Langage, Langues et Cultures d'Afrique

Llacan

AREAL PATTERNS OF NOUN/VERB RATIOS IN SUB-SAHARAN AFRICA

Dmitry Idiatov, Guillaume Segerer & Mark Van de Velde

LLACAN (CNRS – INALCO)

dmitry.idiatov@cnrs.fr

guillaume.segerer@cnrs.fr

mark.vandevelde@cnrs.fr

AREAL TYPOLOGY

- 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).

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
- … but garbage in, garbage out
- let the data speak for themselves (☹ binning)
- non-binary
- non-/ŋ͡mằmò/*: Start from a clear research hypothesis, define the null hypothesis and be aware of the possible bias that a given decision may induce on the result

*/ŋ͡mằmò/ 'commit oneself to something subsequently found embarrassing' (Grebo; Innes 1967)

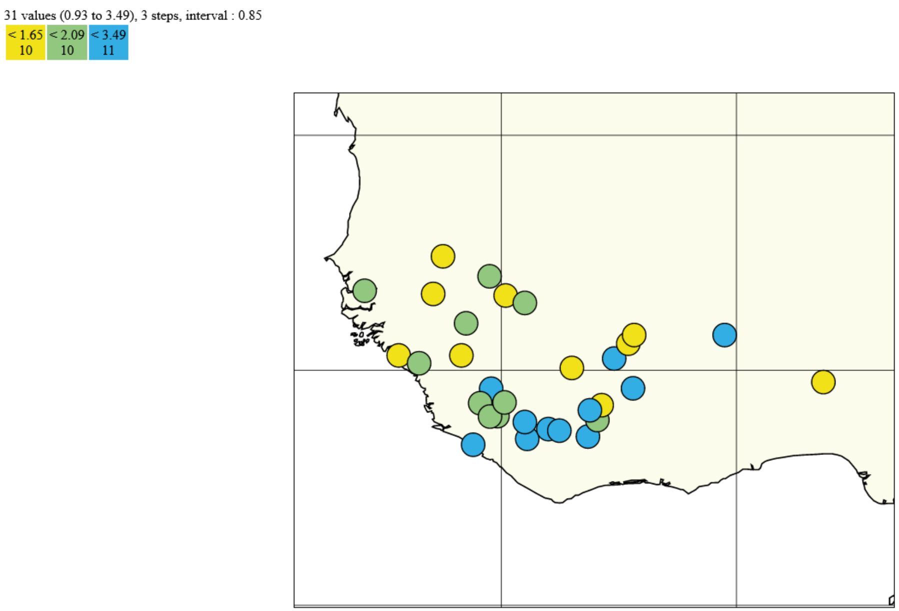
RPREFERRED APPROACH: PRACTICALLY

- 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**

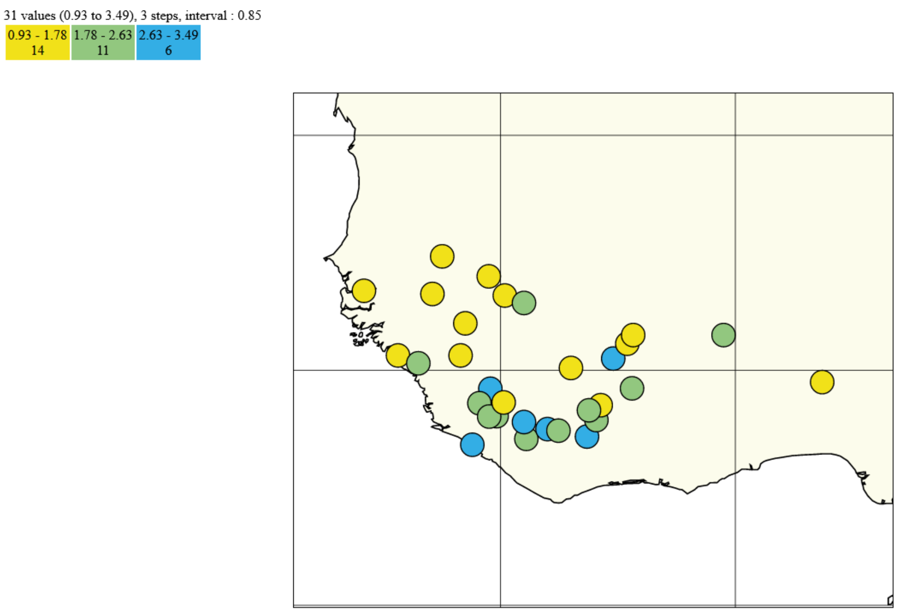
VISUALIZATION: POINTS & BINS

- **Points of different colors** in space as a first approximation

Language Distribution

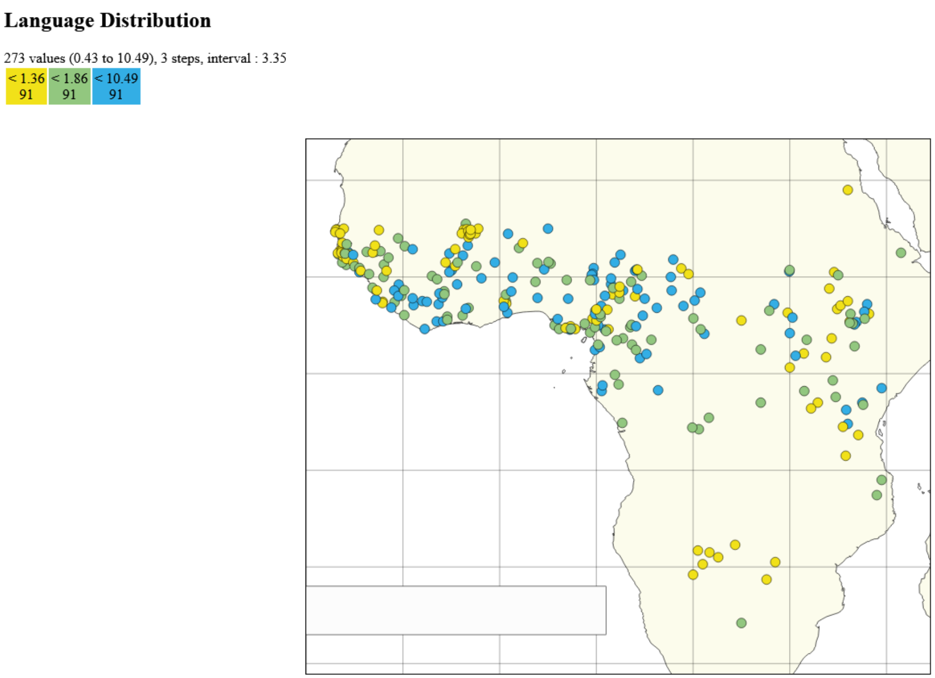


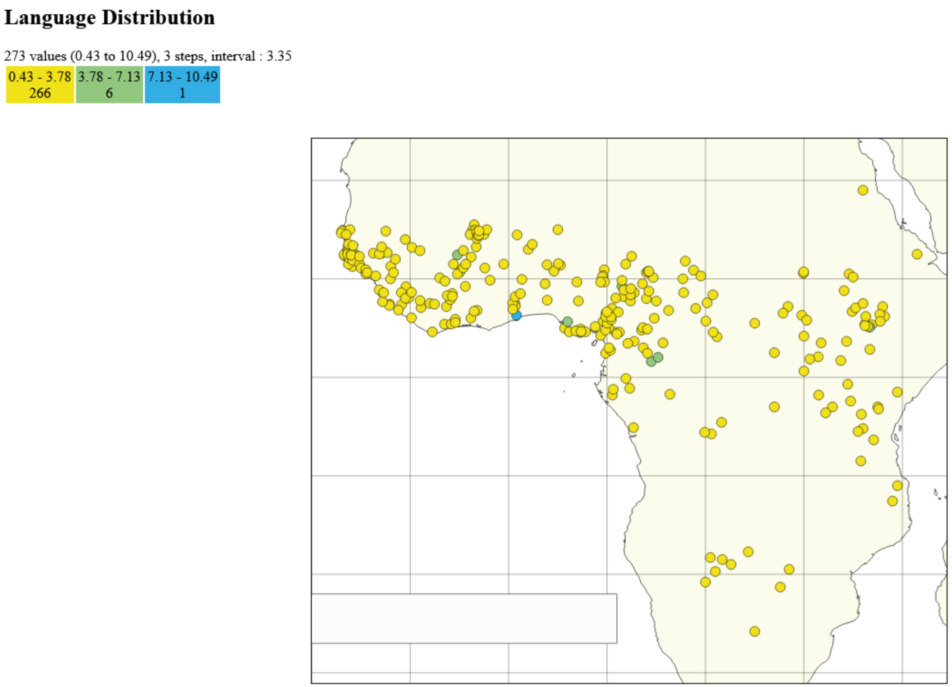
Language Distribution



VISUALIZATION: POINTS & BINS

- **Points of different colors** in space as a first approximation



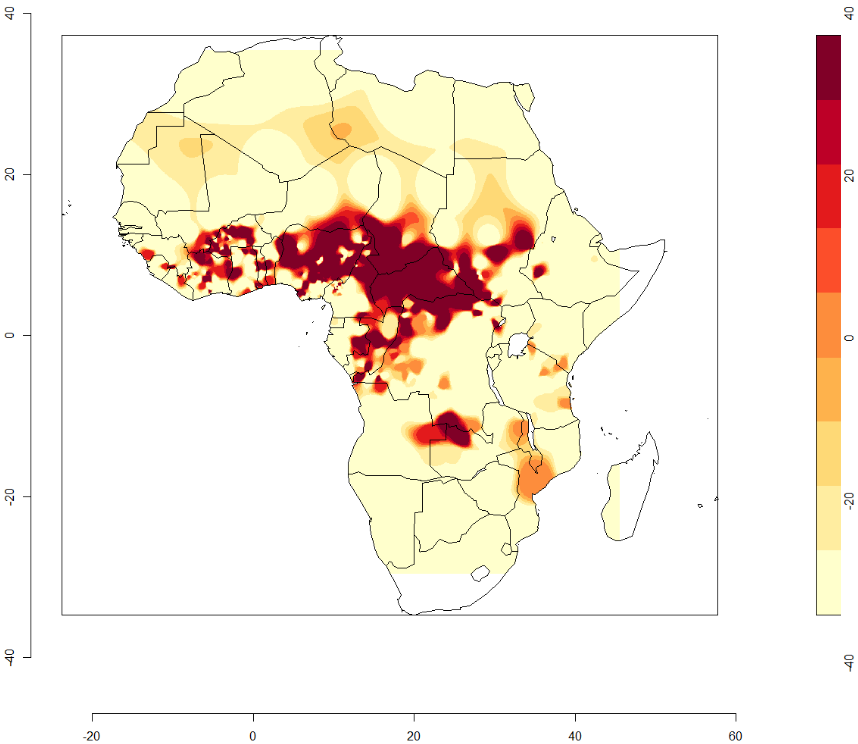


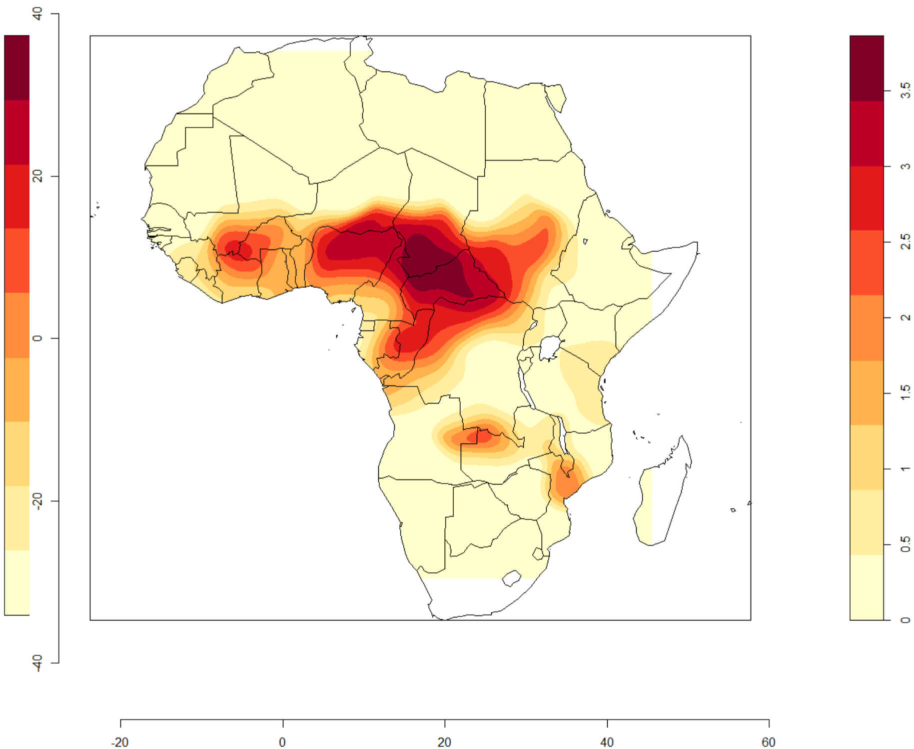
DETERMINISTIC VISUALIZATION: INTERPOLATION

- Spatial interpolation: a (deterministic) tool for visualizing the distribution of a variable in space by estimating the value of a variable at any specific location based on a weighted average of the known values at sampled locations
  - **IDW** (inverse distance weighting): exact, finer structure
  - **Kernel smoothing** : inexact, general trends

DETERMINISTIC VISUALIZATION: INTERPOLATION

- the choice of **bandwidth**
- **visualization artefacts**
  - Idiatov (2018:140-141) on the areal typology of CFNM





STATISTIC VISUALIZATION: GAM

- **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

STATISTIC VISUALIZATION: GAM

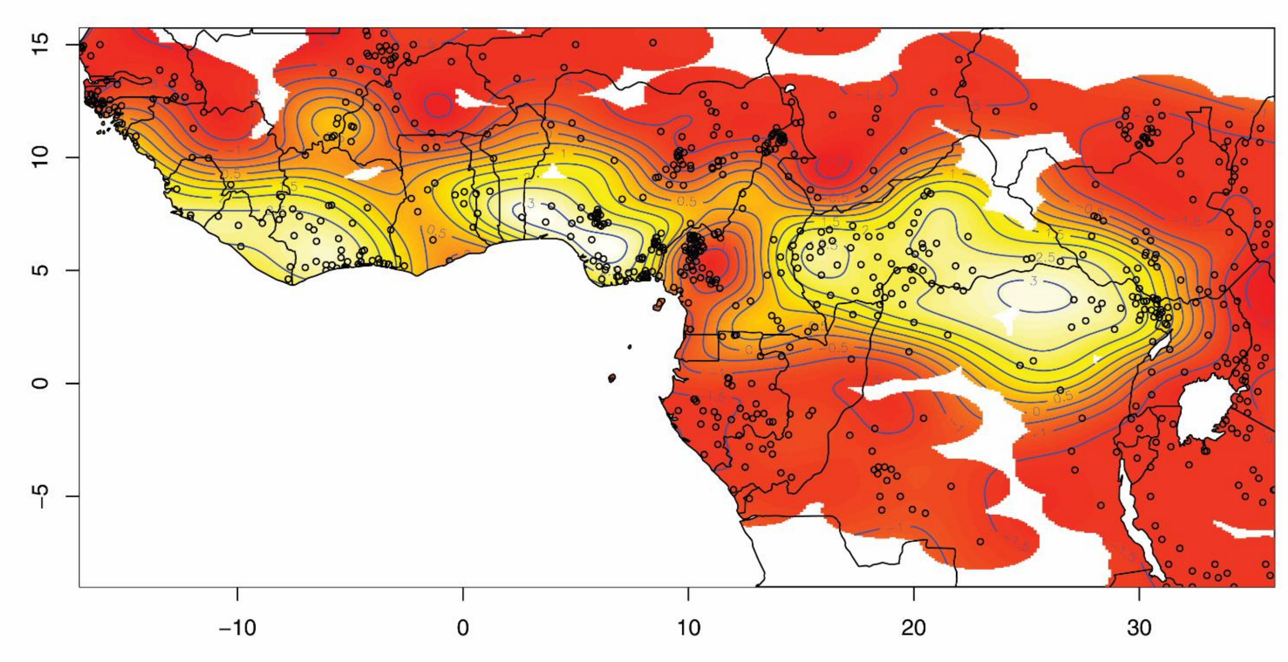
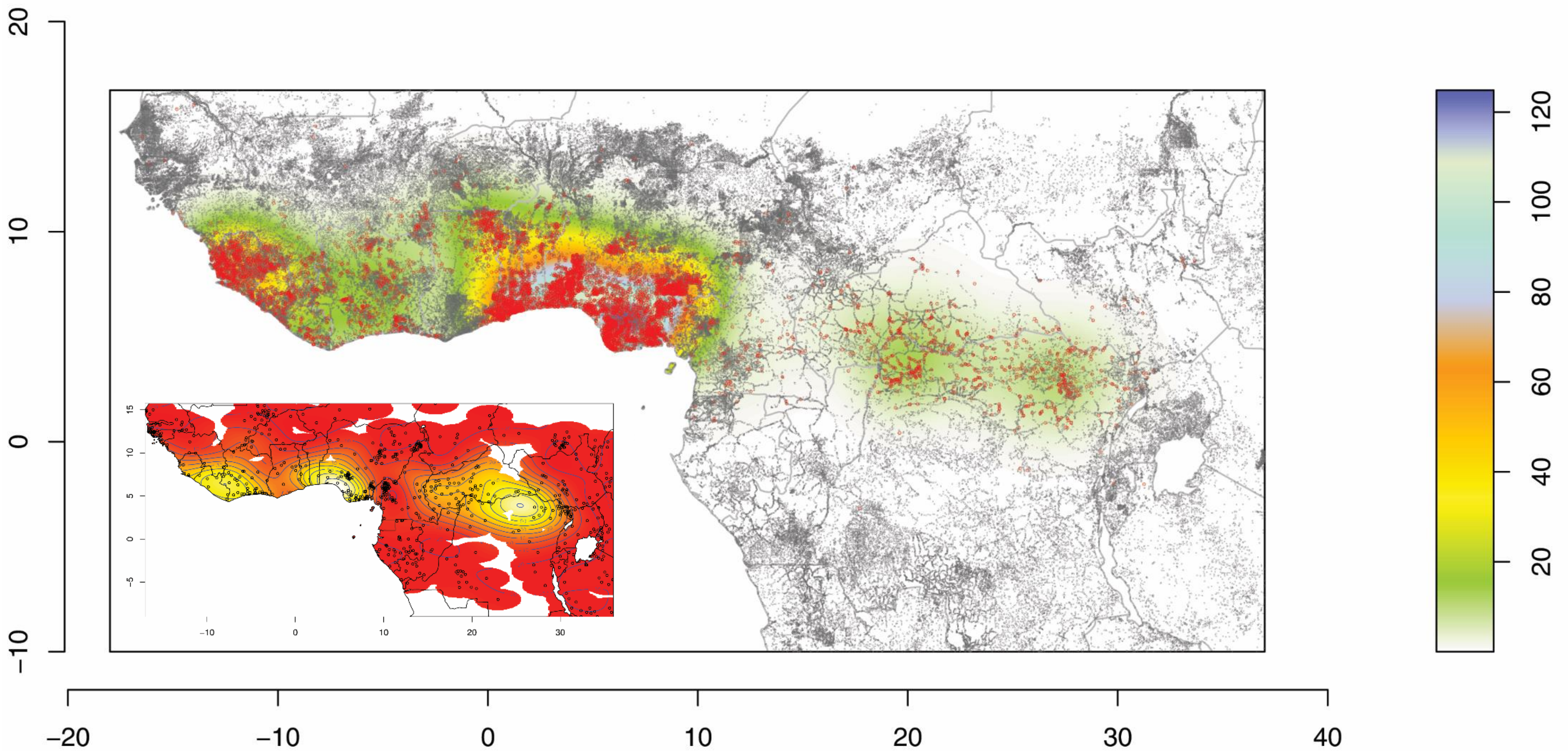


FIGURE 9 from Idiatov & Van de Velde (2021): The heat map color scheme contour plot of the GAM regression surface of the log-transformed (after scaling up by 0.83) F_{LV} frequencies (including the languages without LV stops) as a function of the combination of longitude and latitude using thin-plate 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, $p < .001$.

- Cross-validation with other types of data



LV STOPS (IDIATOV & VAN DE VELDE 2021)

The **main findings** of Idiatov & Van de Velde (2021) with respect to LV stops in NSSA:

- 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
- Languages with higher lexical frequencies of LV stops are grouped into three areal hotbeds
- LV stops have a skewed lexical distribution, both phonotactically (stem-initial position) and semantically (expressive vocabulary)

A **historical interpretation** of the findings with respect to LV stops in NSSA:

- LV stops are a substrate feature and the three hotbeds are areas of retention and refuge zones.
- 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

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

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

DATA PROCESSING: SELECTING THE SOURCES

- For the moment, limited to the data in RefLex
- On 03.11.2021, RefLex has 2074 sources for 1095 languages, but the source are of very uneven quality
- Selecting the sources – the first pass → 316 sources:
  - Removed comparative wordlists (TLS, BCCW, ALGAB, Koelle), grammars, articles, theses
  - Removed sources before 1900
  - Removed very small sources < 400 entries (cf. Dockum & Bowern 2019 on the 400-item threshold to be able to correctly represent the phonology of a language)

DATA PROCESSING: SELECTING THE SOURCES

- The filtering of sources is ongoing → currently at 272 ~ 261 sources
  - Removed sources on tone languages with no tones marked
  - Removed (smaller) sources that are most likely to be based on wordlist elicitation
  - In case of several sources for the same language or several closely related varieties spoken near each other, we kept the most reliable source(s), which tend to also be the biggest
  - Comparative testing: 6 Joola sources, 4 Manjaku sources…

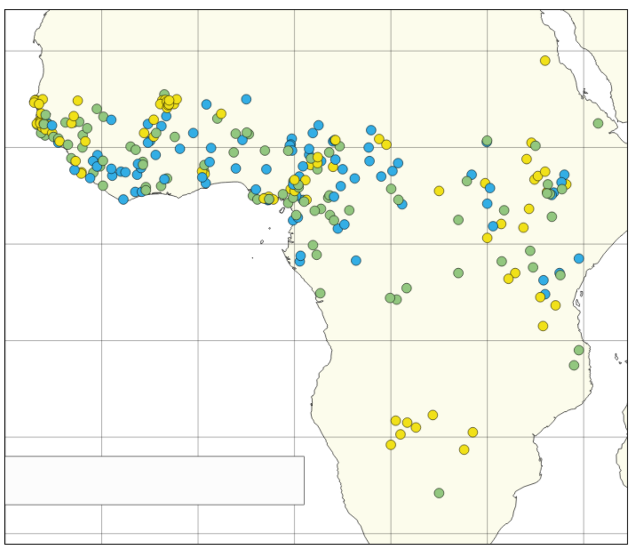
DATA PROCESSING: COUNTING WHAT?

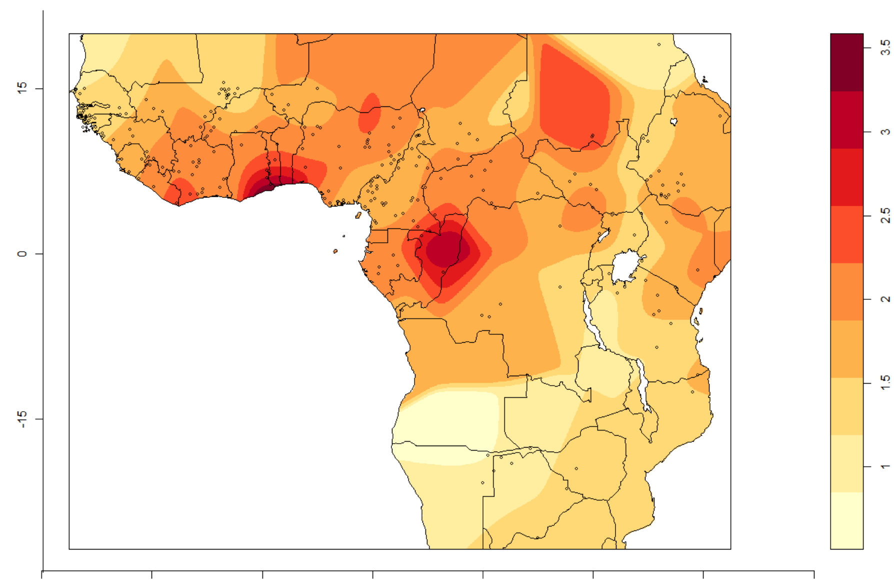
- Option #1 "Raw data": The raw numbers of entries categorized as nouns and verbs in a given source in RefLex
  - Easy to implement
  - But the signal in such data is very weak and muddled

Language Distribution

273 values (0.43 to 10.49), 3 steps, interval : 3.35

< 1.36	< 1.86	< 10.49
91	91	91

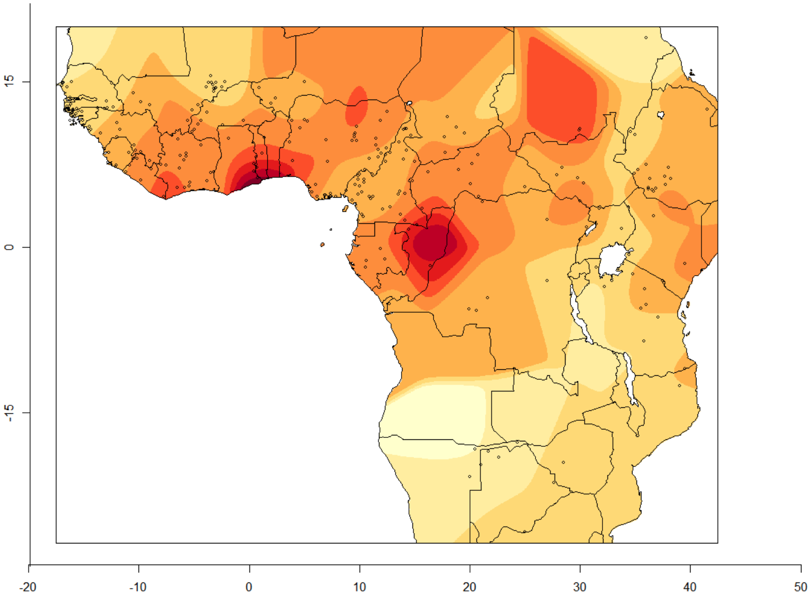




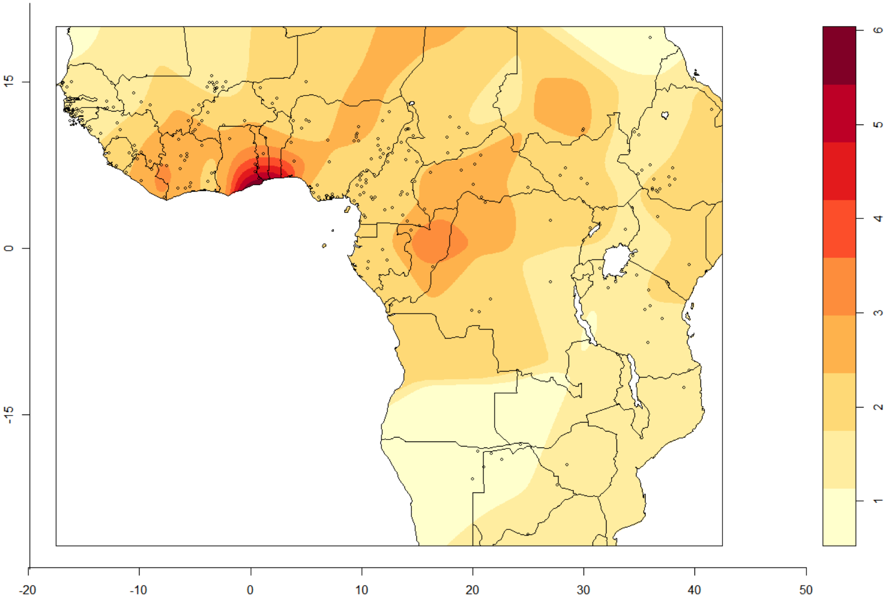
DATA PROCESSING: COUNTING WHAT?

- Option #2 "Unique entries": Count only unique lexical entries categorized as nouns and verbs in a given source in RefLex
- The way RefLex deals with the polysemy and homonymy in the sources:
  - Each meaning of an original multi-sense entry is converted into a separate entry
  - Original homonyms = entries are kept as such
- The same issues as with Option #1:
  - Easy to implement
  - But the signal in such data is very weak and muddled

- Option #2 “Unique entries”: Count only unique lexical entries categorized as nouns and verbs in a given source in RefLex



#1 Raw data: kernel smoothing



#2 Unique entries: kernel smoothing

THE CULPRITS WEAKENING THE SIGNAL

- The main **culprits** muddling the signal in the data:
  - derived forms (primarily: V > V, V > N ; to a lesser extent: N > V, Other > N, Other > V)
  - compounds
  - borrowings
- How do we remove these culprits? (by preference, in a semi-automatic way)
  - Relatively easy when this information is provided by the source and the corresponding fields were filled in RefLex (EML, RAC…)
  - Unfortunately, this the case for only a very small number of sources in our sample

APPROXIMATING THE MONOMORPHEMIC CORE

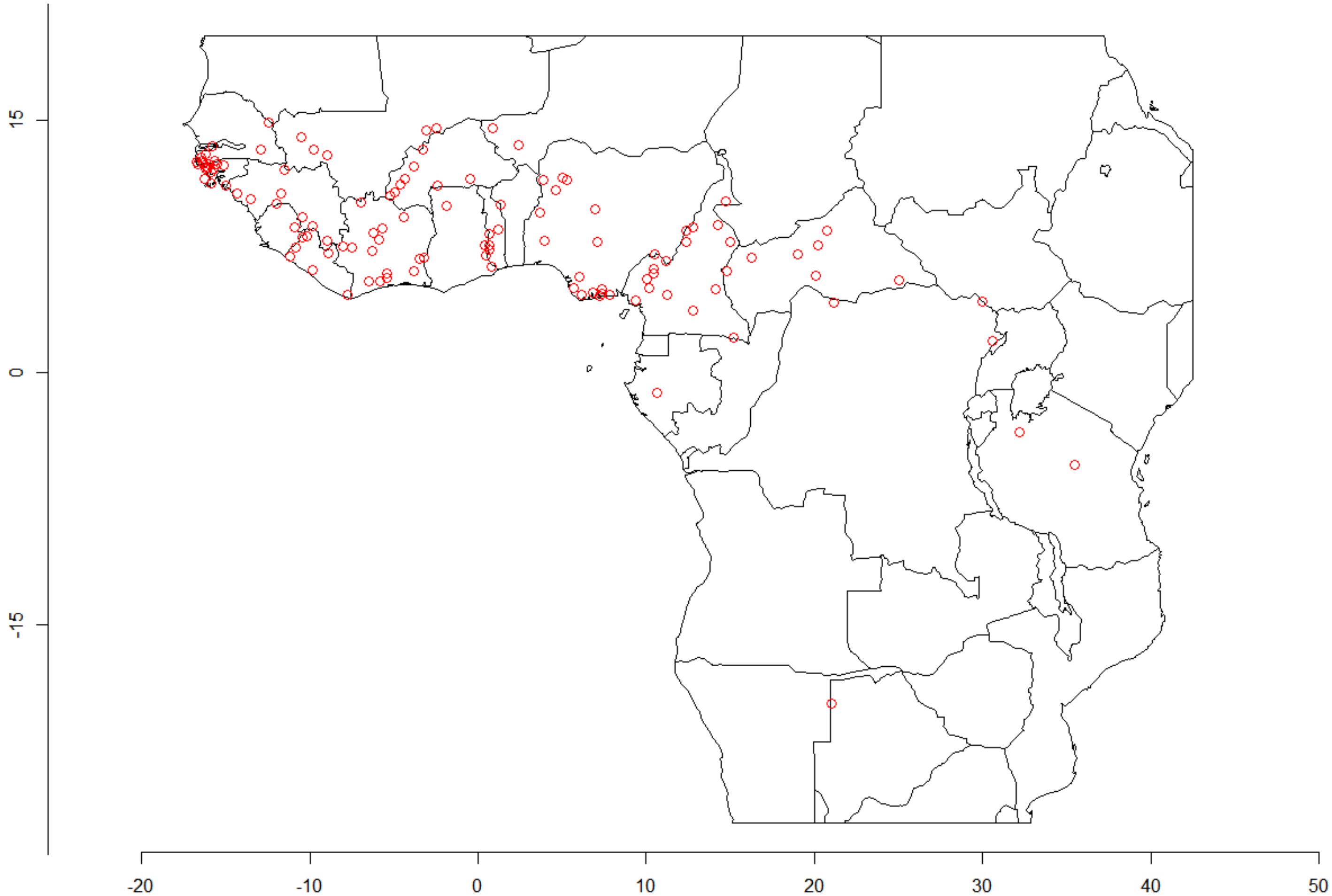
- Option #2 "Approximate 1-morphemic core": Approximate the monomorphemic (non-derived) lexicon of a language by keeping only certain shorter forms, whose range is limited to a predefined set to enhance comparability.
  - Typically for the (N)SSA, N&V monomorphemic roots are 1-2syl
  - Most available reconstructions for the languages of (N)SSA suggest 1-2syl roots for N&V
  - However, it is not always clear what to count as a syllable
  - We know that in many languages, certain types of 1-2syl tend to come from 2-3syl forms

APPROXIMATING THE MONOMORPHEMIC CORE: 1H2L

- Option #2 "1h2l": a (somewhat arbitrary) fixed list of syllabic shapes of maximally 1 heavy or 2 light syllables
  - Follow the conventions of RefLex with respect to long vowels (VV) and homorganic N-stop clusters (one C).
  - no C-clusters, except initial CCV
  - no super-heavy syllables, such as CVVC
  - Only: C, CV, CVV, CC, CCV, CVC, CVCV, V, VC, VV, VVC, VCV, VCVC
  - ✋ For languages with frozen or active class affixes, these shapes refer to **stems** (☞ lots of manual cleaning)
  - Manually remove the remaining borrowings, derivates and compounds

DATA PROCESSING: 1H2L CLEANING

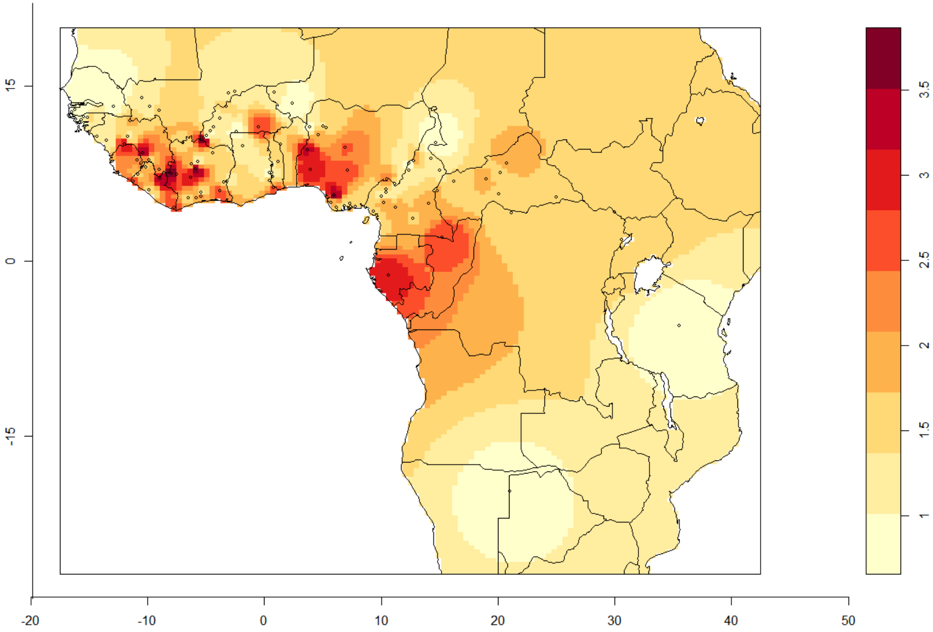
- So far, we have 1h2l cleaned 123 sources

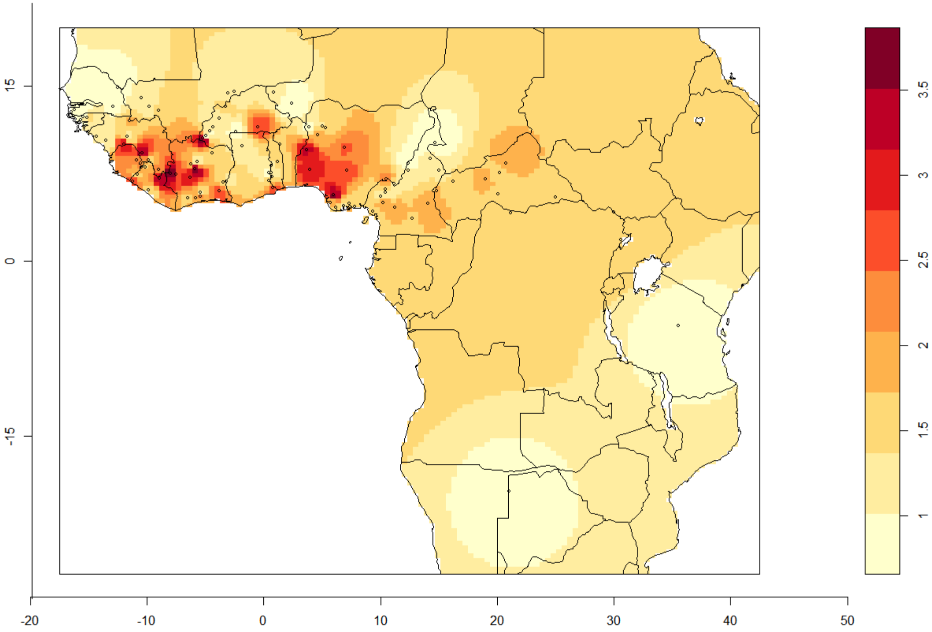


- **Additional advantages**:
  - Normalized some outliers (e.g., Rongier 2003 : ewe)
  - Generally, normalized differences in source size between related languages for medium-sized and big sources
  - For smaller sources (≈ < 1000 entries) and some less reliable bigger sources, the effect is less pronounced
  - Revealed a **clear and coherent signal** in the data…

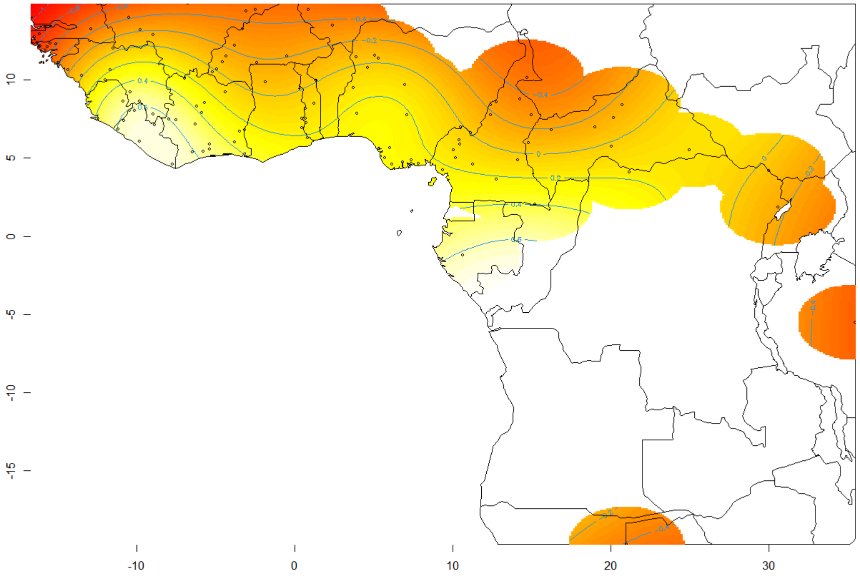
- "Encyclopedic" sources :
  - Such as: Van der Veen & Bodinga 2000 : Gevia ; Brisson & Boursier 1979 : baka ; Dieu & Perrois 2016 : koma ; Innes 1967 : Grebo ; Dumestre 2011 : bambara
  - Create new outliers by inflating (as compared to the other sources) the number of 1h2l nouns with flora and fauna names, specialist technical and cultural vocabulary, neologisms, slang, etc.
  - When they are few other data points in their vicinity, some of these outliers cause problems for visualizations and are best removed for the time being (esp., Gevia and Baka)

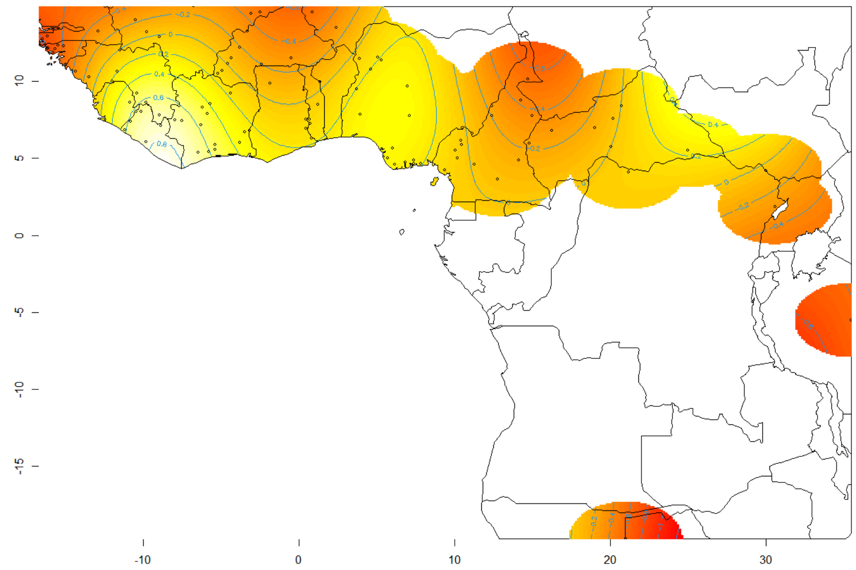
1H2L: SPATIAL INTERPOLATION





1H2L: GAM





1H2L vs LV HOTBEDS