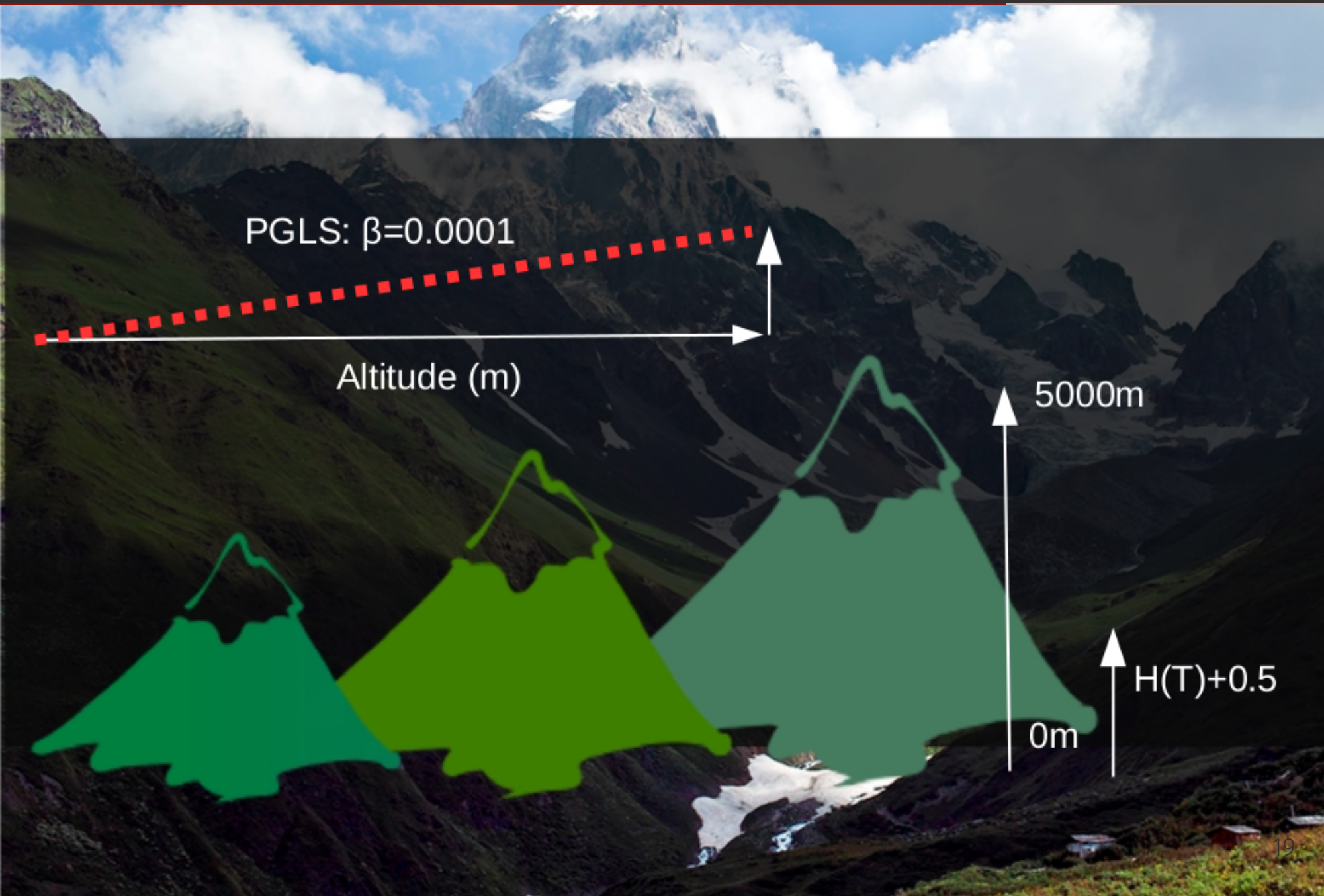
Dependent	Predictor	coef. (β)	SE	t-value
Unigram H	altitude	0.0001	0.00003	3.4***
Nichols Compl.	altitude	0.00005	0.00007	1.63
WALS Compl.	altitude	0.00004	0.00003	0.52

*** p<0.001; ** p<0.01; *p<0.05

Jäger & Wichmann (2016). Inferring the world tree of languages from word lists.

Discussion

Discussion: Effect Size



Discussion: Effect Size

$\hat{H}$(fin) ≈ 1.3

$\hat{H}$(hun) ≈ 1

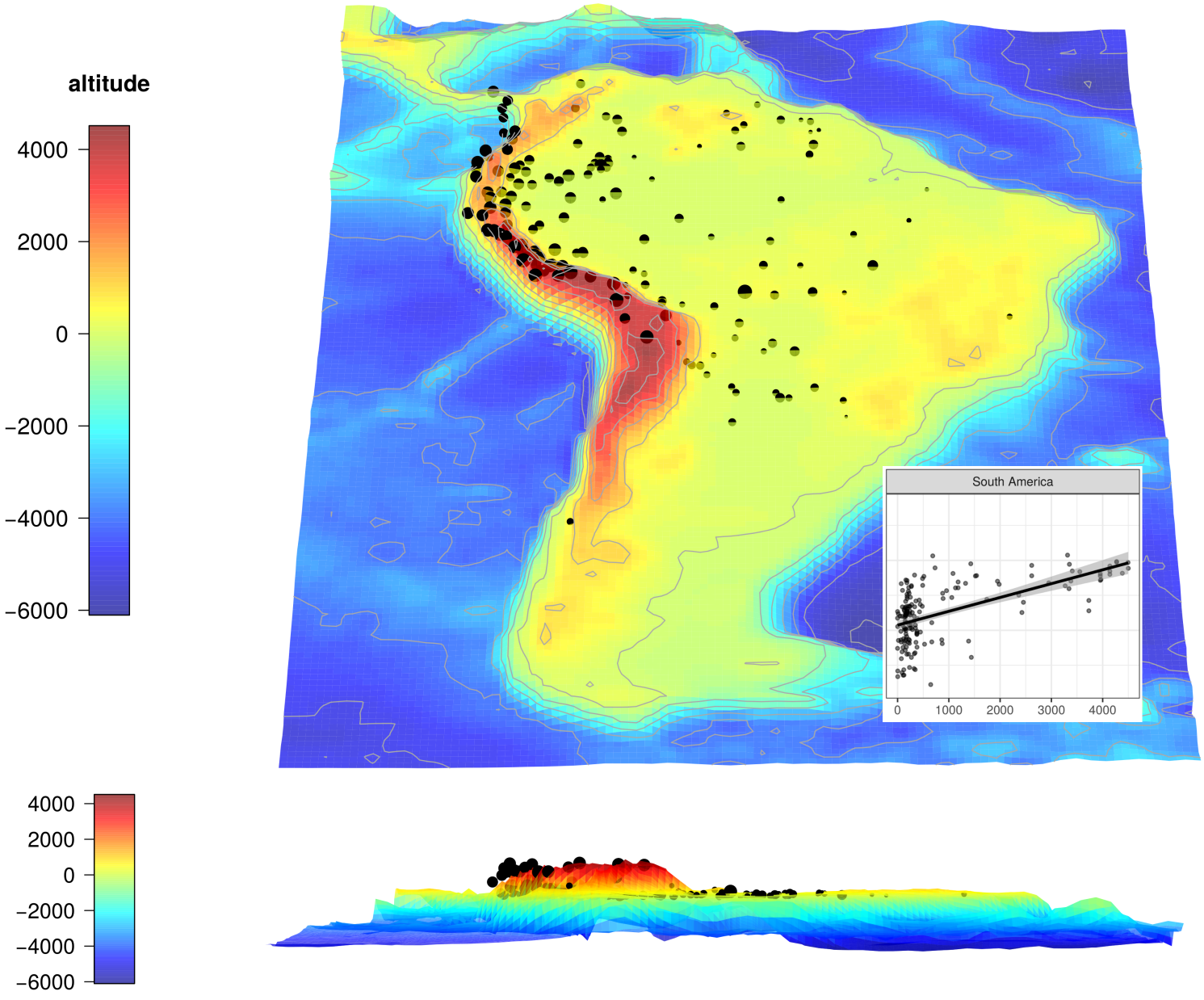
$\hat{H}$(isl) ≈ 0.4

$\hat{H}$(deu) ≈ 0.2
$\hat{H}$(fra) ≈ 0.1

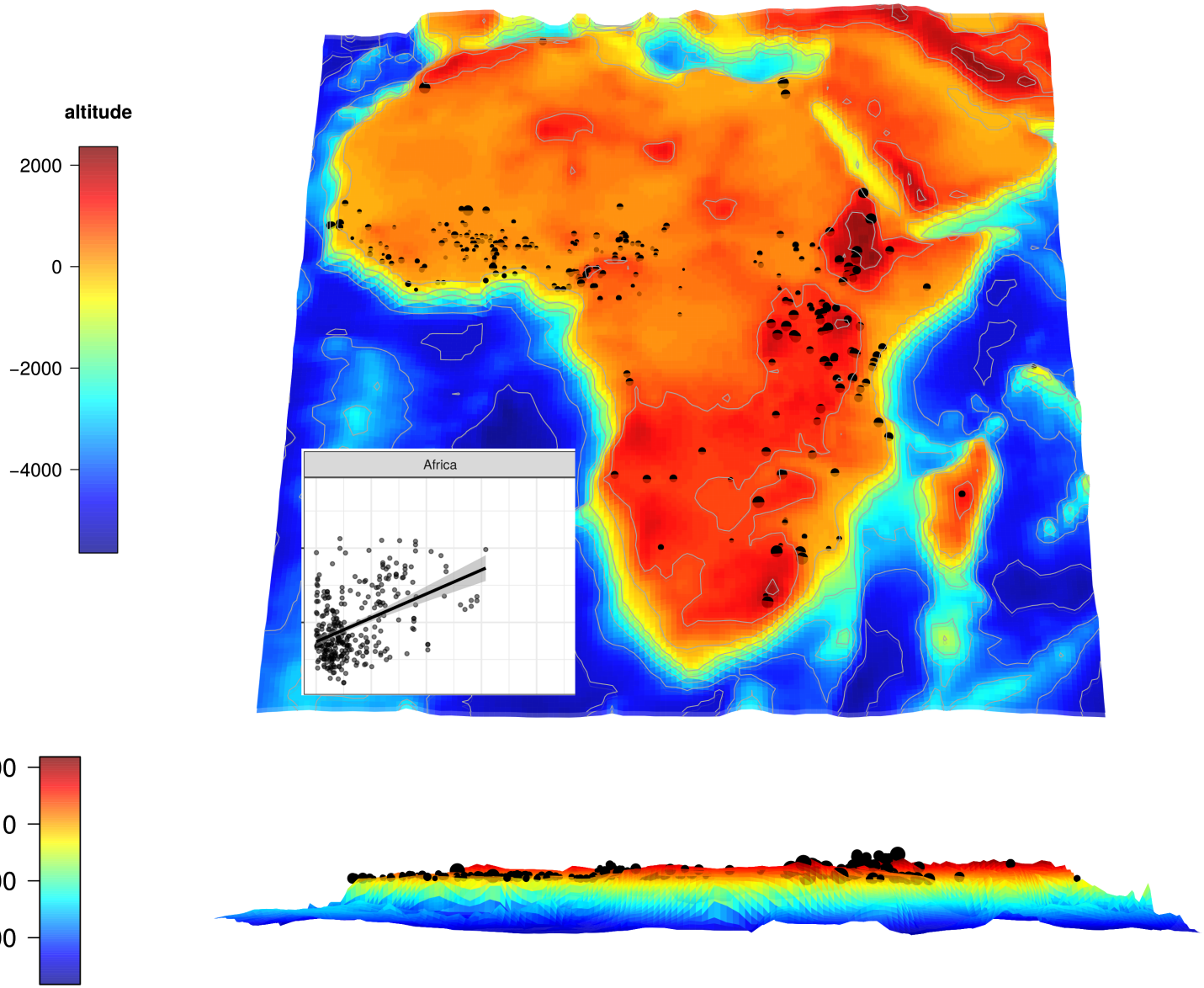
$\hat{H}$(spa) ≈ -0.1
$\hat{H}$(nld) ≈ -0.2

$\hat{H}$(eng) ≈ -0.6

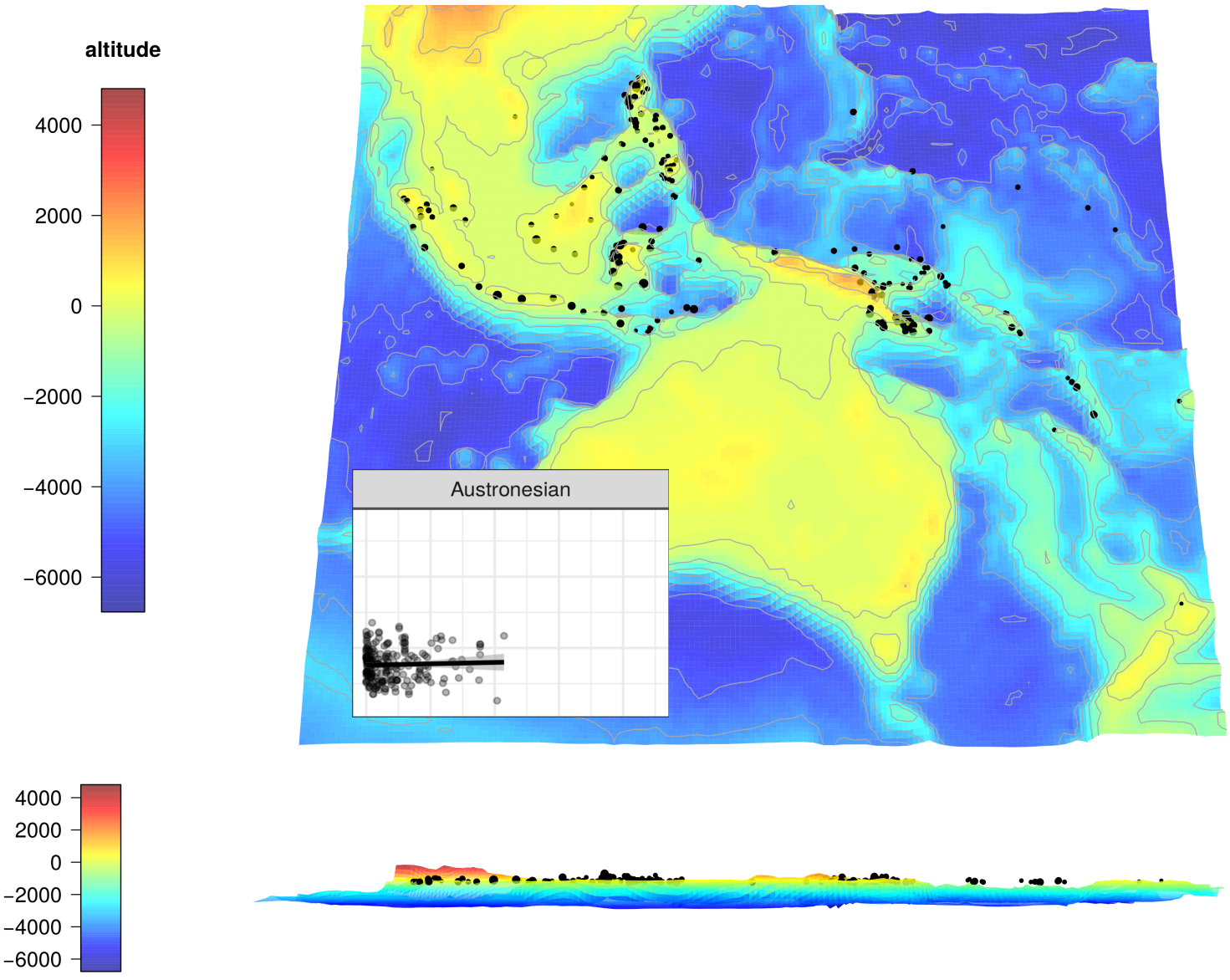
South America



Africa

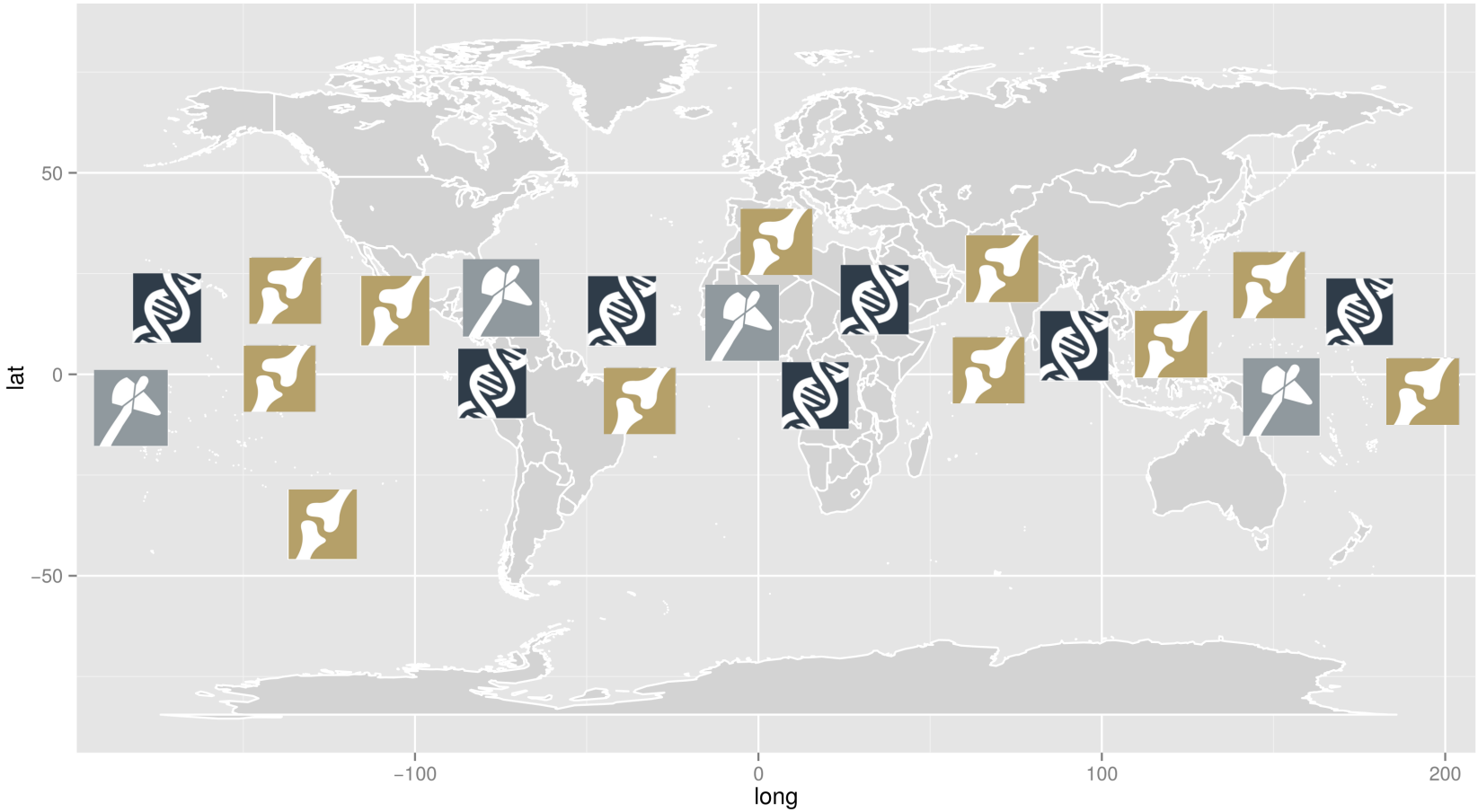


Austronesian



Conclusion

Conclusion

- There is a world wide effect higher altitude $\rightarrow$ higher complexity, reflected in WALS complexity and unigram entropy, but not for Nichols' complexity (sample size?)
- For WALS this is maintained in a mixed-effects model controlling for family and area variation, but not in a finer-grained phylogenetic regression
- For unigram entropy the effect is robust at all levels of phylogenetic and areal controls

Thank You