

1 Language and phoneme distributions

The distribution of the languages in the final dataset is shown in Figure 1. This figure shows that there are two main areas of high language density. One of these is in West Africa, with one hotspot around Ghana and another around Cameroon and Nigeria. These fall into the macro-area of the Sudanic belt (Clements & Rialland 2007). The other area of high density, which also comprises two major hotspots, encompasses the East and Rift valley areas (as well as the eastern part of the Sudanic belt), and includes parts of Ethiopia, Uganda, Sudan, South Sudan, Malawi, Kenya, and Tanzania.

It should be noted that this distribution is only an approximation. For instance, the southern tip of Africa is depicted as being empty of languages, which is obviously not the case. According to our map, Afrikaans is spoken somewhere around the border between South Africa and Zimbabwe, although in fact it is spoken all across South Africa, concentrated in the west and coastal areas around Cape Town. Another caveat is that some languages cover a much larger area than others, which is not reflected in this data. These points reflect a general limitation of point location data: it is necessarily a simplified version of reality.

The phoneme distribution is shown in Figure 2.

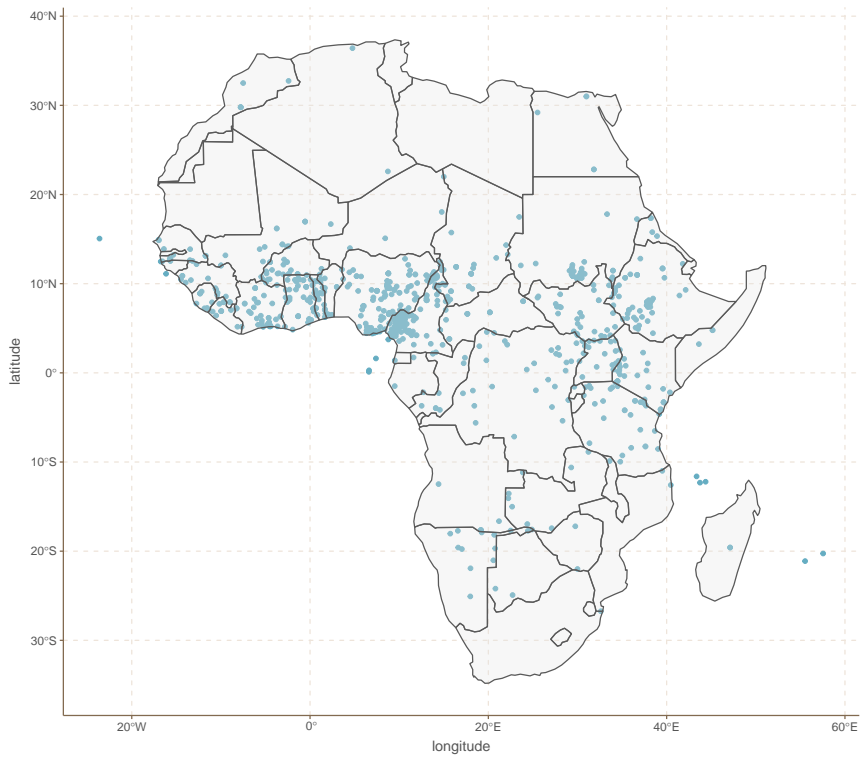


Fig. 1: Language distribution

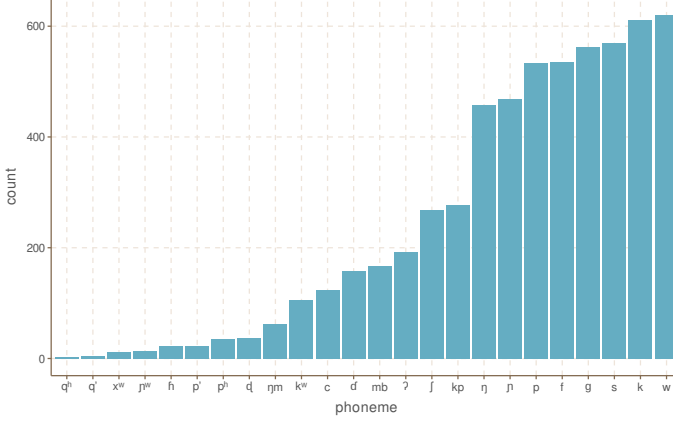


Fig. 2: Phoneme distribution

2 Multi-probit model

The model is easiest to explain in the case when we have two binary outcomes: y_1 and y_2 (for example, the presence or absence of /k/ and /p/). To model both outcomes simultaneously in a multivariate probit model, we estimate two latent (i.e. not directly observed) variables, y^*_1 and y^*_2 , as coming from a multivariate normal distribution:

$$\begin{bmatrix} y^*_1 \\ y^*_2 \end{bmatrix} \sim \text{MultiNormal} \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1^2\sigma_2^2 \\ \rho\sigma_1^2\sigma_2^2 & \sigma_2^2 \end{bmatrix} \right) \quad (1)$$

with the constraints:

$$y_1 = \begin{cases} 1, & \text{if } y^*_{11} \geq 0 \\ 0, & \text{if } y^*_{11} < 0 \end{cases} \quad (2)$$

$$y_2 = \begin{cases} 1, & \text{if } y^*_{22} \geq 0 \\ 0, & \text{if } y^*_{22} < 0 \end{cases} \quad (3)$$

where μ_1 and μ_2 , and σ_1 σ_2 are the means and standard deviations of y^*_{11} and y^*_{22} , respectively, and ρ is their correlation. This model can be extended to arbitrarily many outcomes.

The full model contains a phylogenetic term for each outcome, and a partially shared gaussian process. With these two terms the rest of model specification is as follows:

$$\mu_1 = \alpha_1 + xi_1 + \eta_1 \quad (4)$$

$$\mu_2 = \alpha_2 + xi_2 + \eta_2 \quad (5)$$

$$xi_1 \sim Normal(0, \sigma_{p1}^2 \Sigma_1) \quad (6)$$

$$xi_2 \sim Normal(0, \sigma_{p2}^2 \Sigma_2) \quad (7)$$

$$\eta_1 \sim MultiNormal(0, \Sigma_{GP}) \quad (8)$$

$$\eta_2 \sim MultiNormal(0, \Sigma_{GP}) \quad (9)$$

$$\Sigma_{GP} = K(x|\lambda, \delta, D) \quad (10)$$

$$\lambda \sim N(10, 8) \quad (11)$$

$$\rho \sim N(0, 2) \quad (12)$$

$$K_{j,i}(\lambda, \delta, D) = \delta^2 \exp \left(-\frac{D_{j,i}^2}{2\lambda^2} \right) + \delta^2 \quad (13)$$

Where ξ_1 and ξ_2 are varying intercepts centered around 0, with σ_{p1}^2 and σ_{p2}^2 are the standard deviation of the phylogenetic intercepts; and Σ_1 and Σ_2 are the phylogenetic correlation.

The Gaussian Process is captured by η_1 and η_2 , which are latent variables sampled from a Multivariate Normal distribution, and with the spatial correlation in Σ_{GP} . The point here, is that even though η_1 and η_2 are different, Σ_{GP} is the same for both. Σ_{GP} is as a function of two parameters, the length-scale λ and the standard deviation δ . Finally, D is a distance matrix between all points, and $D_{j,i}$ is the distance between points j and i .

3 Results

3.1 Correlation uncertainty

Figure 3 shows the uncertainty of the correlation values. This plot supplements Figure 6 in the paper. It shows that for most correlation estimates, there is considerable uncertainty about the estimates.

3.2 Phylogenetic effects

Exploring the phylogenetic effects is not as straightforward as the areal patterns because the model estimates an individual intercept for each language and outcome.

Figure 4 shows the group-level intercepts for some languages belonging to three families: Ijoid, Kadugli-Krongo, and Ta-Ne-Omoti. The important point here is that although we estimate an individual intercept for each language, we also impose a strict covariance structure between these, which means that the model does not allow the intercepts to vary greatly from the mean of other closely related languages. This can be clearly seen for a phoneme like /f/. Because all the languages in the Ijoid and Kadugli-Krongo families have /f/, the group-level intercepts for these languages are all close to 2 (a positive value means that it is likely the language will contain the phoneme, while a negative value means it is likely that the language will not contain the phoneme). In contrast, for the Ta-Ne-Omoti family, only four languages have /f/, and these four languages have intercepts close to -1.

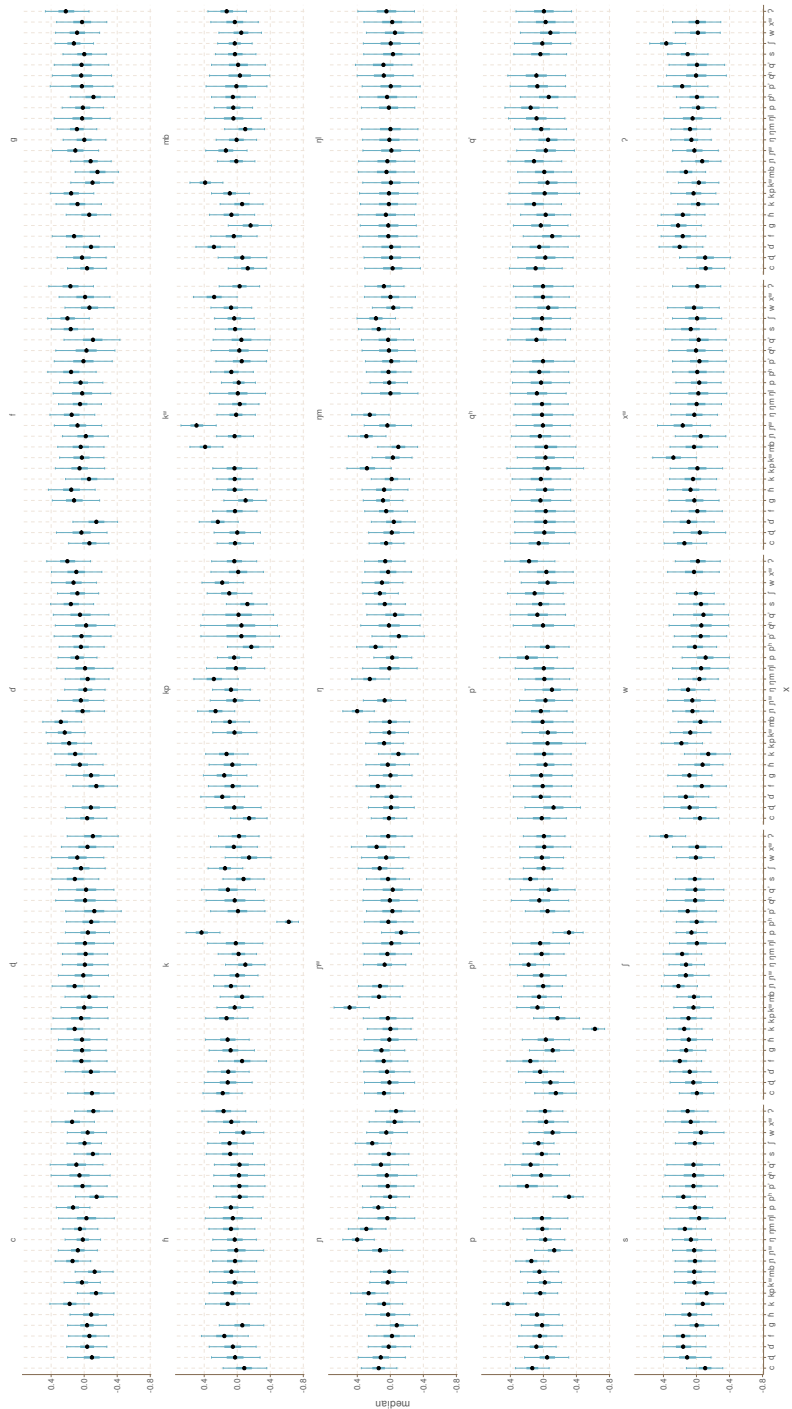


Fig. 3: Posterior correlation effects

The languages in Ta-Ne-Omoti which do not have /f/ have a higher group-level intercept, but its value is closer to 0, not 2. The effect of constraining variation within families is that there is shrinkage to the mean, such that the group-level intercepts of all Ta-Ne-Omoti languages are close to each other, but each individual language is allowed some variation. How much each language can vary from its closely related languages depends on the tree and the behaviour of the languages in it, as well as the areal effects. Overall, including a phylogenetic term is more flexible than simply adding family or genus as a group-level effect, while avoiding overfitting.

Another interesting result is the standard deviation (sd) of the phylogenetic effects. The greater the sd of the phylogenetic effects, the more variance the model assigns to the phylogenetic term. Therefore, it is interesting to compare the sd of the phylogenetic term for each phoneme for the model with and without areal effects. This comparison is shown in Figure 5. This plot shows, in dark brown, the sd for the model which has both phylogenetic and areal effects, and in yellow, the sd for the model including only phylogenetic effects. The plot also shows the 50 (in thick lines) and 95 (in thin lines) uncertainty intervals for the estimates. For some phonemes, like /f/, /s/ or /w/, the sd of both models is almost identical, while for other phonemes like /kp/, /mb/ and /p/, the sd varies considerably between the different models. This change in sd happens when a portion of the variance which the first model assigns to the phylogenetic effects ends up being explained by the areal effects in the second model. This does not immediately suggest that the phoneme in question has or does not have strong areal effects, but rather whether the areal patterns found for that

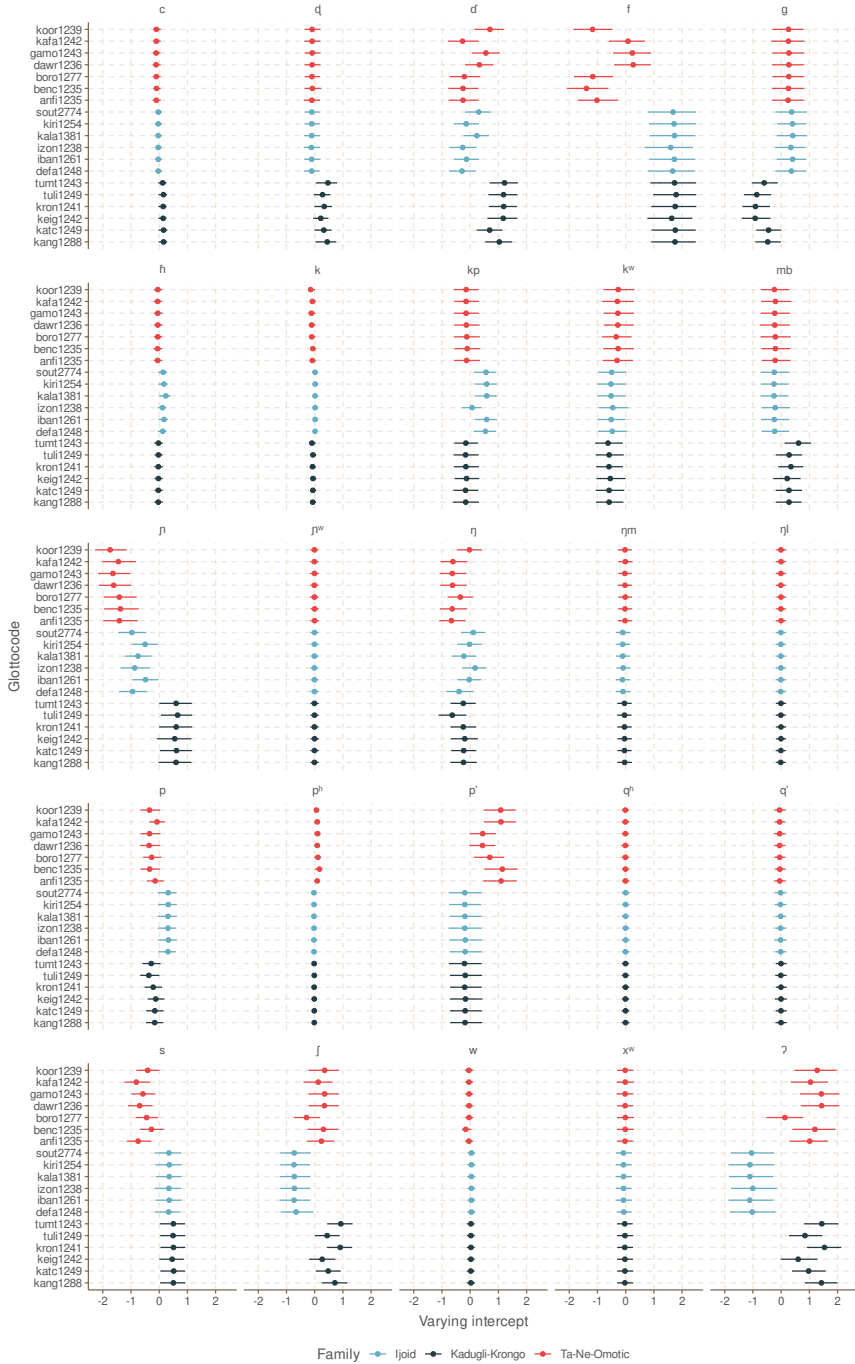


Fig. 4: Phylogenetic effects for Ijoid, Kadugli-Krongo, and Ta-Ne-Omotic

phoneme overlap with the genetic effects. Figure 5 shows that for phonemes like /f/ or /s/, the areal effects are either expected to be very weak (like /s/, see the additional plots in the zenodo repository), or largely non-overlapping with the genetic effect (like /f/, see the additional plots in the zenodo repository). In contrast, for phonemes like /kp/ or /mb/, we expect that these have very strong effects, as we see for /kp/ (in the paper), or areal effects largely overlapping with the genetic effects (like /p/, in the paper). This result is interesting because while we cannot fully disentangle contact and genetic patterns, we can estimate the degree of overlap between the two for any given outcome.

4 Model evaluation and cross-validation

The main purpose of this paper is not to develop the best possible predictive model. Nonetheless, it is often useful to evaluate the performance of the model using prediction. To do this, we performed 5-fold cross-validation.¹ In cross-validation, we split the dataset into groups and train the model on all but one group and then predict the group that was left out. We then repeat this process for all the groups. Figure 6 shows the balanced accuracy for each phoneme

¹ The usual number of iterations in cross-validation is 10. The main reason why we decided to do 5 iterations instead of 10 is computational time. Each model at a sample size of 90% of the data takes about a week to run. With a sample size of 80% of the data, each model takes a couple of days. Doing 10-fold cross-validation would have taken several months.

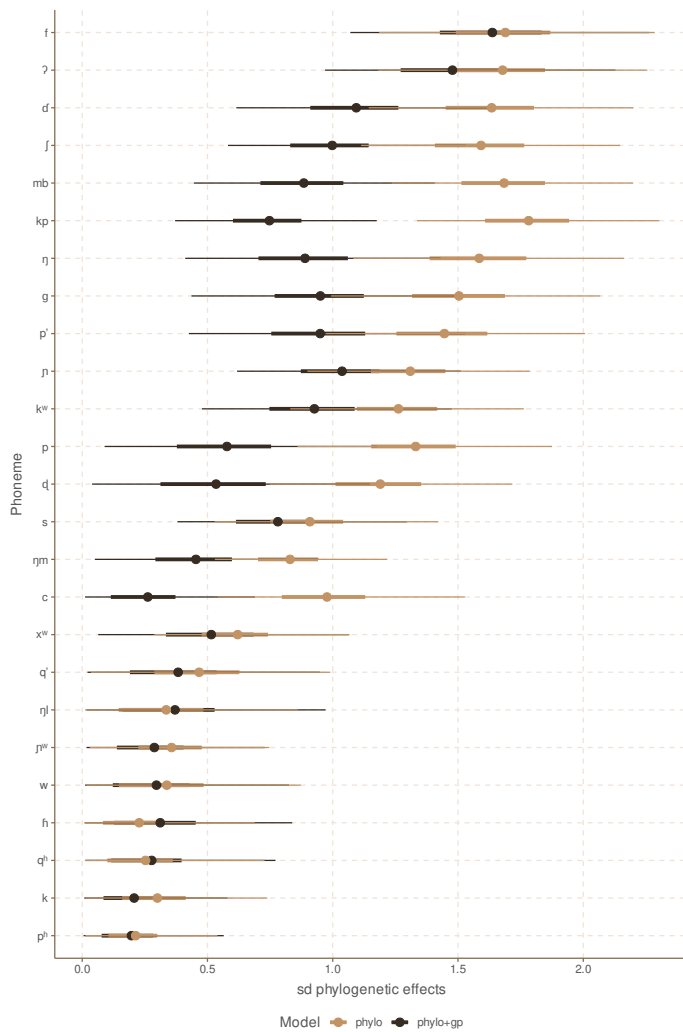


Fig. 5: Standard deviation of phylogenetic effects

for each model.^{2,3} What we see is that there is no one model which performs consistently better than the other models. For reference, this figure also shows the proportion of each phoneme in our dataset on the right y -axis. For some phonemes, like /kp/ or /d/, the model without areal effects is clearly better, while for /p'/, the model with only areal effects is better, and for /ŋ/ the model including both areal and phylogenetic effects is better. In terms of mean balanced accuracy across all phonemes, the model with areal and phylogenetic effects and the model with only phylogenetic effects are tied at 0.58, the model with only areal effects comes in third at 0.57, and the model without predictors has an expected balanced accuracy of 0.5. This suggests that phylogeny and area are equally good predictors of phoneme distributions in general, although this naturally depends a great deal on the specific phoneme. Another (perhaps expected) observation is that phonemes which are either very common or very rare in the dataset were harder to predict overall than phonemes with a more balanced distribution. This is because there will often be no neighbourhood or areal information that helps the model to identify which languages have rare phonemes or lack common ones.

2 The balanced accuracy is useful to evaluate a binary classifier when the two classes are imbalanced. In our case, since some phonemes are either very rare or very common, a simple accuracy metric would be biased. The balanced accuracy is calculated as (sensitivity + specificity)/2. The balanced accuracy is 0.5 if the classifier does not perform better than simply predicting the majority class.

3 An alternative method for model predictive evaluation in Bayesian statistics is ELPD (estimated log predictive density) as discussed by Vehtari et al. (2017). For our case study, we could not use this technique because there is no straightforward way of calculating the ELPD for multivariate probit models.

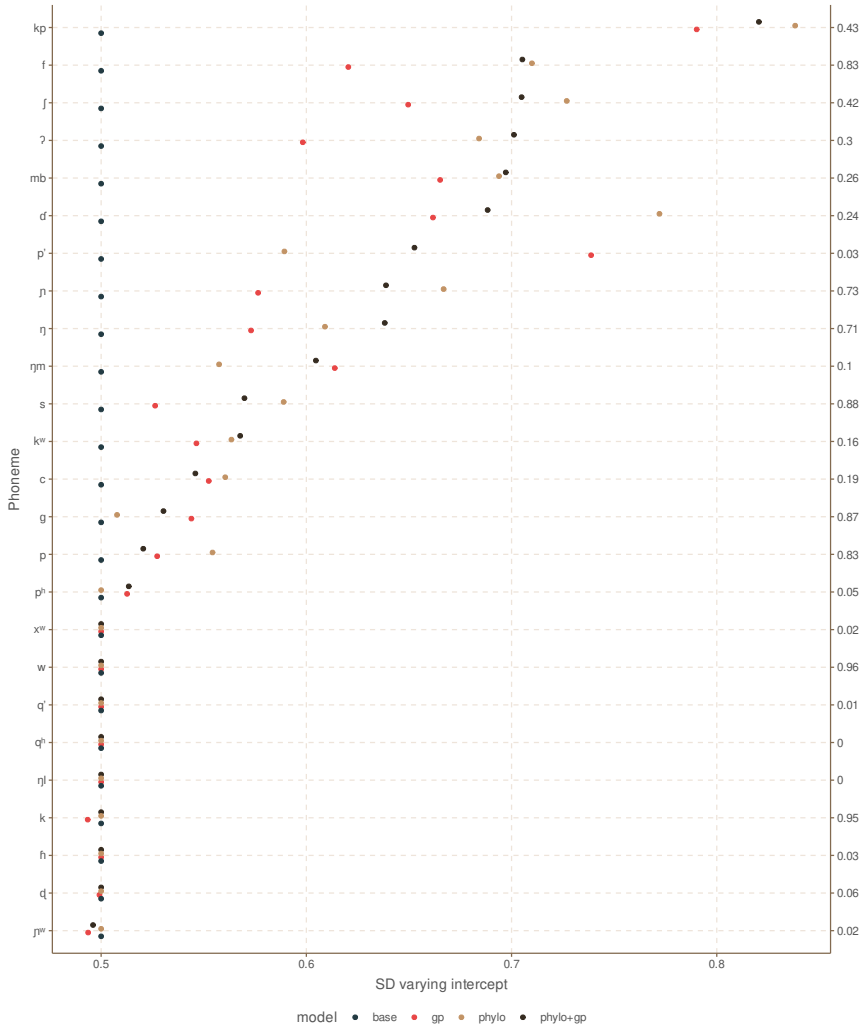


Fig. 6: Model accuracy by phoneme

While we do not explore this model here, it is likely that having independent GPs for each phoneme would improve the overall accuracy of the model. However, such an approach would take us in a different direction. Perhaps a better approach would be to study phonemic *categories* rather than specific segments, as this would allow us to capture more general information about phoneme inventories without making the model computationally intractable.

References

- Clements, G. N. & Annie Rialland. 2007. Africa as a phonological area. In Bernd Heine & Derek Nurse (eds.), *A Linguistic Geography of Africa* (Cambridge Approaches to Language Contact), 36–85. Cambridge: Cambridge University Press.
- Vehtari, Aki, Andrew Gelman & Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27(5). 1413–1432.