# 7 Exploring the Geographical Distribution of Missing Data Using Approximate Gaussian Processes

Miri Mertner[1], Matías Guzmán Naranjo[1]

[1] Institute of Linguistics, University of Tübingen

Gaussian processes (GPs) have several qualities that make them well-suited to spatial statistics, as they allow us to add non-linear effects to a model in a flexible way (see e.g. McElreath, 2020, Chapter 14, for an explanatory example). A GP essentially estimates the effect that every observation has on every other observation in the form of a covariance matrix, which can then be used, for example, as a predictor in a model. In linguistic typology, they have been used as a way to control for spatial autocorrelation between languages, as well as for inferring probable ranges of contact between languages (Guzmán Naranjo & Mertner, 2022). However, they can be prohibitively slow to use with large datasets, such as the global sample of languages included in WALS or Glottolog. Therefore, in order to use them on such large datasets, an approximation of the GP is required.

One of the cases in which a large dataset is necessary to make meaningful inferences is in the exploration of the distribution of missing data in linguistic databases such as WALS (Dryer & Haspelmath, 2013) and ASJP (Wichmann et al., 2022). Using approximate GPs implemented in the programming language Stan (Stan Development Team), the present study will focus on uncovering areal biases in the distribution of missing linguistic data. Geographical and social correlates which could help explain the causal factors behind a higher or lower density of missing data in a particular area will also be tested, such as landscape roughness, climate, and population size. A better understanding of the factors which lead to geographical imbalances in the distribution of missing data could, among other things, improve our ability to impute missing data as part of statistical modelling work.

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, accessed on 2023-01-01.)

Guzmán Naranjo, Matías and Mertner, Miri. 2022. Estimating areal effects in typology: a case study of African phoneme inventories. Linguistic Typology. https://doi.org/10.1515/lingty-2022-0037

McElreath, R. (2020). Statistical Rethinking: A Bayesian Course with Examples in R and Stan (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315372495

Stan Development Team. 2022. Stan Modeling Language Users Guide and Reference Manual, 2.31. https://mc-stan.org

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown (eds.). 2022. The ASJP Database (version 20).