Controlling for geographical, areal and family biases

Matías Guzmán Naranjo & Laura Becker
Université de Paris & Universität Erlangen Nürnberg

Introduction

Sampling and controlling for areal and family biases has always been an important part of language typology. Most sampling methods have focused on establishing principled ways of selecting languages so that it is possible to formulate crosslinguistic generalizations from a small number of languages without being biased by family, areal, and geographic effects. Prominent sampling methods have been proposed by Bakker (2012), Bickel (2008), Dryer (1991, 1992, 2009, 2011), Dunn et al. (2011), Greenberg (1966), and Song (2014). An important related question that has received much attention from the typological literature concerns the definition and establishment of linguistic areas, both on linguistic features and non-linguistic features as well as a combination thereof (Aikhenvald and Dixon, 2006; Bickel and Nichols, 2006; Dahl, 2001; Donohue and Whiting, 2011; Enfield, 2005; Hammarström and Donohue, 2014).

Controlling for areal effects on a global scale has not been as widely studied as family effects, but it often relies on similar techniques: selecting a subset of languages in a way that minimizes language contact and larger areal effects between those languages. One important point made in (Dryer, 2018) is that distance between languages should be viewed as a relative measure and cannot be accounted for in absolute terms. Two languages with a distance of 100 kilometers between them may be in a contact situation in Siberia but not in the Amazon region, the Caucasus or Papua New Guinea. In order to control for areal effects on a global scale, most large-scale typological studies can only afford to include a simplified approach to geographical biases, using macro-areas or other large and coarse geographical splits. While all these methods provide reasonable ways of controlling for family and areal effects, they all face the same issue: the researcher can only include a portion of their data, and categorical splits in terms of language family and area are required. In this talk we present three statistical methods which allow us to include all data available while controlling for the effects mentioned.

Materials and methods

We propose techniques that have not yet been used in typology for controlling for geographical, areal and family biases. Our methods can be applied to (e.g.) regression models of universals. For instance, in order to test for universal correlations between OV/VO and noun-genitive orders, we can use the OV/VO orders of the languages in the sample to predict their noun-genitive orders.

To control for geographical biases in the sample we include latitude and longitude information of each language directly into our model. We do this by using a bivariate spline. Instead of manually...
dividing the sample into macro-areas, the bivariate spline automatically detects those geographical
zones that have an effect on the predicted feature (e.g. the noun-genitive order).

To control for areal biases we first define the closest neighbor of language L1 as the language which is
closest to L1. We then use the distance between L1 and its closest neighbor to define other neighbors
of L1. Any language which is at most four times as far from L1 as its closest neighbor of L1 will be
counted as neighboring language. In other words, if L1 has its closest neighbor at a distance of 10 km,
all languages within a 40 km radius will be considered neighbors of L1. However, if the nearest
neighbor of L1 is 100 km away, all languages within a 400 km radius will be considered neighbors of L1.
Having determined the neighbors of each language in the sample, we can include the behavior of the
neighboring languages regarding the predicted feature (e.g. the noun-genitive order) and estimate to
what extent the value of each language depends on the behavior of its neighbors.

Finally, to control for family biases, we include a phylogenetic term in our regression model. This
means that instead of simply grouping languages into families or genera (at any given level), we can
include the complete (known or assumed) genealogical structure between all the languages in the
sample. This is important because even though genera are defined in a way that they ideally represent
a comparable time depth, we know that families or genera are not entirely comparable. Thus, including
the entire known genealogical relations into the model provides a more accurate representation of
family effects.

Results

We apply the methods described above to three well known cases from the literature: the relation
between OV/VO order and noun-genitive order; family, geographical, and areal effects on the size of
the phoneme inventory; and the prevalence of prefixes vs. suffixes. In all three cases we find that areal
effects are the strongest, meaning that languages tend to act like their neighbors. Additionally, our
approach allows us to quantify the individual importance of these factors. For instance, for the size of
phoneme inventories, both geographical and family effects play a major role, for word order, on the
other hand, we find that family and geographical effects do not a large impact on the relation between
the two-word orders.

References

AIKENVALD, ALEXANDRA Y. AND ROBERT MALCOLM WARD DIXON (2006). Areal Diffusion and Genetic Inheritance:
Press.

BAKKER, DIK (2012). “Language sampling”. In: The Oxford handbook of linguistic typology. Ed. by JAE JUNG SONG
AND DIK BAKKER. Oxford University Press.

BICKEL, BALTHasar (2008). “A Refined Sampling Procedure for Genealogical Control”. In: Sprachtypologie und Uni-
versalienforschung 61, pp. 221–233.


