Identifying covariate traits of trance and possession phenomena in the Ethnographic Atlas using multiple imputation and nested sparse regression

Péter Rácz

23 April, 2020

Structure of the SI

The EA dataset is challenging in three ways. (1) Many, correlated, predictors might explain the outcome, (2) Data are structured by cultural and geographic distance, (3) Data are missing.

The analysis attempts to meet the challenge by the (1) Use of shrinkage for predictor selection, (2) Use of a nested model to account for cultural and geographic structure (in a relatively simple way), (3) Use of imputed data in model selection.

Models are fit in R (R Core Team 2019) using Stan (Stan Development Team 2019) and brms (Bürkner 2018). Figures are created using ggplot (Wickham 2016), tidybayes (Kay 2020), and bayesplot (Gabry et al. 2019).

Data come from the Ethnographic Atlas (Murdock 1967) made available in the D-Place database (Kirby et al. 2016) (see helper file).

Analysis was heavily inspired by Kevin Stadler (*link*) and Michael Betancourt (*link*), as well as by Andrew Gelman's collection of prior recommendations (*link*).

This Supplementary Information is structured in the following way. First, we consider the starting predictors. Second, we code the outcome. Third, we consider autocorrelation in the data. Fourth, we consider missing data using imputed datasets. Fifth, we consider variable selection using sparse regression. Sixth, and finally, we refit the best model on the available EA data and perform some additional robustness checks.

Starting predictor variables

The EA contains values of 94 variables for 1291 societies.

The existing literature identifies a number of promising correlates of trance states in the Ethnographic Atlas. We select twenty-one of these as starting variables. Categorical variables will be fit with a default intercept, ordinal variables, where we can posit a hierarchy between factor levels, will be fit as numeric scales. This runs the risk of missing non-linear effects, but greatly contributes to model health in the long run.

For many of these variables, a large amount of data are missing.

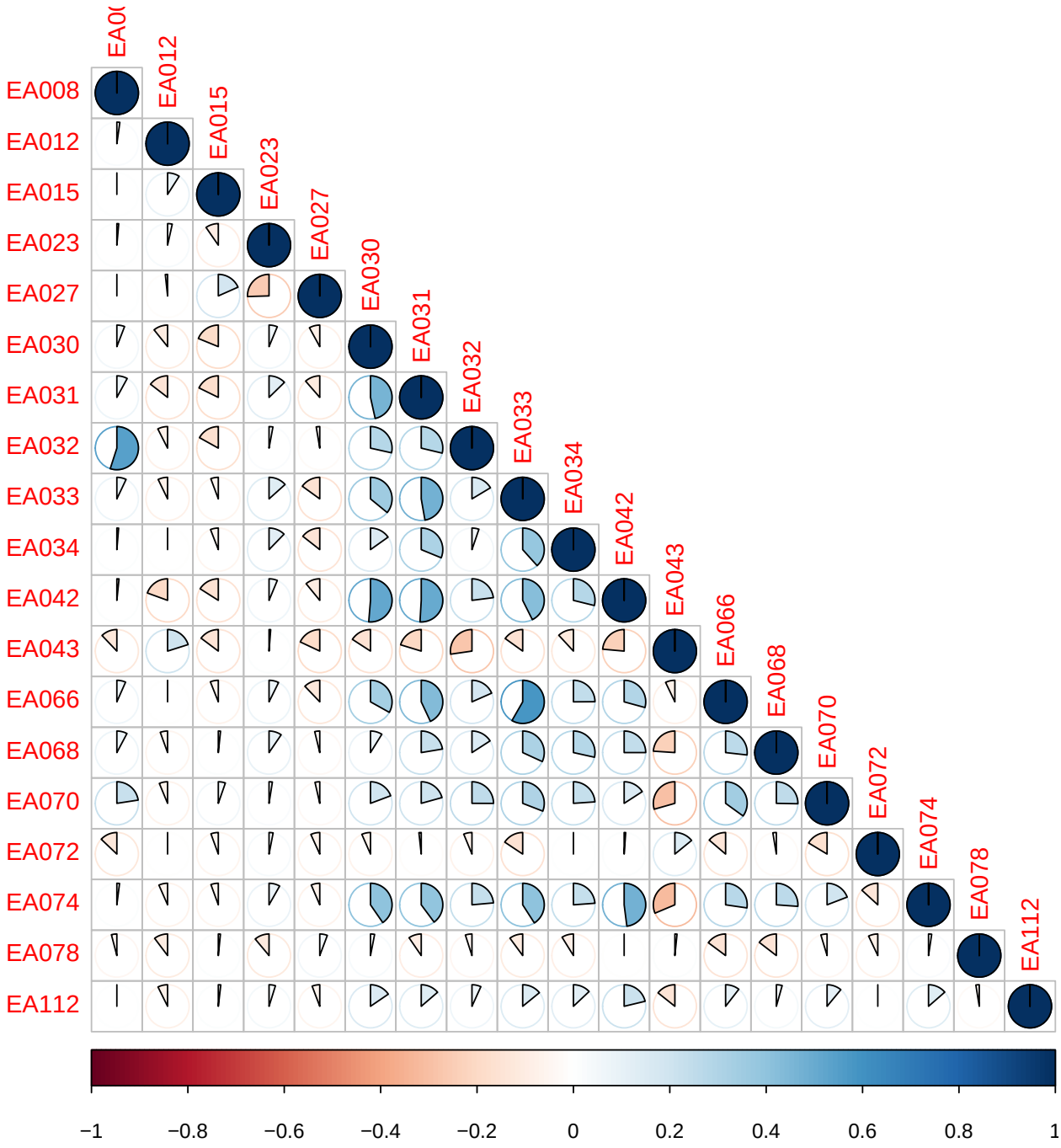
var_id	var_title	type	percent_missing
EA008	Domestic organization	ordinal	2
EA012	Marital residence with kin: prevailing pattern	ordinal	2
EA015	Community marriage organization	categorical	15
EA023	Cousin marriages permitted	ordinal	19
EA027	Kin terms for cousins	ordinal	26
EA030	Settlement patterns	ordinal	8
EA031	Mean size of local communities	ordinal	53
EA032	Jurisdictional hierarchy of local community	ordinal	10
EA033	Jurisdictional hierarchy beyond local community	ordinal	11
EA034	Religion: high gods	ordinal	40
EA042	Subsistence economy: dominant activity	categorical	0
EA043	Descent: major type	ordinal	1
EA066	Class differentiation: primary	ordinal	14
EA068	Caste differentiation: primary	ordinal	15
EA070	Slavery: type	ordinal	13
EA072	Political succession	ordinal	27
EA074	Inheritance rule for real property (land)	categorical	34
EA078	Norms of premarital sexual behavior of girls	ordinal	54
EA112	Trance states	ordinal	49
EA113	Societal rigidity	categorical	98
EA202	Population size	ordinal	26

Thompson, Roberts, and Lupyan (2018) use multiple imputation using classification and regression trees (**???**) to impute missing data in the EA dataset. As they point out, the rough accuracy of this method for unseen data is 74%, compared to a random sampling baseline of 19%.

We borrow one of their imputed datasets to help us visualise correlations between our variables. We assume that 74% of the imputations for the missing 25% will be accurate. This means that 94% of the underlying data are accurate.

This is good enough for this figure. We will take data imputation more seriously in our models.

We visualise Spearman correlations between the relevant variables in our imputed dataset. We can see that some of these are indeed correlated. For example, primary class differentiations in a society (EA066) increase with the increase of structure in the larger community (EA033). Different primary subsistence economies (EA042) go together with different levels of local and larger social hierarchical complexity. This makes sense.



Outcome variable

Trance and possession states are coded in the Ethnographic Atlas under EA112.

description
No trance states of any kind are known to occur, and there is no belief in possession.
Trance behavior is known to occur, but there is no belief in possession.
A belief in possession exists.
There is both a trance state and a belief in possession, but this belief refers to phenomena other than trance, which is expl
Two types of trance states are known to occur. One which is explained as due to possession and one which is given anothe
Trance explained as due to possession is known to occur, and there are no other trance states, but cases of possession outs
Trance states of two kinds are known to occur, some of which are explained by possession. No other phenomena are explai
Trance behavior is known to occur and is explained as due to possession. There is no possession belief referring to other ex
Missing data

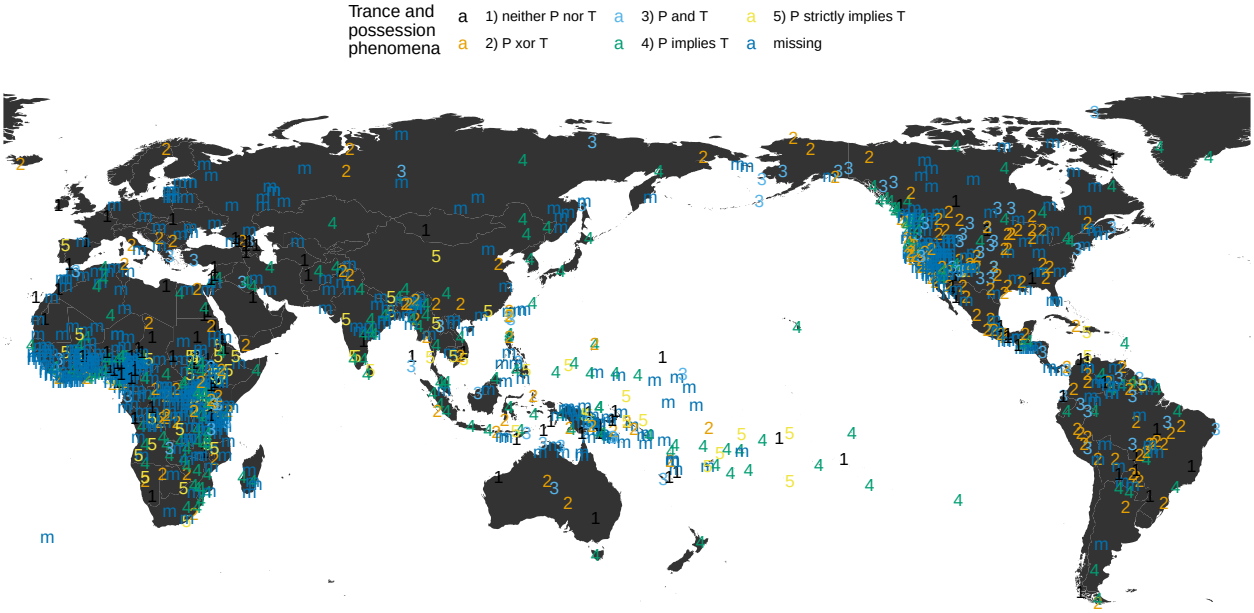
The levels of this categorical variable hint at the difficulty of encoding complex relationship between trance states and possession in human cultures.

For the purposes of this broad-brush analysis, we can recognise a number of implicative structures. Societies can be grouped according to whether:

1. *No* trance states are known to occur and *no* belief in possession exists.
2. *Either* trance states are known to occur *or* belief in possession exists.
3. *Both* trance states are known to occur *and* belief in possession exists. *No relationship is stipulated between the two.*
4. *Both* trance states are known to occur *and* belief in possession exists. *Not all* trance states are explained through possession **or** *not all* cases of possession are related to trance.
5. *Both* trance states are known to occur *and* belief in possession exists. *All* trance states are explained through possession **and** *all* cases of possession are related to trance.

Here we stipulated a hierarchy of types of trance and possession beliefs across the societies recorded in the Ethnographic Atlas.

Let's put this on a map.



Data are mostly absent for the Eurasian landmass. Possession and/or Trance states are posited (2-3) primarily in the Americas in the dataset, whereas strong links are present for the two (4 and especially 5) primarily in Sub-Saharan Africa and Oceania.

Modelling: coding the outcome

Without making a deep ontological commitment to this hierarchy we can use it to model the distribution of trance and possession beliefs in the dataset by using an ordered categorical model. This is an extension of the logistic generalised linear regression model which predicts the relative likelihood of one outcome (say, presence of trance states) over another (say, absence of trance states).

The ordered model assumes that the levels of this hierarchy are ordered from 1 to 5 and predicts the cumulative likelihood of responses in an expanding set of the levels in our hierarchy. It makes a prediction for

- 1 (viz that trance and possession states are categorically absent) vs 2:5 (they are present on some level),
- 1:2 (viz that trance or possession states can be observed but not together) vs 3:5 (they can be observed together)
- 1:3 (viz that trance or possession states can be observed but without a relationship between the two) vs 4:5 (there is a relationship)
- 1:4 (viz that trance or possession states and a relationship between them can be observed but if such a relationship exists it's not exclusive) vs 5 (the relationship between possession and trance is exclusive)

What we do stipulate is this ordering of the levels. However this is justified to some extent by the implicational structure.

Modelling: accounting for Galton's problem

The state of the art in accounting for cultural distance between observations relies on the use of phylogenetic comparative methods that use language trees to calculate distances between cultural units and take these distances into account when correcting for autocorrelation. These are difficult to deploy for the EA data, since the 1291 societies in the dataset speak 1211 languages that come from 139 language families.

Following standard practice, we assume a grouping factor for language family and one for geographic-cultural region as defined in the Ethnographic Atlas. These will go to some length in accounting for autocorrelation in cultural behaviour, viz that trance and possession states might be more typical for certain cultural groups or regions.

Modelling: variable selection

With 21, collinear, predictor variables, it becomes difficult for a model to capture those that are relevant in explaining variation in the outcome. The use of stepwise regression is generally discouraged (Flom and Cassell 2007). Frequentist methods, like L1 (Lasso) and L2 (Ridge) regularisation, exist for variable selection and regularisation.

If we want to assume that, among a large set of possibly covariate predictors, only a few have any robust correlation with our outcome, the Bayesian approach is to build this assumption into our model in the form of an informative prior distribution. Such sparsity-inducing distributions will allow for variable selection in the vein of the frequentist LASSO method (Tibshirani 1996), except that we introduce sparsity through the distribution rather than by using a penalised maximum likelihood estimator.

We consider three types of prior distributions in our models: a weakly informative Student's t distribution (which does not push the model towards variable selection), a Laplace distribution (effectively, the Bayesian equivalent of the LASSO method), and the Horseshoe prior, a heavy-tailed Cauchy prior distribution that generally pushes weaker posteriors towards zero (Carvalho, Polson, and Scott 2009), which we use by adjusting the global scale to .5 to favour shrinkage (Piironen and Vehtari 2016).

We use the Widely Applicable Information Criterion (WAIC) and leave-one-out cross-validation for model selection.

Modelling: missing data

Thompson, Roberts, and Lupyan (2018) published 100 iterations of the imputed EA dataset. These have, on average, 74% accuracy in the imputed values. We draw a random dataset from this set of 100 to fit our model and draw a full set of ten random datasets to refit the model. We walk through these steps below.

Fitting the model

We draw ten imputed datasets at random from Thompson, Roberts, and Lupyan (2018) and use the **first** of these for model selection.

We fit a multilevel ordered categorical model on the dataset. It predicts the presence of trance states and possession phenomena in a society as a five-step scale (see above) using the 21 pre-selected variables as predictors (see above). We assume a grouping factor for language family and geographic region, drawn from the Ethnographic Atlas.

We fit the model thrice, with a Laplace prior (Fit 1), a Student-t prior (Fit 2), and a Horseshoe prior (Fit 3). We use four mcmc chains with 2000 iterations each.

We estimate the expected log pointwise predictive density for each model and consider their differences:

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
fit3	0.00	0.00	-1372.30	16.27	47.91	1.64	2744.60	32.55
fit1	-11.69	3.58	-1383.99	17.07	62.09	2.05	2767.98	34.13
fit2	-13.95	4.15	-1386.25	17.21	63.53	2.09	2772.49	34.43

This indicates that using the Horseshoe prior provides the best fit on the dataset.

Next, we refit Fit 3 on the dataset without a grouping factor for geographic region (Fit 3b), language family (Fit 3c), or both (Fit 3d) and compare these as well:

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
fit3c	0.00	0.00	-1371.71	16.20	36.26	1.23	2743.42	32.40
fit3	-0.59	2.18	-1372.30	16.27	47.91	1.64	2744.60	32.55
fit3b	-5.02	5.53	-1376.73	15.80	40.01	1.63	2753.46	31.61
fit3d	-14.15	5.65	-1385.87	15.03	15.69	0.46	2771.73	30.06

The results indicate that grouping observations by language family doesn't account for a substantial amount of variation. Geographic region, however, is a useful grouping factor.

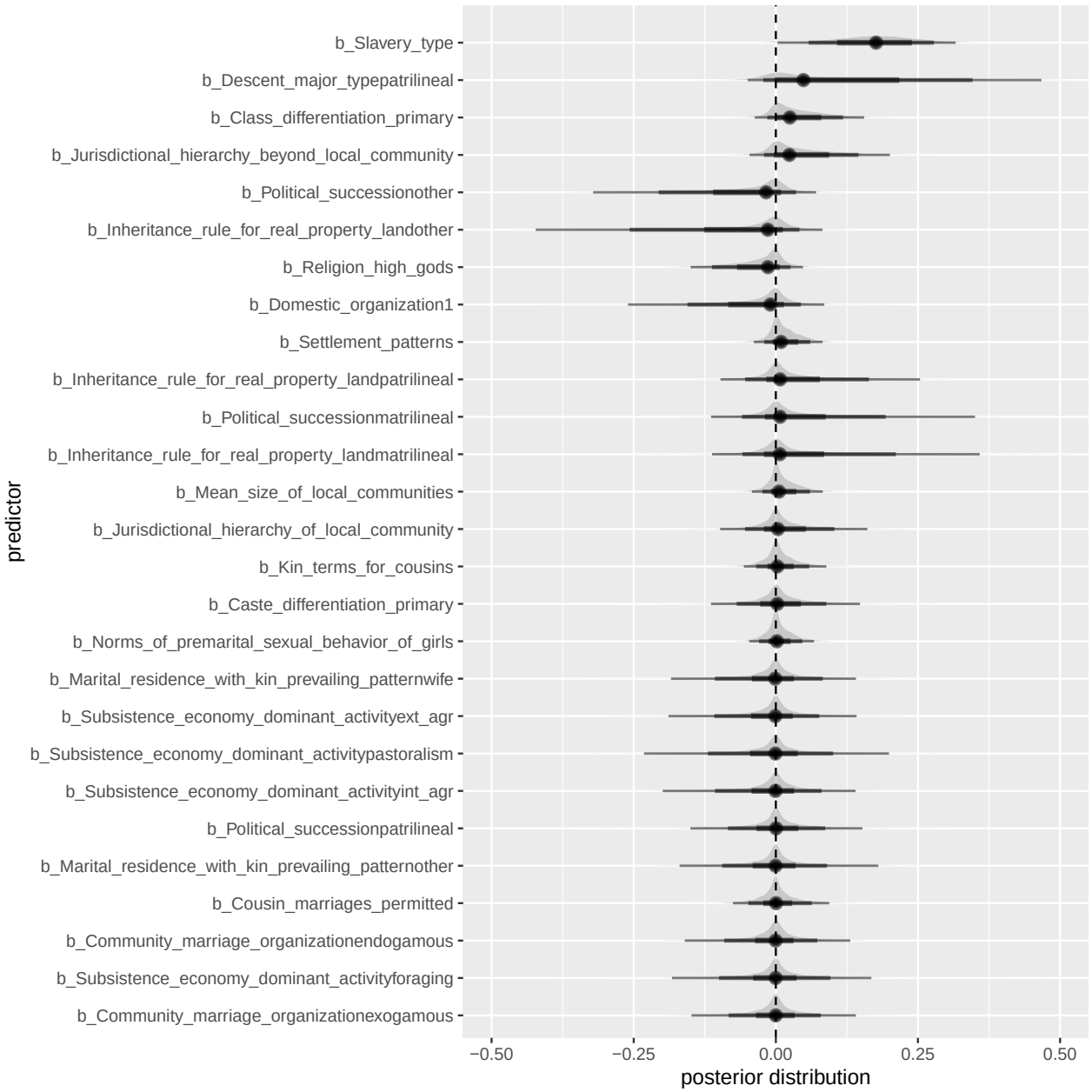
We use the WAMBS checklist to collect diagnostics on the best model, Fit 3c. This includes refitting the model using flat priors (Fit 3b unif), with 5000 iterations per chain (Fit 3b long), and with the order of observations in the data reshuffled (Fit 3b rand) and checking diagnostics of chain convergence.

This process leads us to a best model (Fit 3c) which uses Horseshoe priors and region as grouping factor.

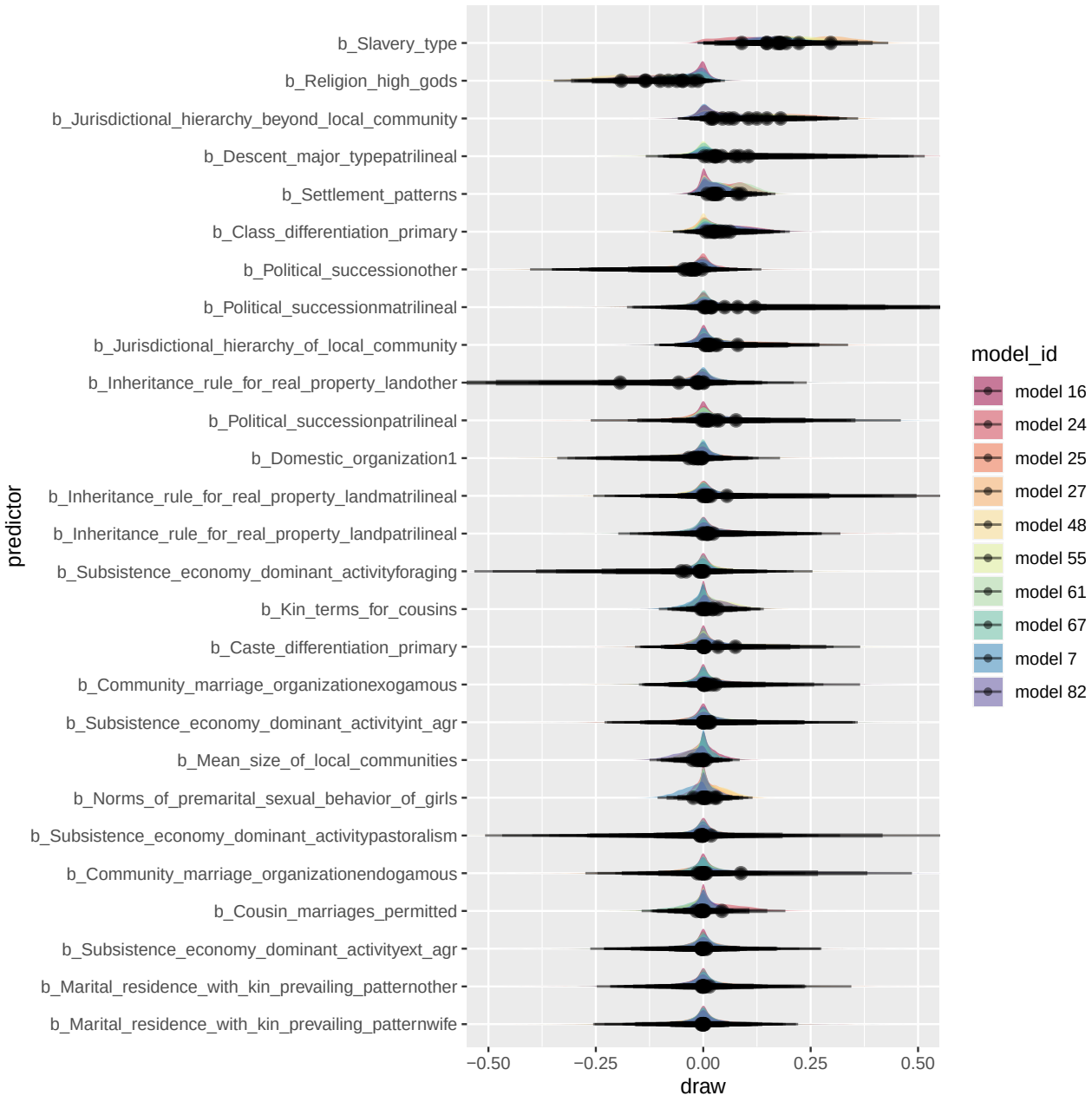
Robustness checks on imputed data

We take this model and fit it on ten random datasets from the 100 imputed datasets published by Thompson, Roberts, and Lupyan (2018).

First, we look at draws from the posterior distributions for our predictors from Fit 3c. The horizontal lineranges indicate the 66%, 89%, and 97% confidence intervals. 89% is the closest smaller prime number to the default 95% confidence intervals used in frequentist studies, while the other limits also have the advantage of being prime numbers (McElreath 2020). Predictors on the y axis are ordered according to absolute medians: those that shift the prior the most are on top.

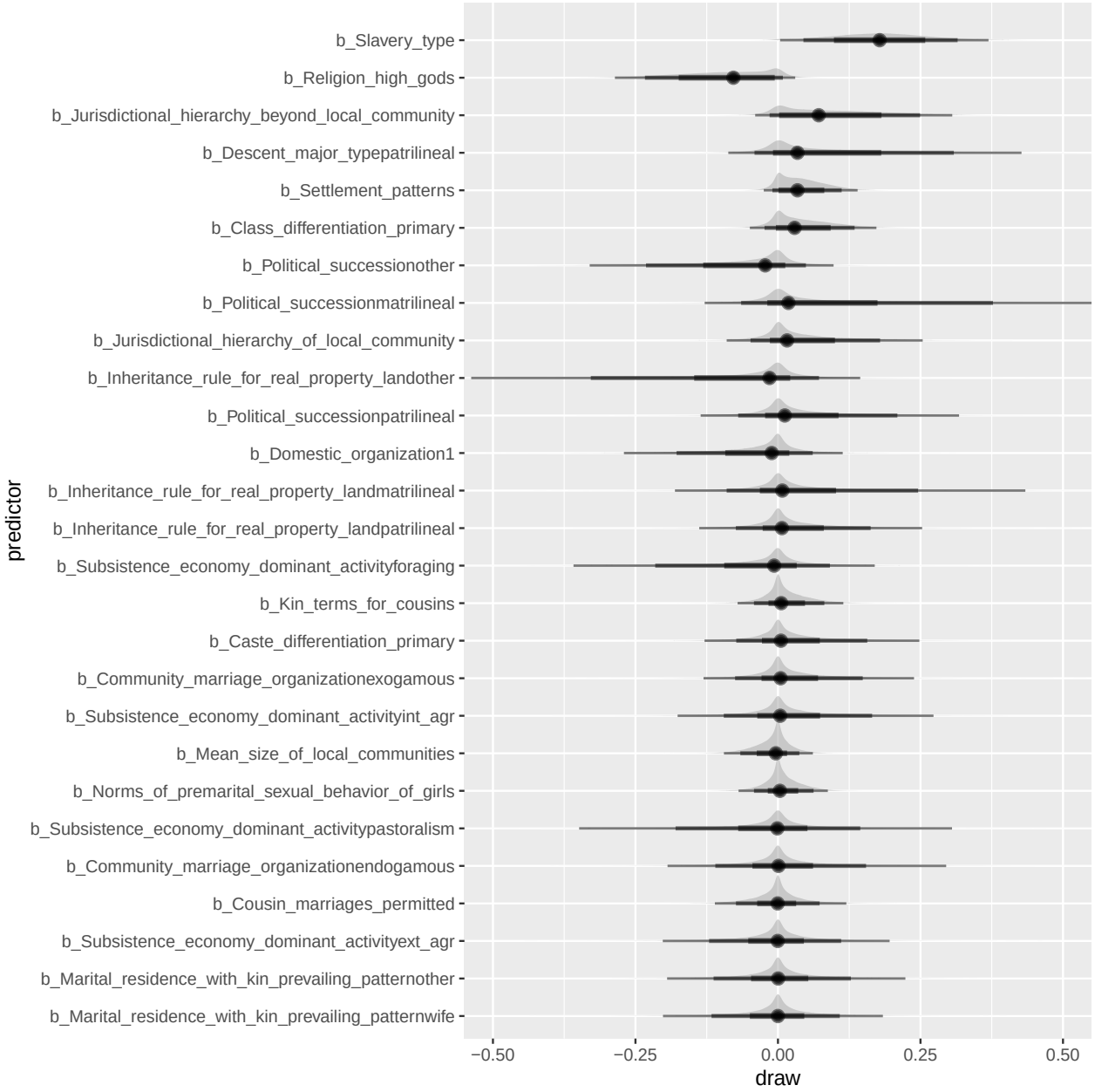


As we see, most predictors have relatively little to contribute in this sparse model, except for some notable exceptions.



This is the same plot of posteriors ordered by absolute median, except that each row plots ten distributions instead of one. This is a dense figure, but what the reader can garner is that most variation is visible for those predictors at the top. These are the ones that generall have non-zero effects.

The ten models are identical with the exception of the datasets. The missing values of the EA data were imputed by Thompson, Roberts, and Lupyan (2018) for each imputed dataset. We can assess the overall effects across these data by pooling the draws from the posterior distributions across models. The result can be seen below.



Altogether, across the ten models fit on the ten imputed datasets, most of our starting variables explain little variation in the outcome.

We pick the top six variables. Note that this is, to some degree, arbitrary. We pick the first five because the the 67% CI of the pooled distribution excludes zero. We pick class differentiation because it sounds interesting.

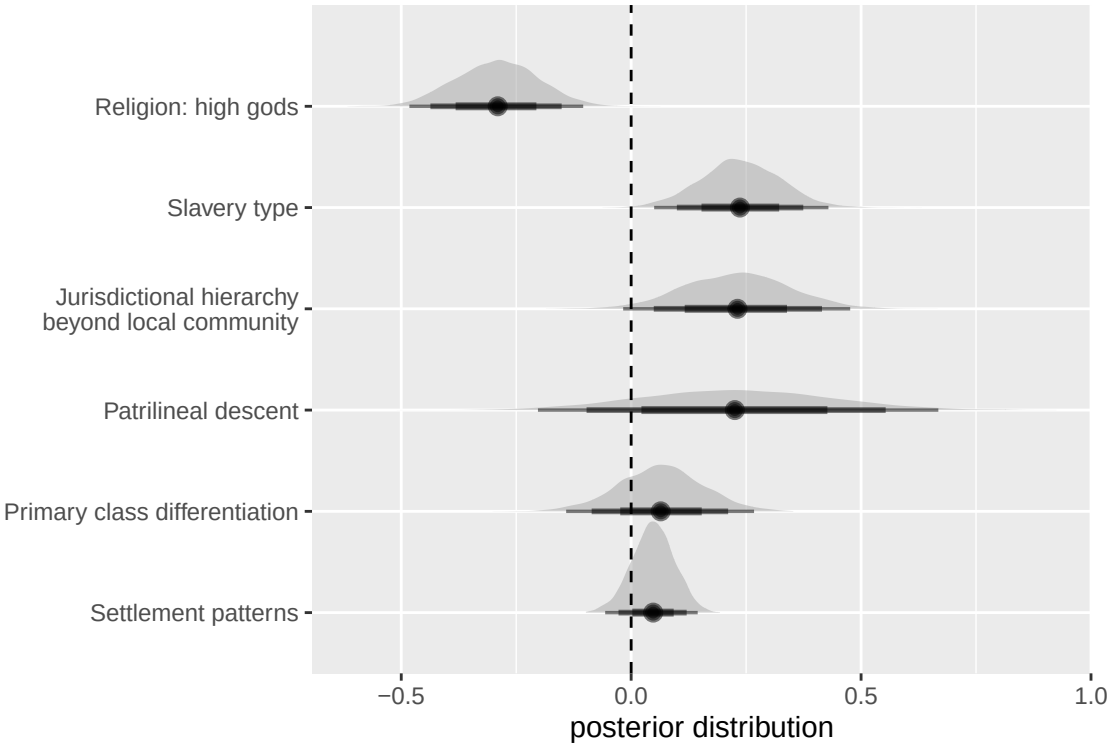
var_id	var_title
EA030	Settlement patterns
EA033	Jurisdictional hierarchy beyond local community
EA034	Religion: high gods
EA043	Descent: major type
EA066	Class differentiation: primary
EA070	Slavery: type

Refitting the model on real data

We refit the model on the complete observations of the EA (428/1291 societies). We use weakly informative student-t priors for both betas and intercepts. We are now done with variable selection and want to see how our model holds up on the real thing.

We check the model using the WAMBS checklist (selectively).

The posterior distributions of the model can be seen below.



We can see that we have strong evidence from the model on the effect of slavery, religion, and hierarchy beyond the local community on trance and possession states in the data. Evidence is weaker for the effect of descent, class differentiation, and settlement patterns. We can calculate the strength of the evidence. We have robust evidence (alpha = 0.03) for the first three, much weaker evidence (alpha = 0.33) for the remaining two.

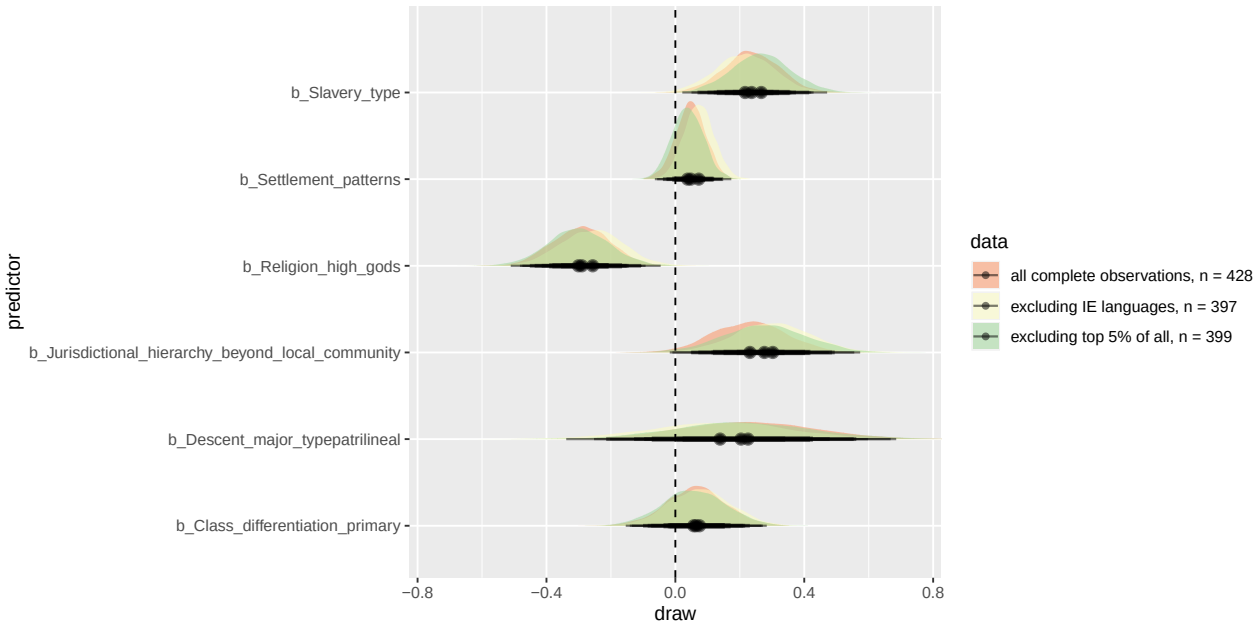
Hypothesis	Estimate	Est.Error	CI.Lower	CI.Upper	alpha
Religion high gods < 0	-0.29	0.09	-0.46	-0.12	0.03
Slavery type > 0	0.24	0.09	0.07	0.40	0.03
Jurisdictional hierarchy beyond local community > 0	0.23	0.11	0.02	0.44	0.03
Descent major typepatrilineal > 0	0.23	0.20	0.13	0.32	0.33
Class differentiation primary > 0	0.06	0.09	0.02	0.10	0.33

Robustness checks

We perform two robustness checks by filtering the dataset.

We filter the data excluding societies speaking languages in the Indo-European family and refit Fit 4 (Fit 5). We also filter the data excluding societies with population size in the highest 5% of natural log population size and refit Fit 4 again (Fit 6).

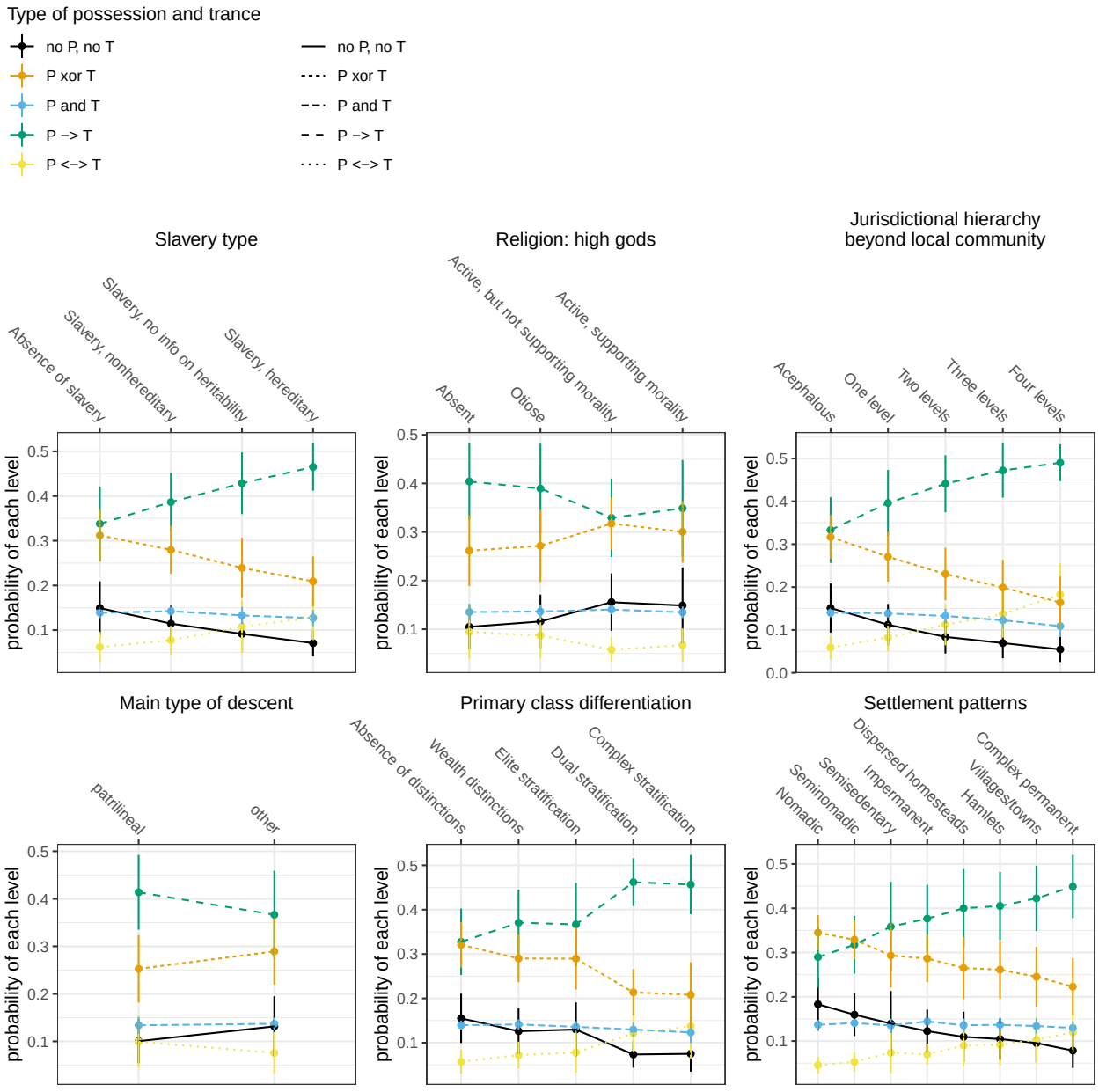
We draw posteriors for the estimates in these three models and visualise the results below.



Some effects are shrunk in the refit models, but the differences are not very large.

Visualising predictions

We take each predictor and draw predictions from Fit 4 for societies at each of its levels. Then, we calculate the mean and the standard deviation and plot these against each other, levels of the predictor on the x axis, predictions for the five levels of the outcome on the y axis.



References

Bürkner, Paul-Christian. 2018. "Advanced Bayesian Multilevel Modeling with the R Package brms." *The R Journal* 10 (1): 395–411. https://doi.org/10.32614/RJ-2018-017.

Carvalho, Carlos M, Nicholas G Polson, and James G Scott. 2009. "Handling Sparsity via the Horseshoe." In *Artificial Intelligence and Statistics*, 73–80.

Flom, Peter L, and David L Cassell. 2007. "Stopping Stepwise: Why Stepwise and Similar Selection Methods Are Bad, and What You Should Use." In *NorthEast Sas Users Group Inc 20th Annual Conference*, 11–14.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. "Visualization in Bayesian Workflow." *J. R. Stat. Soc. A* 182 (2): 389–402. https://doi.org/10.1111/rssa.12378.

Kay, Matthew. 2020. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151.

Kirby, Kathryn R, Russell D Gray, Simon J Greenhill, Fiona M Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E Blasi, et al. 2016. "D-Place: A Global Database of Cultural, Linguistic and Environmental Diversity." *PloS One* 11 (7): e0158391.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC press.

Murdock, George Peter. 1967. "Ethnographic Atlas: A Summary." *Ethnology* 6 (2): 109–236.

Piironen, Juho, and Aki Vehtari. 2016. "On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior." *arXiv Preprint arXiv:1610.05559*.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Stan Development Team. 2019. "RStan: The R Interface to Stan." http://mc-stan.org/.

Thompson, Bill, Sean Roberts, and Gary Lupyan. 2018. "Quantifying Semantic Similarity Across Languages." In *Proceedings of the 40th Annual Conference of the Cognitive Science Society (Cogsci 2018)*.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.