

Commentary

Harald Hammarström*

Commentary: Replication, robustness or methodological competition?

<https://doi.org/10.1515/lingty-2025-0041>

Received April 16, 2025; accepted May 3, 2025; published online July 30, 2025

The authors (Becker and Guzmán Naranjo 2025) frame their study as replication (of the methods part) of four previous studies in quantitative typology. They have adopted a rather broad definition of the term replication, and what they do more specifically is to test the robustness of these studies against a different method. They characterize robustness as the stability of the result of a study “across different methods used for analysis” and “under variations in ... experimental procedures”. Unfortunately no qualification to “variations” or “different” is provided, but surely robustness to just any different method is not interesting. If a method is known to be better than another (for a given question), robustness to the poorer method is not relevant. Presumably the authors mean that if there are different reasonable methods where it is not settled which one is better (for a given question at the current state of knowledge), then robustness is of value. Ideally we should find the better method as far as possible and resort to robustness only in the special case that several methods are equally arguable. Most of the previous work highlighted as replications for robustness in typology would rather count as invalidated¹ using a more suitable method, namely one with more appropriate controls for independence against a previous one with weaker or no such controls.

Rather than robustness, the primary question should be whether or not the method used by the authors is better (for the cases at hand). Indeed, the authors reveal this stance too when interpreting differences in the results.² But are the bias control methods employed by the authors arguably better than their predecessors?

Phylogenetic regression is a relatively well-understood family of methods (see references cited by the authors) and conforms to the theory behind language family

¹ Including Everett et al. (2016) which appears to have been misread by the authors.

² As evidenced in phrasings such as “actual proportions”, “little evidence for the conclusion drawn by ...”, “method for phylogenetic bias used by ... likely overestimated the proportions”.

***Corresponding author: Harald Hammarström** [*ˈhɑːrald ˌhamːaːsʈroem*], Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden, E-mail: harald.hammarstrom@lingfil.uu.se

relationships. If the language family relationships are known, it is in principle the appropriate approach and can be argued a priori – with no robustness experiments needed – to trump previous work without rigorous genealogical control. It also has an edge over the classic one-per-family sampling strategy in making full use of any sample with more than one datapoint from some family/families. The authors, as is common, assume classificationary tree topologies, but do not try to estimate branch lengths from lexical (or any other) data with the hazy motivation that “our experience is that these methods do not improve model performance”. But different branch lengths are found in reality. It remains an open question whether any error introduced by truncating branch lengths is a good trade-off against the weaker use of information in the classic one-per-family sampling strategy.

The authors employ a Gaussian process (GP) to attempt to control for contact bias. GPs have cubic computational complexity which is borderline prohibitive but are mathematically well-understood, allow for a non-linear decay of influence with increasing distance, and provide confidence intervals “for free” compared to other supervised Machine Learning methods. However, as employed in the replication study, the same hyperparameters apply across the entire sample, which means that the influence decays the same way per kilometer (or, actually, per latitude/longitude degree) over the whole world. That is, a certain distance, say 320 km, implies a specific amount of influence, labeled “strong” in Guzmán Naranjo and Mertner (2023: 468) regardless of how many in-between languages that distance covers. The approach is helped somewhat if the in-between languages happen to be in the sample (Guzmán Naranjo and Becker 2022: 630), but the solution remains a weak match with the reality of language density. Furthermore, similar to other approaches, it is only very vaguely connected to current theory and knowledge of language contact and what aspects thereof it seeks to model (Guzmán Naranjo and Becker 2022: 639–630). Is it recent contact? Deep contact? The average net effect of both as well as intensity of contact?

The authors rightly emphasize transparency as a key requirement for methodological development in the field and suggest *Guidelines for better replicability in typology* (Appendix D). However, the suggestions (D.3) ascribe a significant role to the exact source code used for the statistical analysis. Scientific fields which deal routinely with algorithms and exact specifications, such as computer science, long abandoned the idea of utilizing the source code as a sustainable specification of a method or algorithm. Source code requires the knowledge of a specific programming language, necessitates the specification of distractive implementation details (such as whether something is a list or hash table, a while- or for-loop, etc.), and goes obsolete relatively quickly. Far more to the point, general and sustainable is a method specification in pseudocode. Thus, for transparency, we rather need a high level description of the data and methods. The source code may be helpful for immediate replication but is not sufficient, not necessary and not a substitute for an

intelligible and complete description of a method. Furthermore, in papers such as the present one, where the description of the method(s) is spread out in bits and pieces in the paper, cited papers, appendices, R code packages, a concise summary of the input variables as used by the authors for the study at hand (rather than more generally) and the output values obtained would do much for transparency. The list of input variables should explicitly include defaults provided by the packages invoked and separate those variables which are model specification and those which are performance parameters.

In conclusion, the more important question is if the bias control method by the authors is better than its alternatives. If it is not, e.g., if it worse, is it of little interest that one or the other result is robust to it. Furthermore, the bias control method by the authors is more complex than most of its alternatives and we lack model comparisons that take this into account (e.g., using Akaike information criterion or a similar measure).³

References

- Becker, Laura & Matías Guzmán Naranjo. 2025. Replication and methodological robustness in quantitative typology. *Linguistic Typology*. <https://doi.org/10.1515/lingty-2023-0076>.
- Everett, Caleb, Damián E. Blasi & Seán G. Roberts. 2016. Language evolution and climate: The case of desiccation and tone. *Journal of Language Evolution* 1(1). 33–46.
- Guzmán Naranjo, Matías & Laura Becker. 2022. Statistical bias control in typology. *Linguistic Typology* 26(3). 605–670.
- Guzmán Naranjo, Matías & Miri Mertner. 2023. Estimating areal effects in typology: A case study of African phoneme inventories. *Linguistic Typology* 27(2). 455–480.

³ There are also overfitting concerns, e.g., the `m_gp + m_phylo` manifests worse leave-one-out predictions than the special case `m_phylo`, although the difference is rather small.