LABIAL–VELAR STOPS ARE AREAL RETENTIONS BUT GENEALOGICAL INNOVATIONS IN THE NIGER–CONGO LANGUAGES

Dmitry Idiatov & Mark Van de Velde

LLACAN (CNRS – INALCO)

dmitry.idiatov@cnrs.fr

mark.vandevelde@cnrs.fr

- Look for **interesting correlations** in the distribution of values of various linguistic features **in space**
- Try to find **plausible explanations** in terms of **scenarios** which would imply concrete mechanisms of linguistic change (also using data from other disciplines)
- Explanations are fundamentally **diachronic**

  "a theory of why languages are the way they are is fundamentally a theory of language change…" (Dryer 2006:56).

PREFERRED APPROACH

- Following the **methodology** developed in:

  Idiatov, Dmitry. 2018. An areal typology of clause-final negation in Africa: language dynamics in space and time. In Daniël Van Olmen, Tanja Mortelmans & Frank Brisard (eds.), *Aspects of linguistic variation*, 115–163. Berlin: De Gruyter Mouton.

  Idiatov, Dmitry & Mark L.O. Van de Velde. 2021. The lexical distribution of labial-velar stops is a window into the linguistic prehistory of Northern Sub-Saharan Africa. *Language* 97(1). 72–107.

PREFERRED APPROACH: CONCEPTUALLY

- bottom-up
- big data
- garbage in, garbage out
- let the data speak for themselves (☹ binning)
- non-binary
- spell out the rules first

RPREFERRED APPROACH: PRACTICALLY

- Use the **databases that exist** to harvest the data (depending on the feature of interest: **RefLex**, Phoible, Geonames…)
- **Enrich** the harvested data with manually collected data if need be
- **Clean** and **format** the data given research questions and hypotheses and your theoretical assumptions
- Visualize the data **with different visualization methods** to confirm that the results are **qualitatively robust**

VISUALIZATIONS: SPATIAL ANALYSIS

- **deterministic** methods
  - **spatial interpolation by IDW** (inverse distance weighting): exact, finer structure
  - **spatial interpolation by Kernel smoothing** : inexact, general trends
- **statistic** (non-deterministic) methods, such as
  - GAM (generalized additive modeling)
  - GAMM (+ mixed)

- **Advantages** over deterministic methods:
  - a non-deterministic model that describes **a distribution of possible outcomes**
  - **more stable** to variations in the quantity and quality of the data
  - provides **quantified results**
  - comes with **coefficients** that allow for a more objective evaluation of the visualizations
  - can help to **discover patterns** in the data

STATISTIC VISUALIZATION: GAM

- What is GAM?: an extension of multiple regression that provides flexible tools for modeling complex interactions describing wiggly surfaces
  - **regression**
  - wiggly surfaces
  - thin-plate splines
- A powerful tool, but still with some **limitations**
  - type of the distribution of the data (especially, non-Gaussian distributions)
  - Abrupt changes of the dependent value

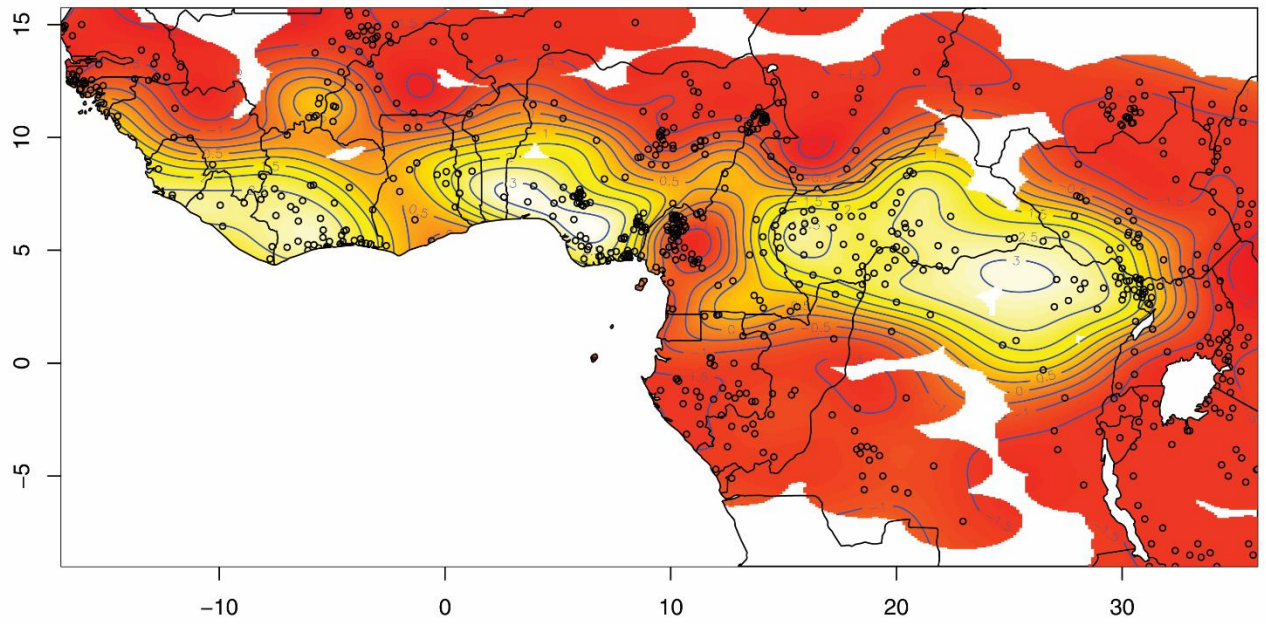
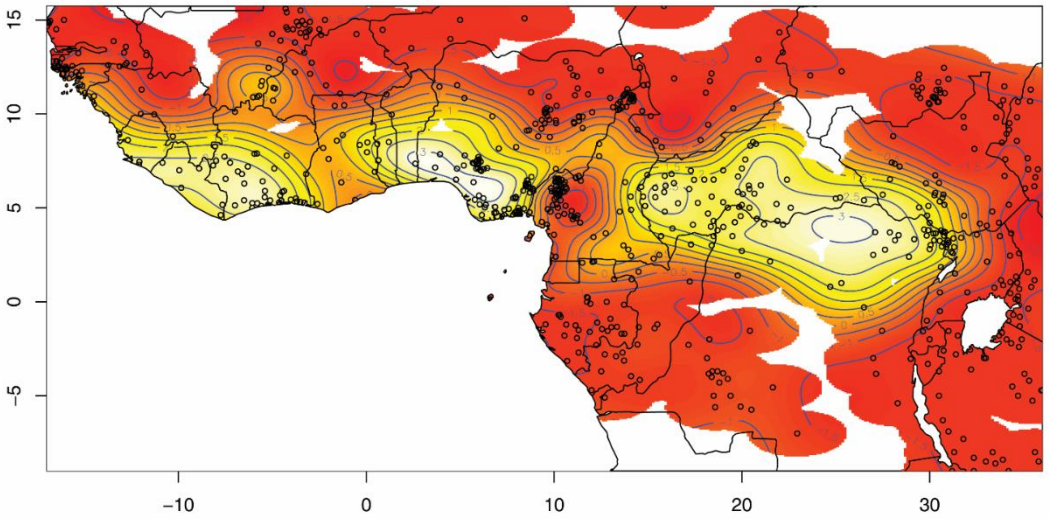
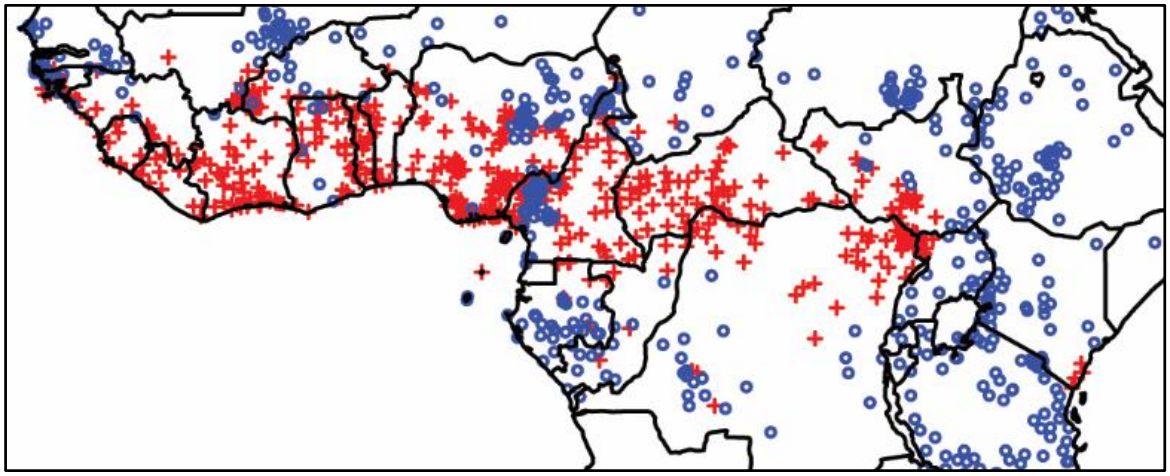


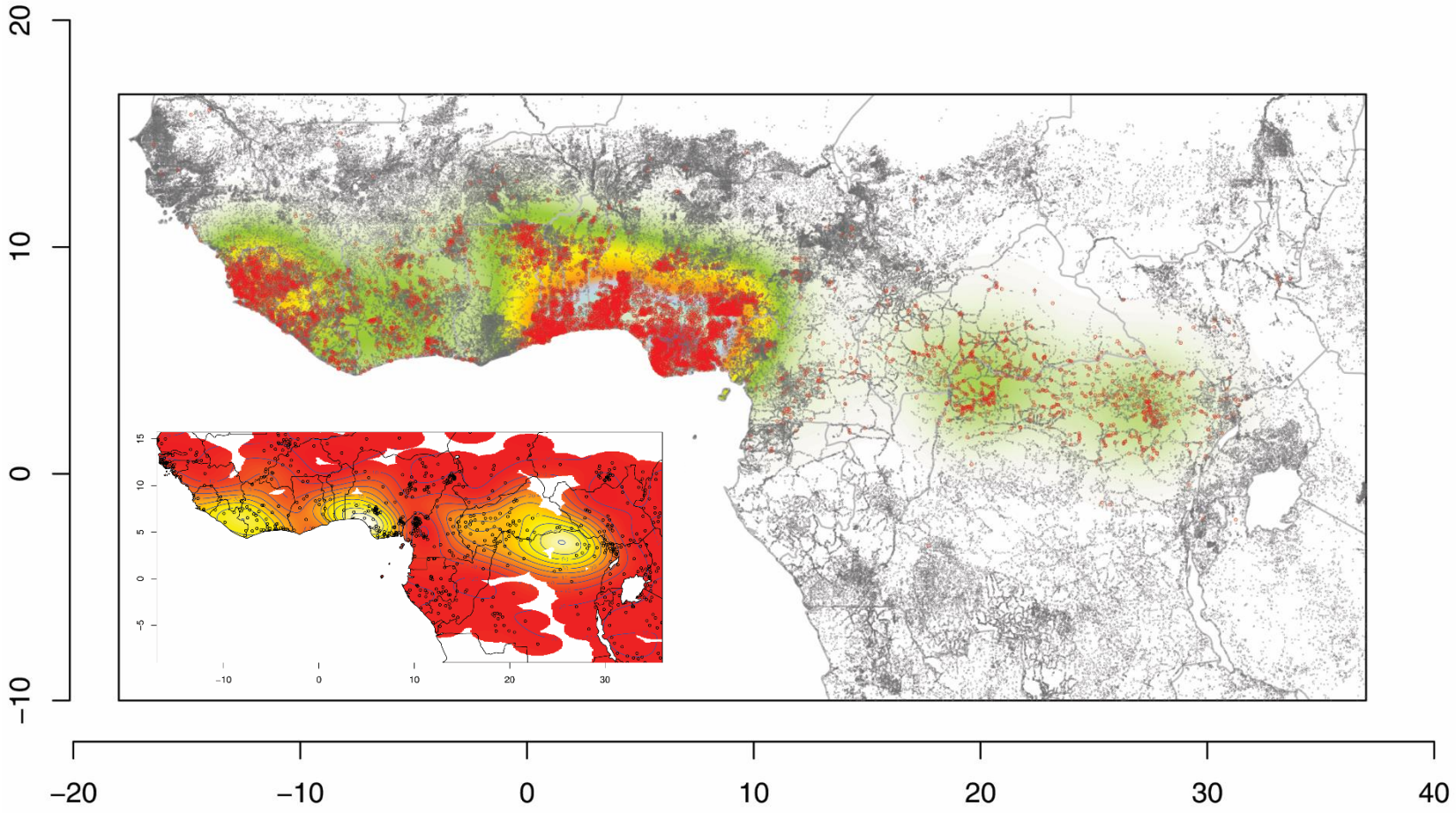
FIGURE 9 from Idiatov & Van de Velde (2021): The heat map color scheme contour plot of the GAM regression surface of the log-transformed (after scaling up by 0.83) F_{LV} frequencies (including the languages without LV stops) as a function of the combination of longitude and latitude using thin-plate regression splines. The model summary: k = 18 (k-index = 1, p-value = 0.53, k′ = 323), family = Gaussian, edf = 108.1, deviance explained = 85.80%, AIC = 1764, intercept log-transformed (after scaling up by 0.83) F_{LV} = 1.54837, $p < .001$.

STATISTIC VISUALIZATION: GAM





- **Cross-validation** with other types of data



LV STOPS: MAIN FINDINGDS

- Languages with higher lexical frequencies of LV stops are grouped into three areal hotbeds
- Languages with LV vary significantly with respect to the status of LV in their phonologies and lexicons
- In many of the languages with LV stops, they have a much lower lexical frequency than average consonant phonemes
- LV stops have a skewed lexical distribution, both phonotactically (stem-initial position) and semantically (expressive vocabulary)

LV STOPS: HISTORICAL INTERPRETATION

- LV stops are a substrate feature and the three hotbeds are areas of retention and refuge zones.
- LV stops are retentions from an areal point of view, but innovations from a genealogical point of view in the great majority of African languages that have them today.
- Detailed hypotheses regarding prehistoric migration patterns of Niger-Congo speaking populations
- Adjusted and refined the scenarios for the Bantu expansion.
- C-emphasis prosody as the primary force driving the emergence, spread, and intra-linguistic distribution of LV stops

SUBSTRATE EFFECTS IN MORPHOSYNTAX: N/V RATIOS

- The same methodology can be applied to morphosyntactic patterns
- N/V ratios in Sub-Saharan languages show striking, areally conditioned differences that reflect substrate effects

WHY N/V RATIOS?

- Like with LV stops, our research question and research hypothesis were informed by our knowledge of many language groups of (N)SSA, especially Mande, "Atlantic", Bantoid
- Examples of languages with few verbs (high N/V ratios):
  - Southern Mande (Tura, Dan ≈ 180-190 underived verbs out of > 3000 lexical entries)
  - ? Bandaic
- Examples of languages with many verbs (low N/V ratios):
  - Bantoid (BLR3 on Proto-Bantu roots: 711 V / 624 N)
  - Northern Atlantic (cf. Christiane Seydou on Fula: hardly any nominal roots)

AN ASIDE ON OMNIPREDICATIVITY

- Very many verbs ≠ "omnipredicativity" (Amerindian or Polynesian-style)
  - N and V are clearly distinguished in morphosyntax
  - Very many N are clearly derived from V
  - True, even for languages where synchronically there seem to be a lot of N/V isomorphism, which (at least, historically) is rather V > N conversion (cf. Idiatov 2018 on Western Mande).

THE NULL HYPOTHESIS

- Minimally: ratios of N/V should be largely constant across related languages
- Maximally: ratios of N/V should be largely constant across the SSA

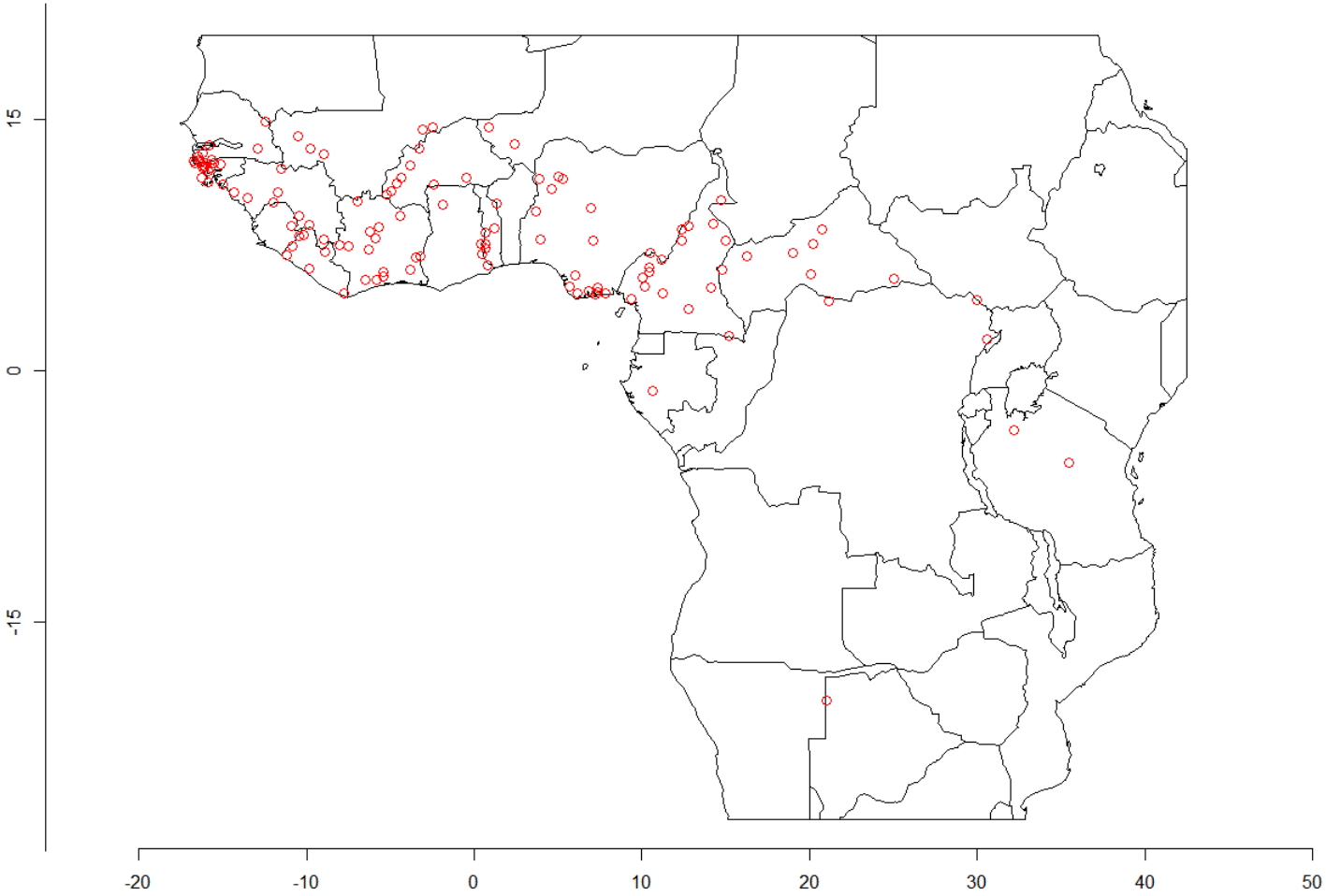
METHODOLOGY

- Largely limited to the data in RefLex (www.reflex.cnrs.fr)
- RefLex has 2074 sources for 1095 languages, but the source are of very uneven quality
- The filtering of sources is ongoing → currently at ≈ 260 sources
- Raw data contain a lot of noise (derivations, compounds, borrowings…) that muddles the signal in the data
- Approximate monomorphemic core → "1h2l":

  C, CV, CVV, CC, CCV, CVC, CVCV, V, VC, VV, VVC, VCV, VCVC

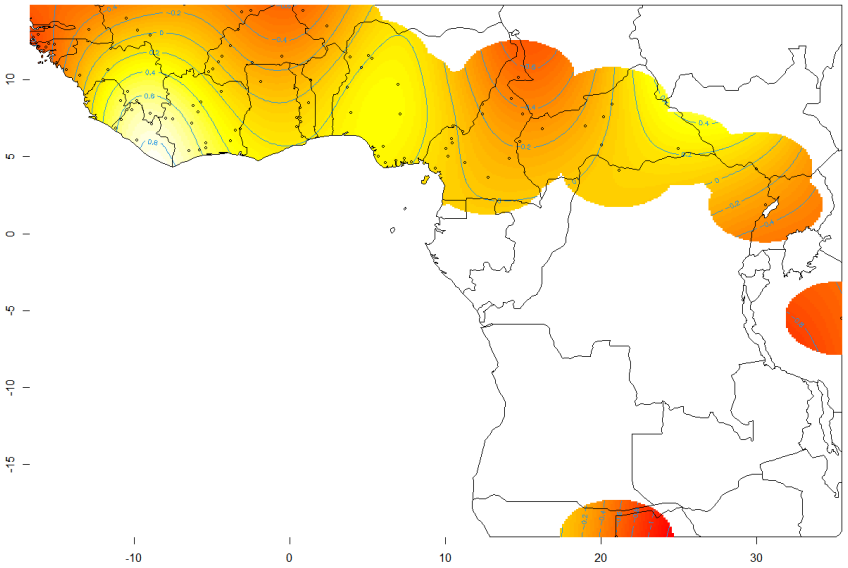
DATA PROCESSING: 1H2L CLEANING

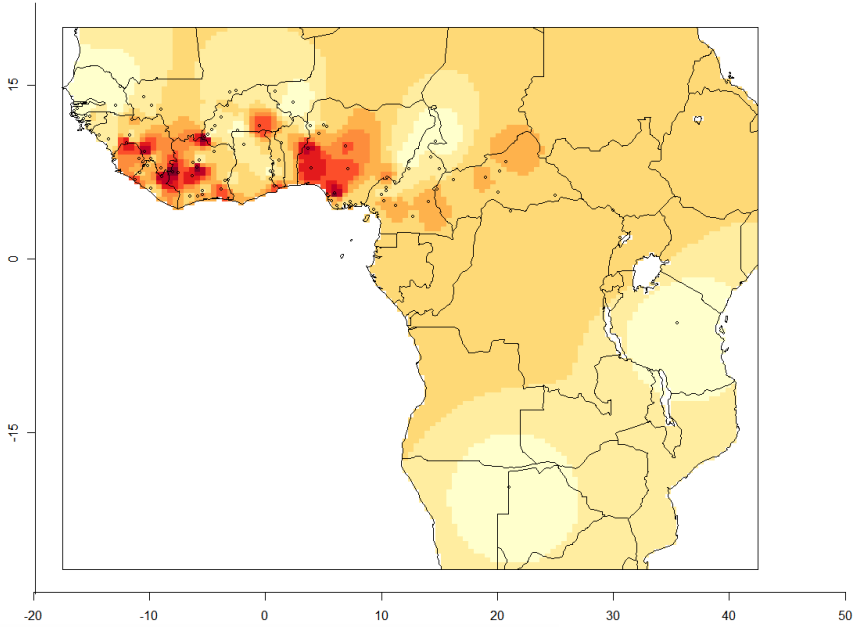
- So far, we have 1h2l cleaned 123 sources

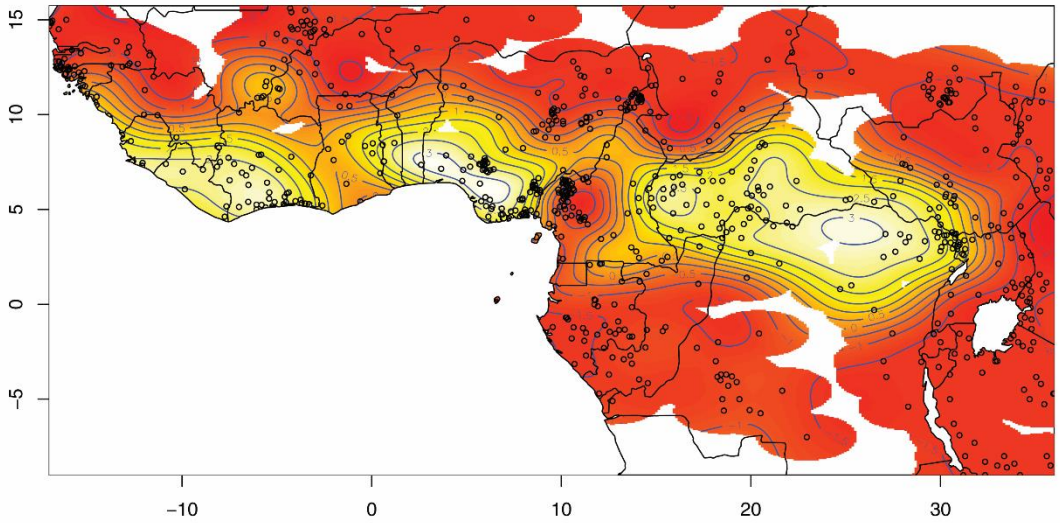


Langage, Langues et Cultures d'Afrique

N/V RATIOS PRELIMINARY RESULTS: 1H2L vs LV HOTBEDS







N/V RATIOS PRELIMINARY RESULTS

Preliminary results with respect to N/V ratios in (N)SSA:

- Languages with few verbs (high N/V ratios) are concentrated in two areal hotbeds
- These two hotbeds largely **coincide with** the Lower and Upper Guinea hotbeds of high lexical frequency of **LV stops**
- The Ubangi Basin hotbed, in contrast, does not clearly correspond to an area with a high N/V ratio