IMPLICATIONS OF THE LEXICAL FREQUENCY OF LABIAL-VELAR STOPS IN NORTHERN SUB-SAHARAN AFRICA FOR NIGER-CONGO RECONSTRUCTION

Dmitry Idiatov & Mark Van de Velde

LLACAN, CNRS, Sorbonne-Paris Cité, INALCO
dmitry.idiatov@cnrs.fr
mark.vandevelde@cnrs.fr

INTRODUCTION

- Northern sub-Saharan Africa is obviously a spread zone with a marked areal distribution of various linguistic features
  - Macro-Sudan belt
  - Sudanic zone
  - …
- LV are common in NSSA languages
- Typologically, LV are known to be rather rare

INTRODUCTION

- Proto-languages with LV

  *Gbaya (Moñino 1995), *Central Sudanic (Boyeldieu 2008), *Southern Mande (Vydrine 2005), *Guang (Snider 1990), *Upper Cross (Dimmendaal 1978), *Lower Cross (Connell 1995), *Igboid (Blench 2016 ms.)…

- Scepticism on the relevance of LV for reconstruction

  “Although labial-velar stops are widespread in Niger-Congo, their historical status is still problematic.”

  (Dimmendaal 2001:377)

  “[t]he presence or absence of labial-velars will not be very useful for the purpose of reconstructing remote proto-languages”

  (Hyman 2011:16)

INTRODUCTION

Q_1: What can the areality of LV tell us about the history of the languages of NSSA?

Q_2: What is the possible historical depth of LV in the languages of NSSA (in particular, in Niger-Congo languages)?

INTRODUCTION

Given that:

- Languages with LV can vary significantly with respect to the status of LV in their phonologies and lexicons

Subquestions:

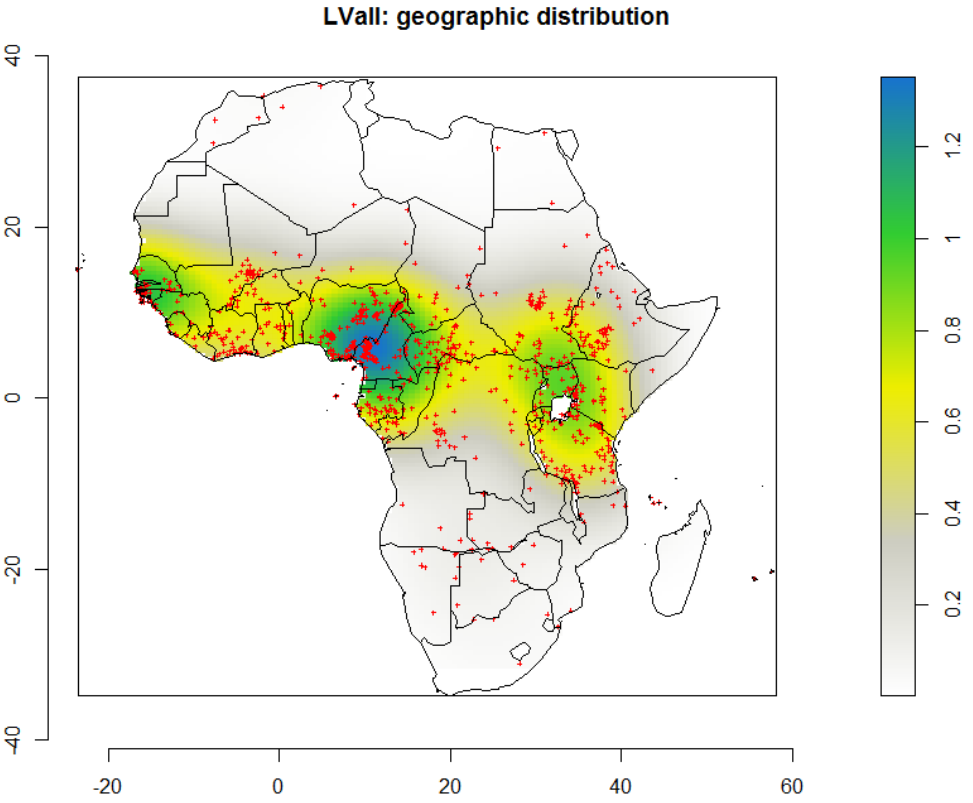
- Are LV "normal" phonemes in NSSA languages?
- Are there differences between languages in the frequencies of LV in their lexicons?
- Are there geographic patterns in the LV frequency distribution?
- Are the distributions of LV within the lexicons random?
- How can we explain the observed patterns?
- What are their implications for the reconstruction of the languages of NSSA, and in particular of Niger-Congo languages?

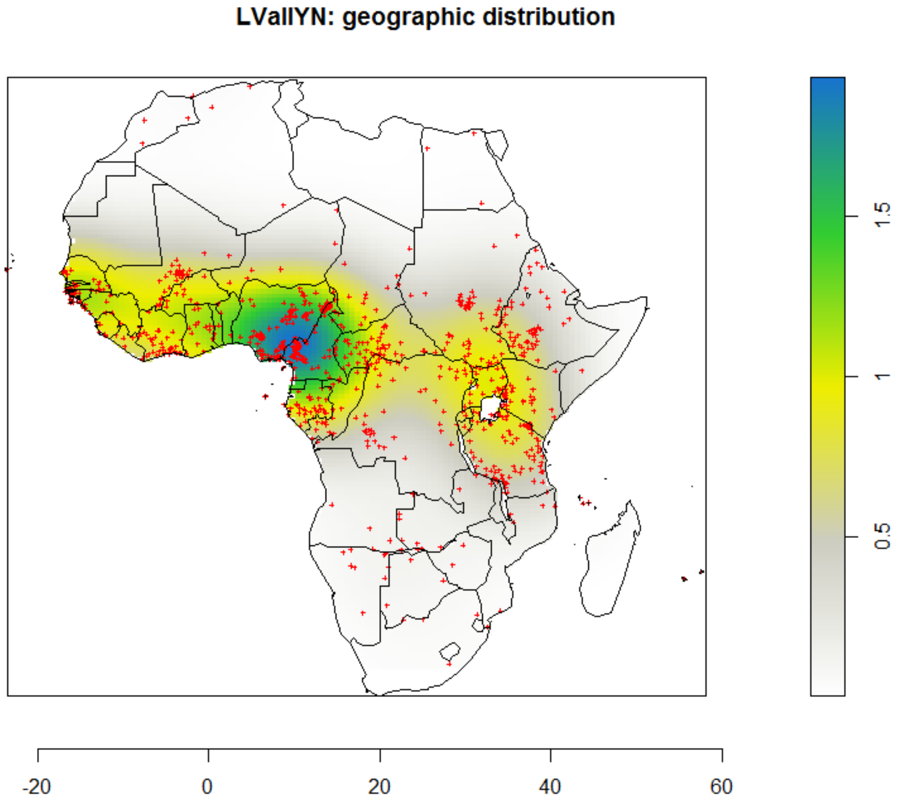
DATA

LV data sources:

- **RefLex**, www.reflex.cnrs.fr, LVFreq data
- Phoible, www.phoible.org, YN data
- Additional LVFreq data for some Mande and Bantu languages

DATA





LVall

1074 languages with frequency data:

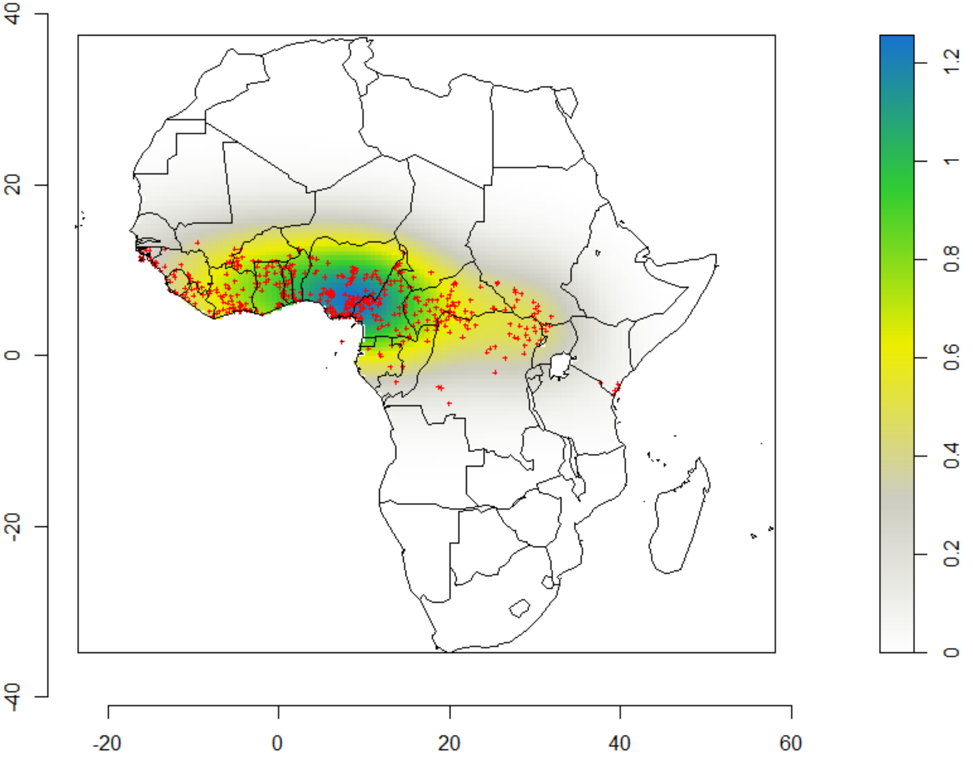
- LV & their frequency is known (336 lgs)
- No LV

LVallYN

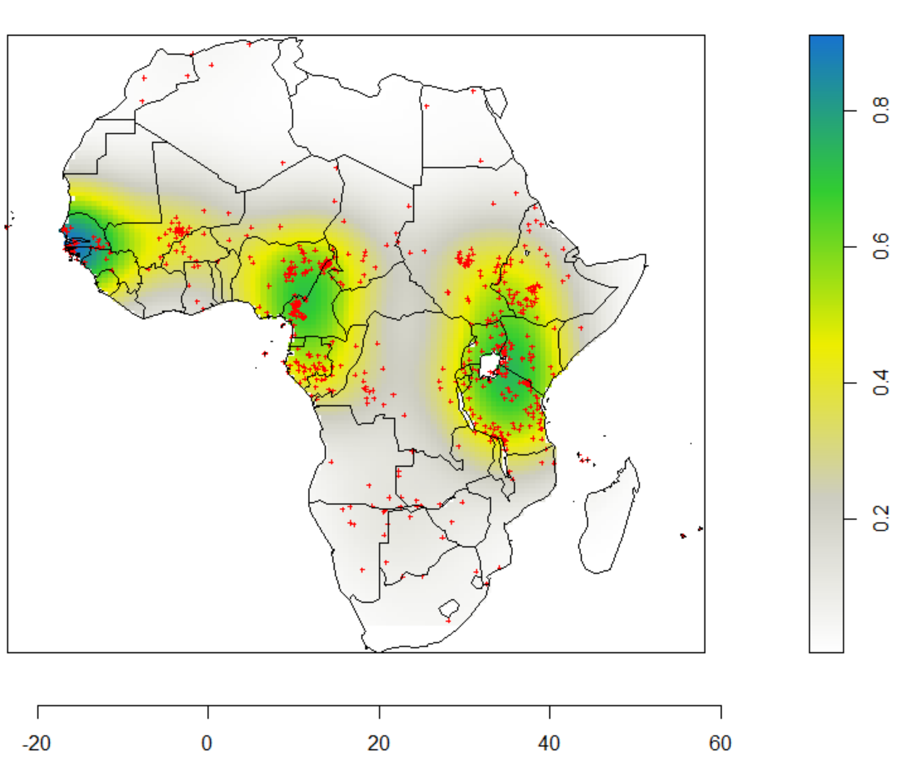
1304 languages:

- LV & their frequency is known (336 lgs)
- LV, but no frequency data (230 lgs)
- No LV

DATA



LVall_Y languages: geographic distribution



LVall_N languages: geographic distribution

LVFreq estimation

H$_0$: In a lexicon, all C phonemes have equal frequency (have equal probability of occurrence)

$$\boldsymbol{LVFreq} = \frac{LV_O}{LV_E * W_{LV}} * 100\% = \frac{\sum T_{LV}}{\frac{\sum T_C}{\sum P_C} * \sum P_{LV}} * 100\%$$

LV_O - observed LV count

LV_E - expected LV count

W_{LV} - LV weighting coefficient

T_{LV} - LV token

T_C - any C token

P_{LV} - LV phoneme

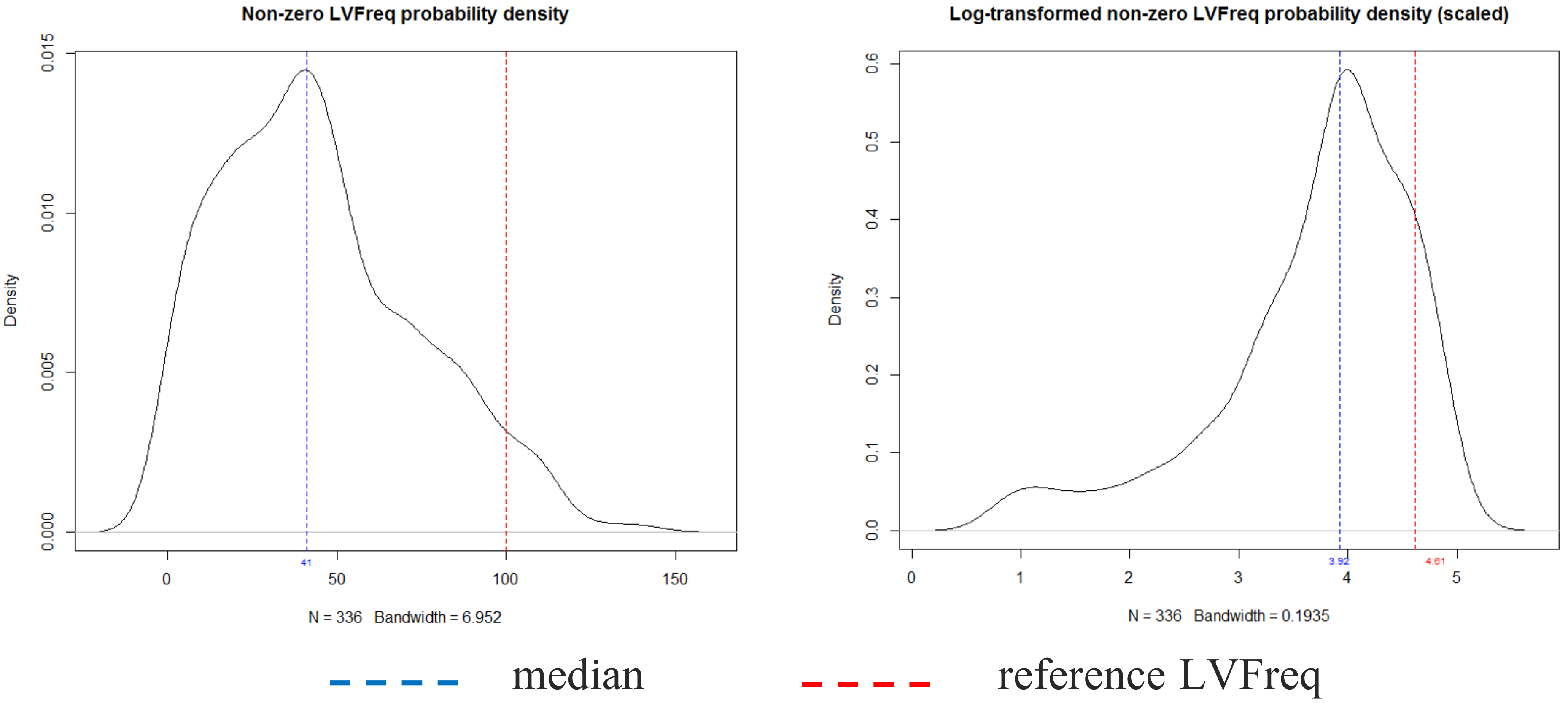
P_C - any C phoneme

LVFreq estimation

LVFreq = 0% no LV

LVFreq = 100% **“reference LVFreq”** - LV are “normal” phonemes, i.e. the observed number of occurrences of LV is the same as would be expected given the H_0

LV FREQUENCY ESTIMATION



- Log-transformation does not help to make the data more normal
- LV are relatively **rare phonemes** in most languages that have them, which is in accordance with their typological rarity

DISTRIBUTION WITHIN THE LEXICON

Are the distributions of LV within the lexicons random?

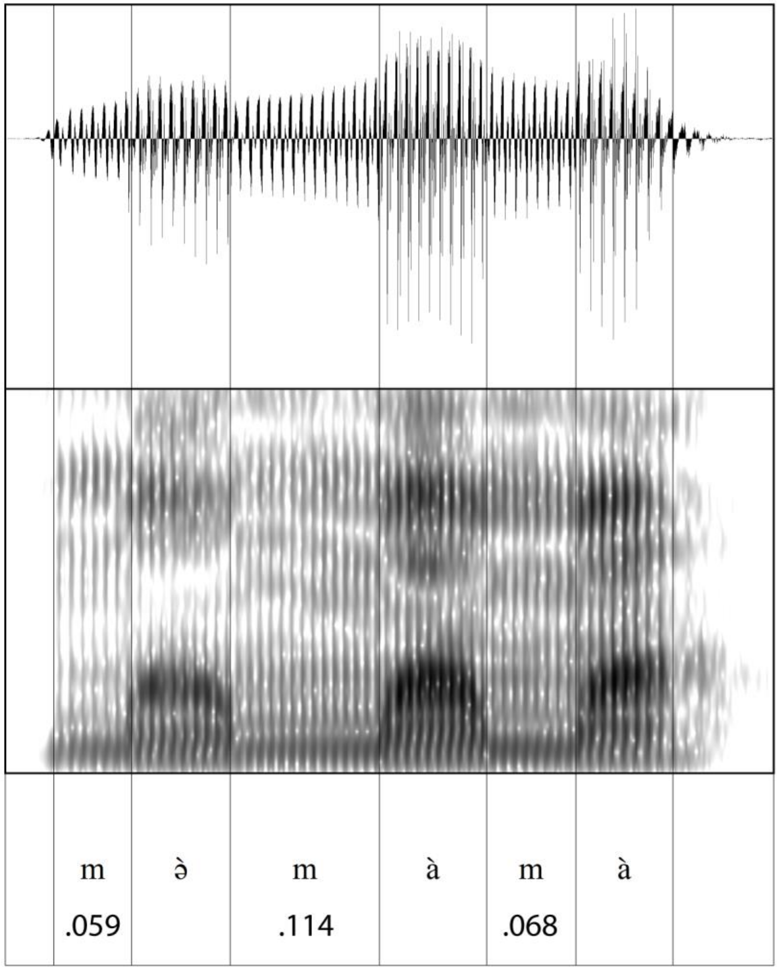
- LV tend to be less common in "basic vocabulary"
- {H}: LV are more common in the "expressive" parts of the lexicon, such as ideophones or property words, rather than referring expressions, such as nouns and verbs
- LV are largely restricted to the stem-initial position

LV, SIC-ACCENT & C-EMPHASIS PROSODY

- The correlation [LV ~ “expressive” vocabulary] is not independent of the correlation [LV ~ stem-initial position]

- SIC-accent (as a manifestation of a more general phenomenon of C-emphasis prosody) is a very important factor behind the emergence of LV in NSSA

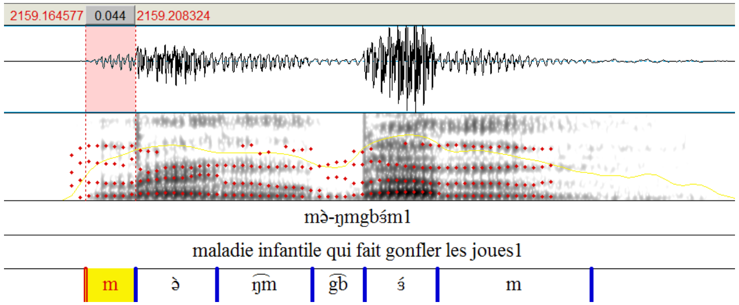
Consonant length in the nonsense word *mə̀-màmà* (Eton, Bantu A70)



LV, SIC-ACCENT & C-EMPHASIS PROSODY

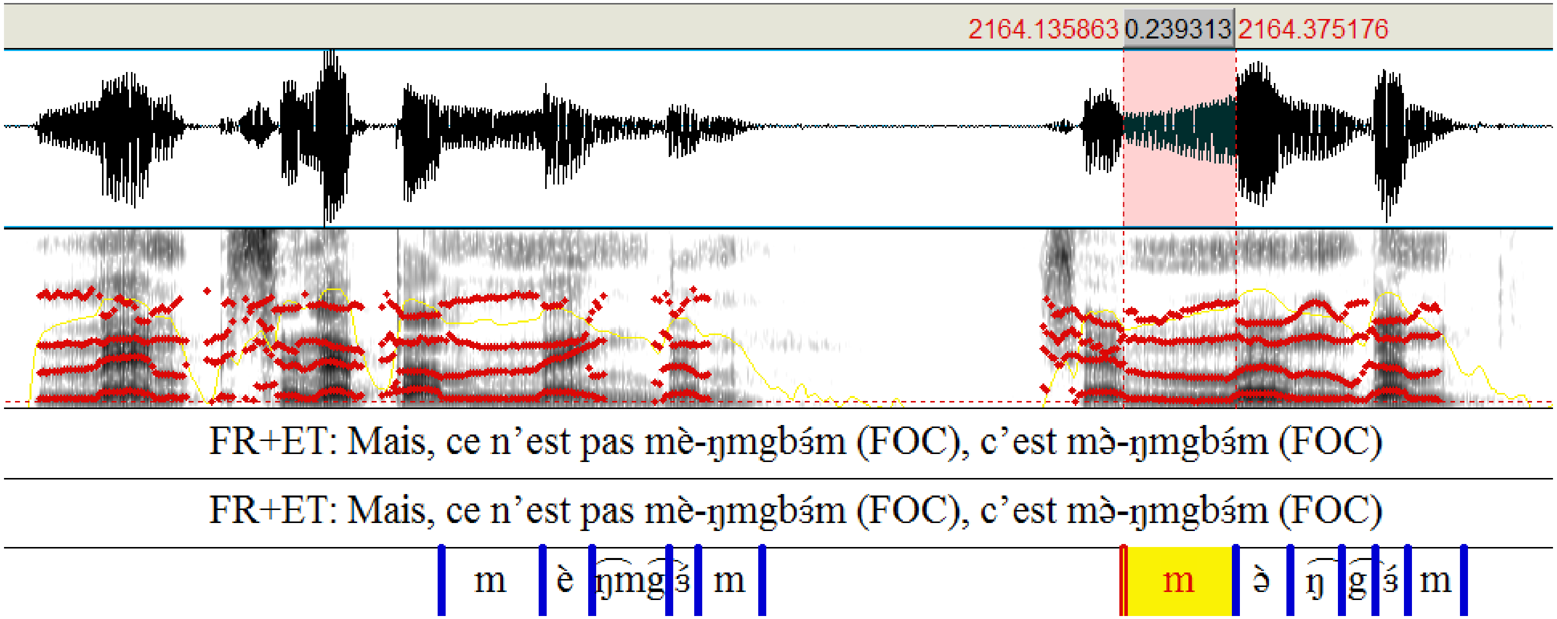
- Corrective focus on the prefix V realized with prefix C-emphasis

Eton (A70)



mə̀-ŋmgbɜ́m1

maladie infantile qui fait gonfler les joues1



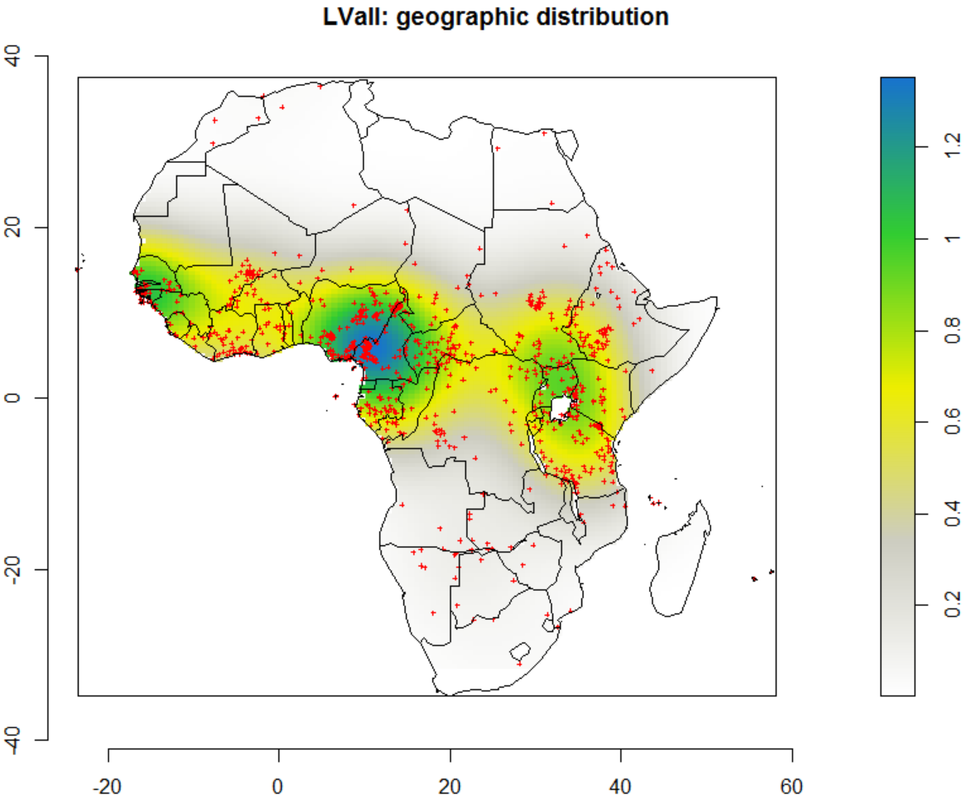
FR+ET: Mais, ce n’est pas mè-ŋmgbɜ́m (FOC), c’est mə̀-ŋmgbɜ́m (FOC)

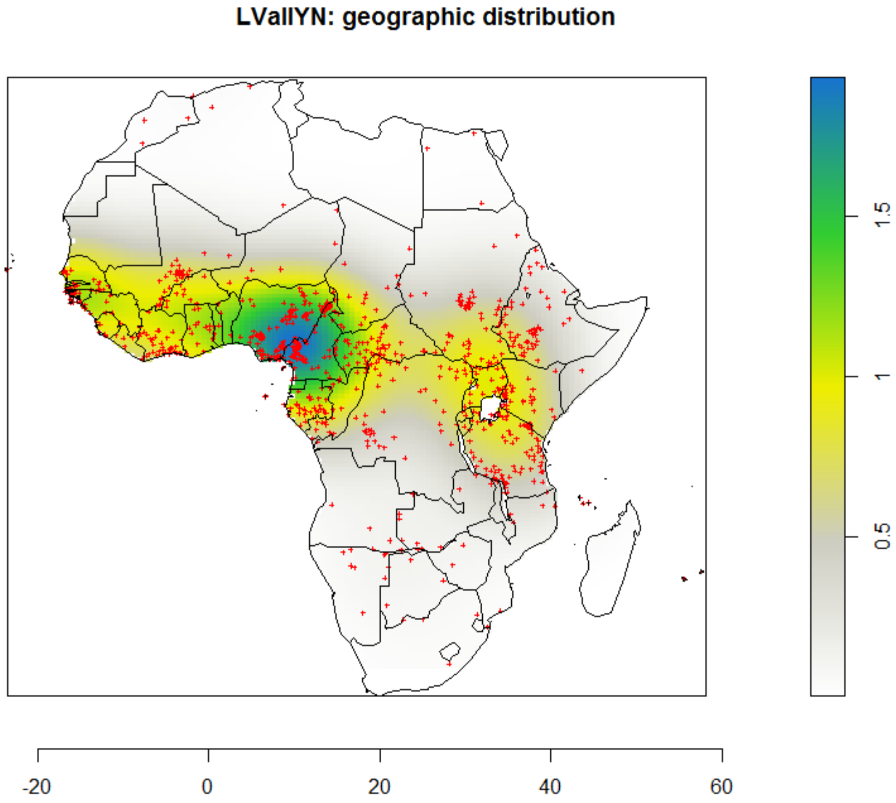
FR+ET: Mais, ce n’est pas mè-ŋmgbɜ́m (FOC), c’est mə̀-ŋmgbɜ́m (FOC)

LV, SIC-ACCENT & C-EMPHASIS PROSODY

- In a broader perspective, **C-emphasis prosody** is a very good candidate for the role of a major driving force behind the emergence of several other types of sounds, such as labial flaps, bilabial trills, and possibly clicks

SPATIAL DISTRIBUTION


LVall: geographic distribution


LVallYN: geographic distribution

LVall

1074 languages with frequency data:

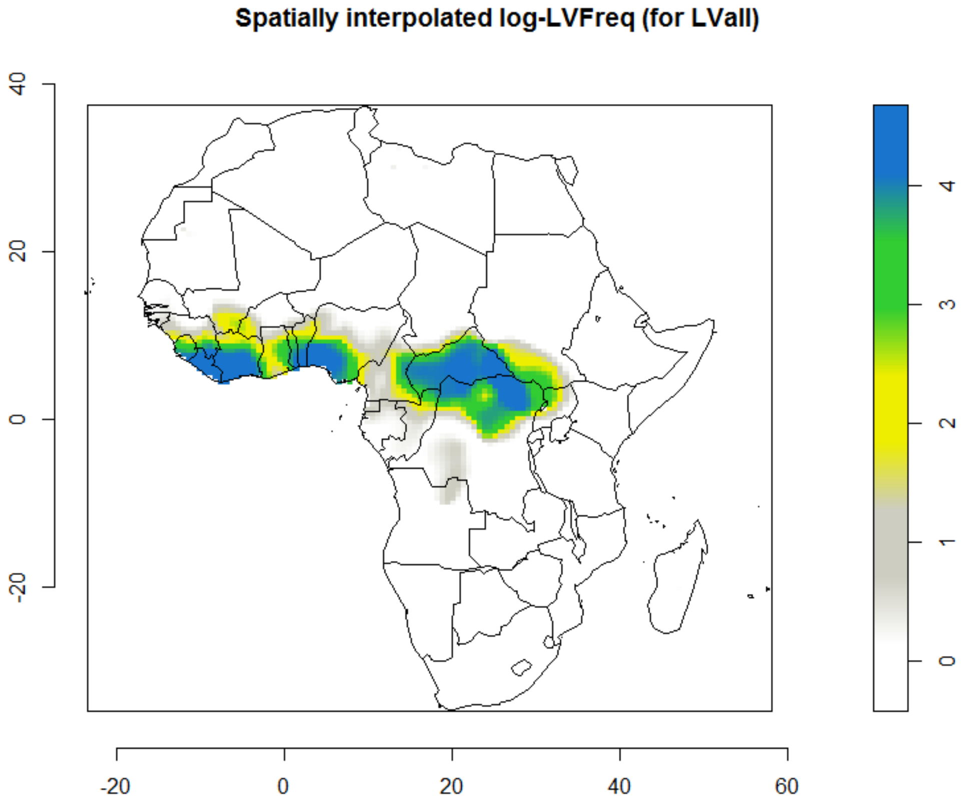
- LV & their frequency is known (336 lgs)
- No LV

LVallYN

1304 languages with LV:

- LV & their frequency is known (336 lgs)
- LV, but no frequency data (230 lgs)
- No LV

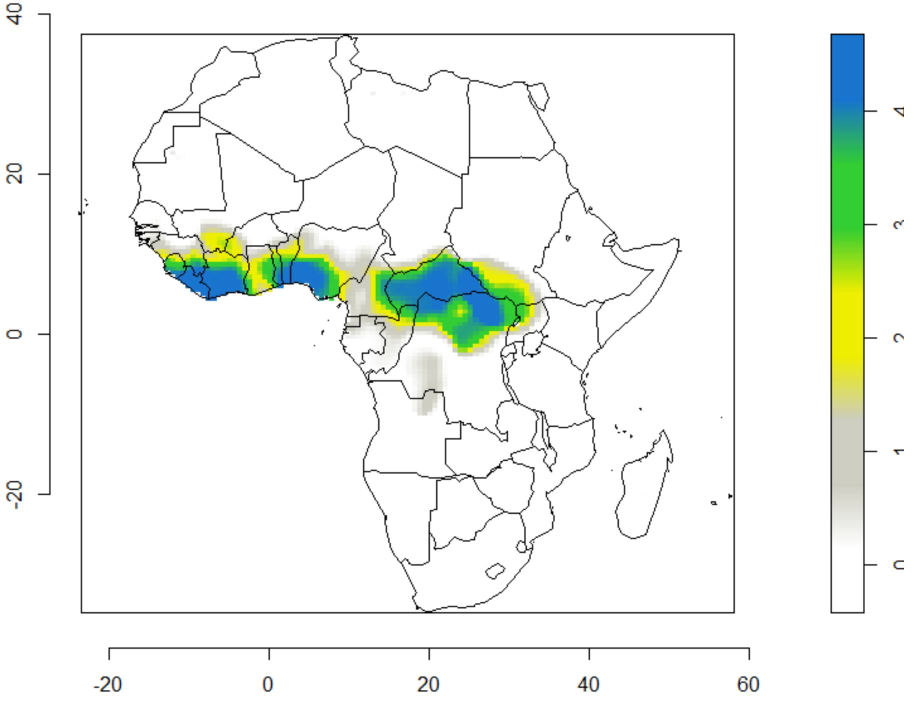
SPATIAL DISTRIBUTION



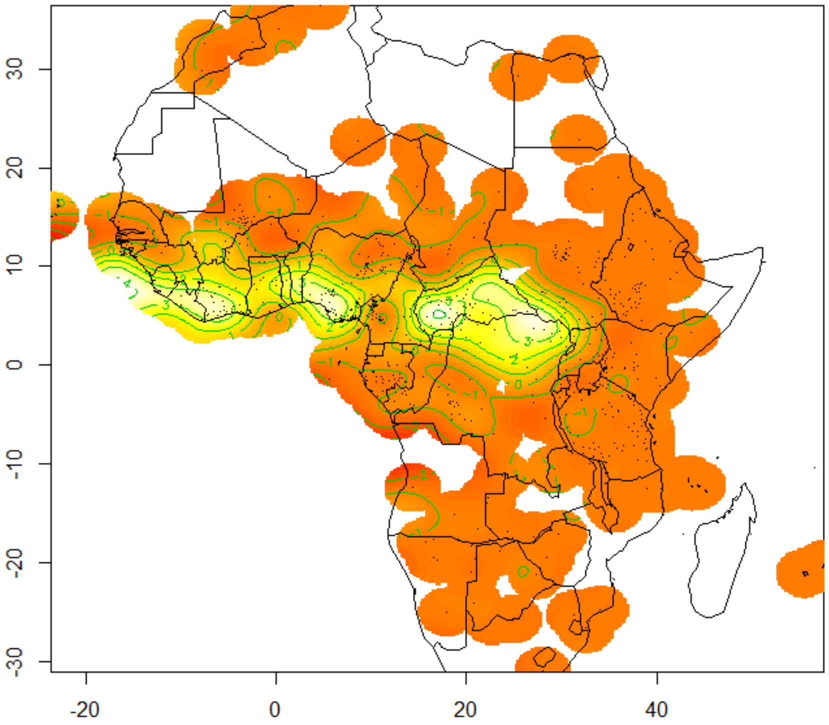
- 2 clearly separated clusters
  - Coastal West Africa (possibly itself composed of 2 sub-clusters)
  - Central Africa
- possibly, + 1 less prominent cluster
  - SE Mali & SW Burkina-Faso
- 1 major spatial discontinuity
  - NE Nigeria & Cameroon
- 1 minor spatial discontinuity
  - Ghana

SPATIAL DISTRIBUTION

Spatially interpolated log-LVFreq (for LVall)



Regression surface of GAM of log-LVFreq
as a function of longitude and latitude



(thin-plate regression splines, k=16, family=Gaussian)

MODEL CROSS-VALIDATION

- How can we cross-validate our model?

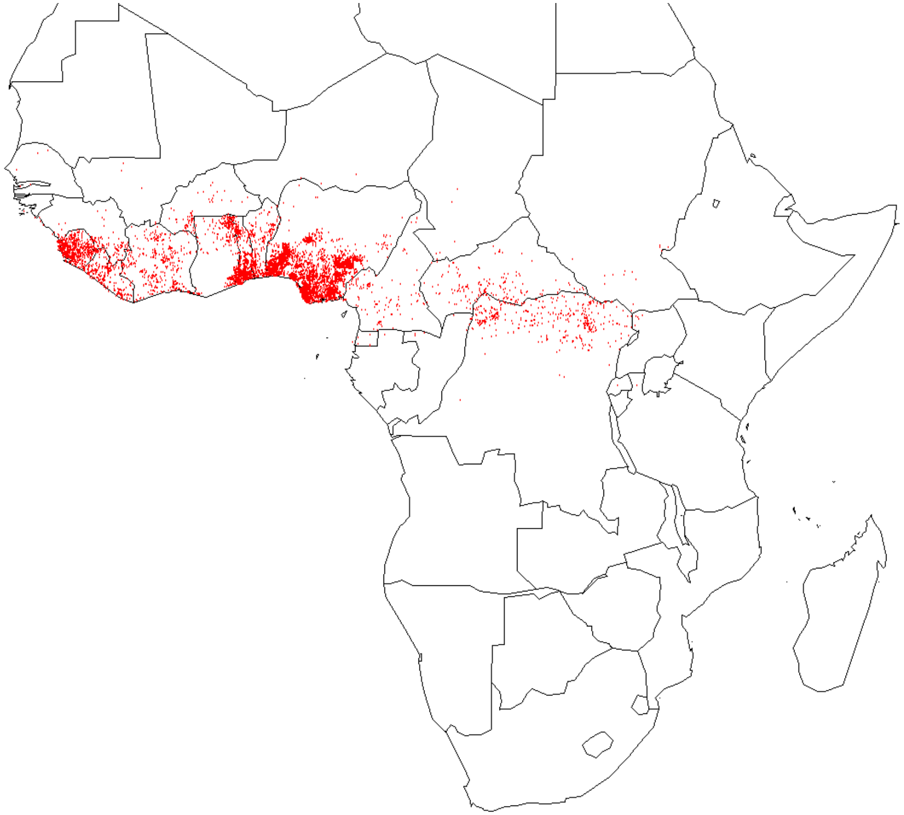
- Spatial distribution of settlement names spelled with a LV (such as “kp”, “gb”, Yoruba “p”) on the assumption that:

  H_0: Frequency of settlement names with LV in a given area should roughly correlate with (be representative of) lexical frequency of LV in the languages spoken in the area

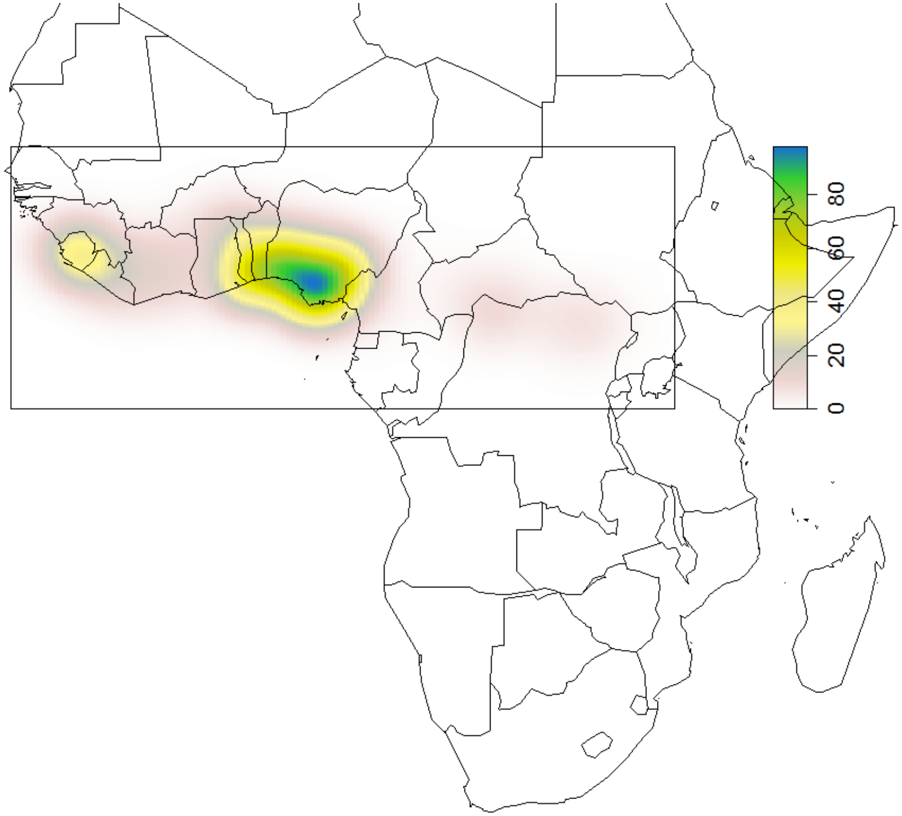
- Big data approach: quantity compensates for quality

- Settlement names data source: GeoNames.org

MODEL CROSS-VALIDATION

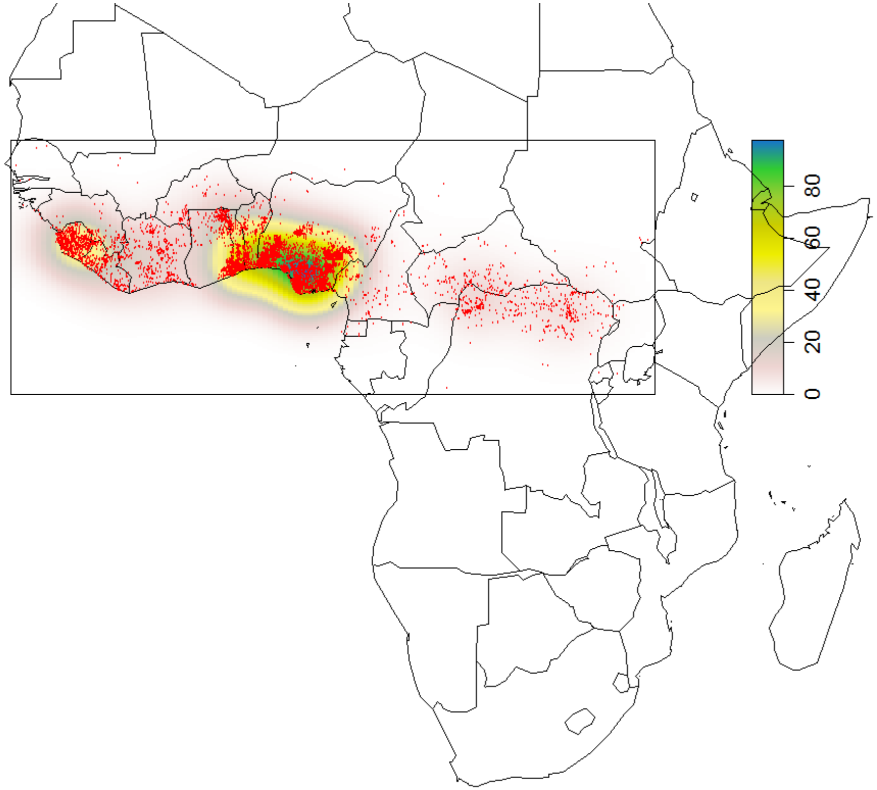


Unique settlement names with a <LV> (<kp>, <gb>, Nigerian Yoruba <p>)

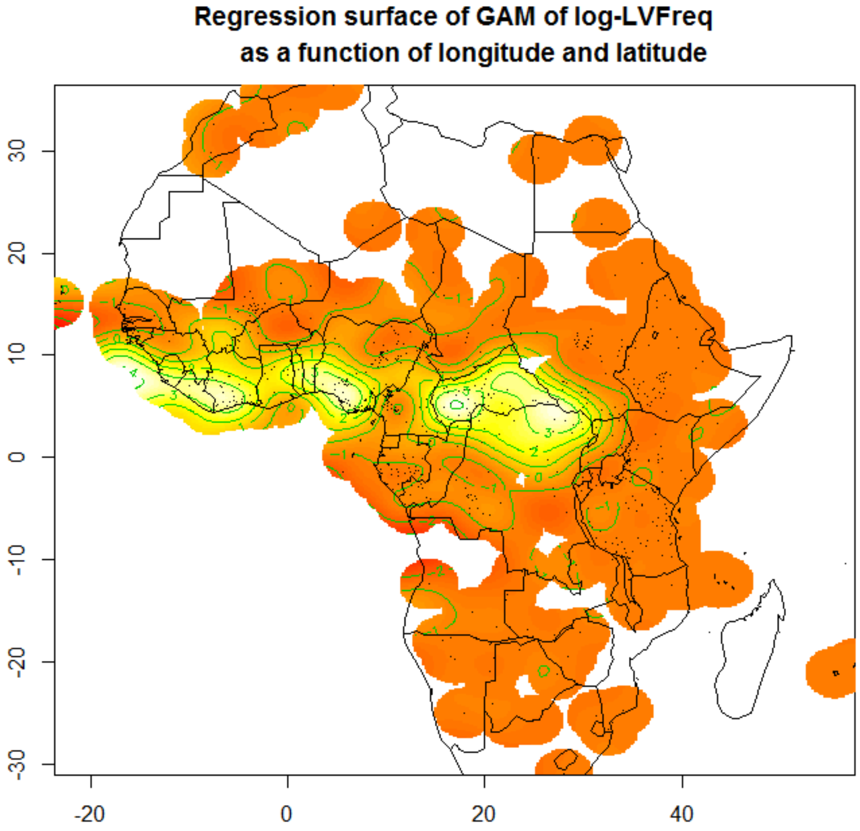


Spatial intensity of unique settlement names with a <LV>

MODEL CROSS-VALIDATION



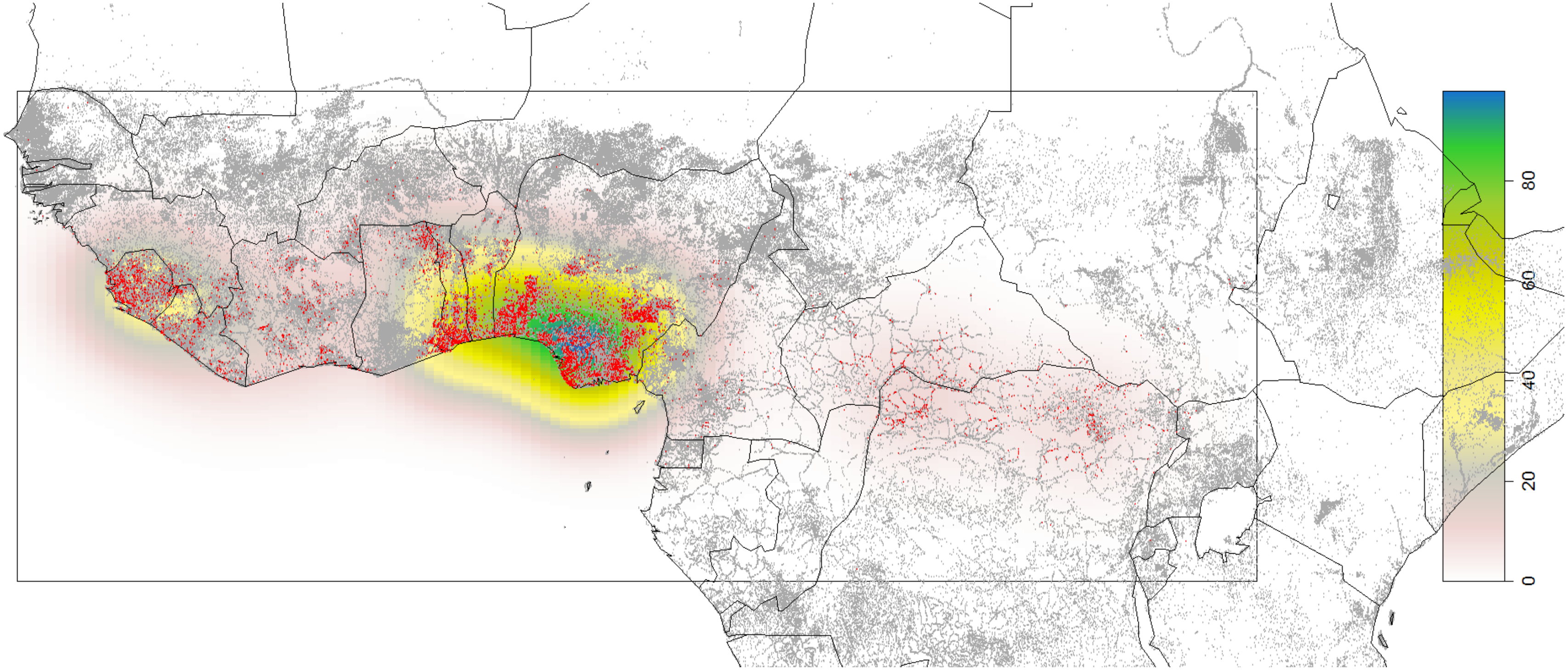
Spatial intensity of unique settlement names
with a <LV>


Regression surface of GAM of log-LVFreq
as a function of longitude and latitude

(thin-plate regression splines, k=16, family=Gaussian)

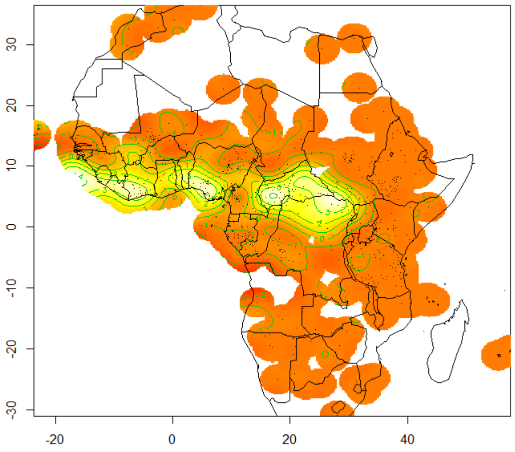
MODEL CROSS-VALIDATION

- The significance of the clusters should be evaluated against the general population density in the respective areas:
  - The seeming weakness of the E-most cluster is an artefact of the low population density in Central Africa
  - Both discontinuities are significant



HISTORICAL IMPLICATIONS

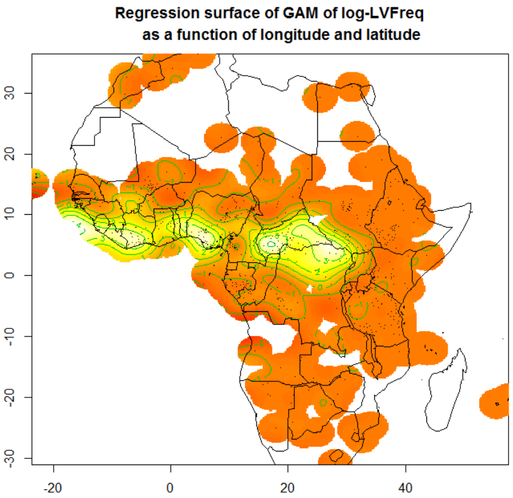
Regression surface of GAM of log-LVFreq as a function of longitude and latitude

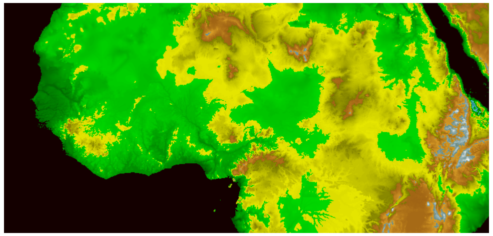


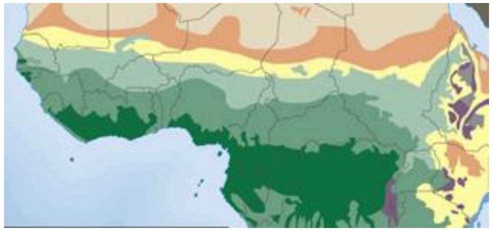
(thin-plate regression splines, k=16, family=Gaussian)

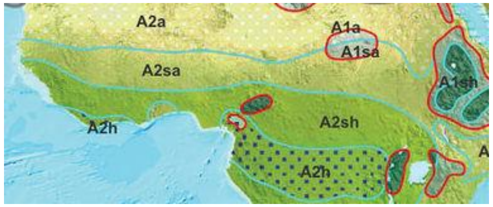
- Logically, the 3 major zones of high LVFreq (and the possible minor zone) are most likely to be **refuge zones**:
  - Typologically, LV are rare
  - Several emergent hotbeds of high LVFreq historically independent of each other are unlikely

HISTORICAL IMPLICATIONS


Regression surface of GAM of log-LVFreq
as a function of longitude and latitude
30
20
10
0
-10
-20
-30
-20
0
20
40



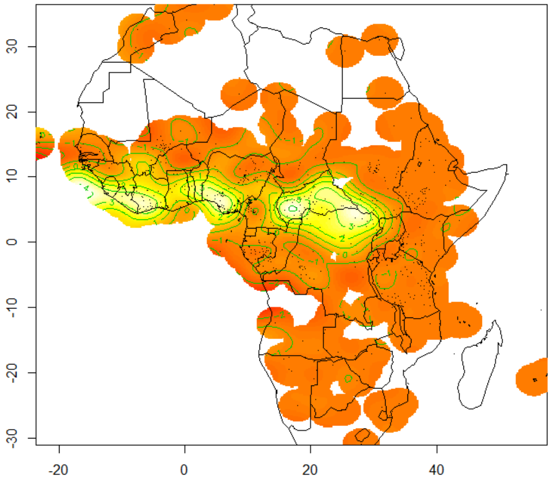




- Geographically, the 3 major zones of high LVFreq (and the possible minor zone) are **refuge zones**: mostly forests delimited by natural boundaries (sea, savanna, mountain ranges)
- Ghana discontinuity ≈ Dahomey forest gap
- NE Nigeria & Cameroon discontinuity ≈ Adamawa Plateau, Cameroon mountains

HISTORICAL IMPLICATIONS

Regression surface of GAM of log-LVFreq
as a function of longitude and latitude

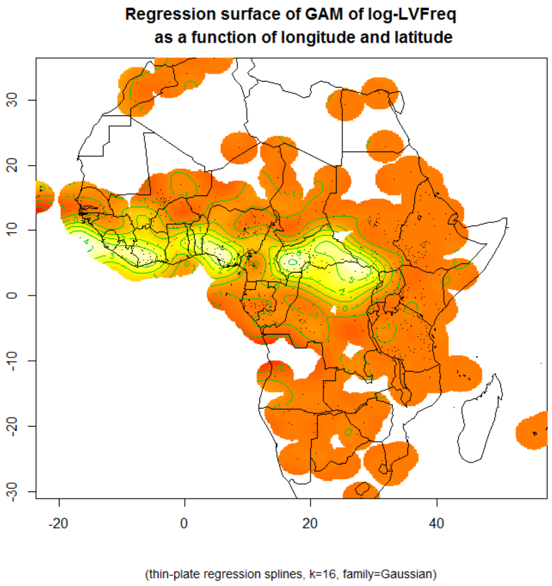


(thin-plate regression splines, k=16, family=Gaussian)

- “hotbeds” → older presence of LV and ultimately SIC-accent and C-emphasis prosody

- Given the refuge zone nature of the “hotbeds”, they are probably “hotbeds” not so much for spread but for **retention** of the feature C-emphasis and derived features, inlcuding SIC-accent & LV, present in the original population

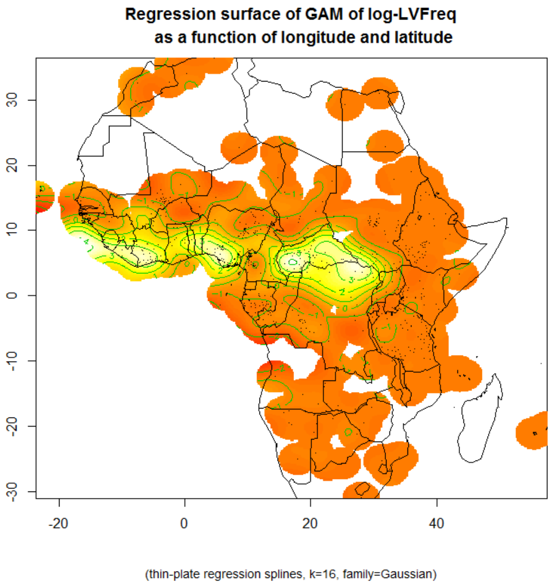
HISTORICAL IMPLICATIONS


Regression surface of GAM of log-LVFreq
as a function of longitude and latitude
30
20
10
0
-10
-20
-30
-20
0
20
40
(thin-plate regression splines, k=16, family=Gaussian)

- Genetic build-up of hotbeds & their outskirts is diverse:
  - W: mostly Niger-Congo, except the extreme W
  - E: Gbaya, Ubangian, parts of Central Sudanic
- Linguistically, the original population with CE-prosody/SIC-accent/LV may be almost any of these (unlikely Niger-Congo or Central Sudanic) or none
- Hotbeds as refuge zones & retention:

- hotbeds || language shift
- outskirts || change in language contact situations

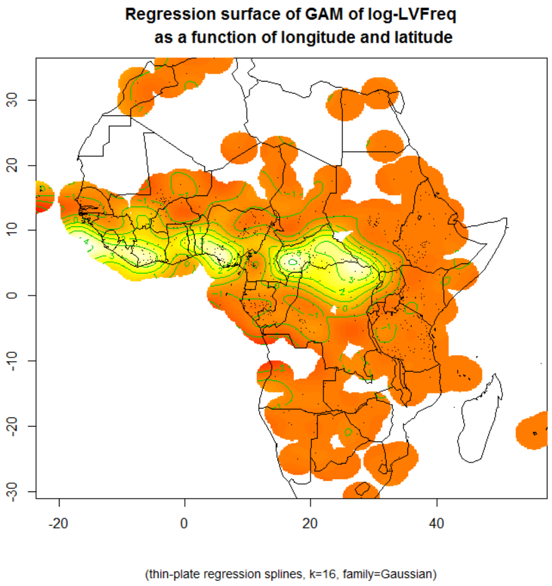
HISTORICAL IMPLICATIONS


Regression surface of GAM of log-LVFreq
as a function of longitude and latitude
30
20
10
0
-10
-20
-30
-20
0
20
40
(thin-plate regression splines, k=16, family=Gaussian)

- LV (and correlated phonetic and phonological features) should not be reconstructed for Proto Niger-Congo or any of its major branches

- We should also be very cautious about reconstructing LV for lower-level branches (problems with “the majority wins” rule)

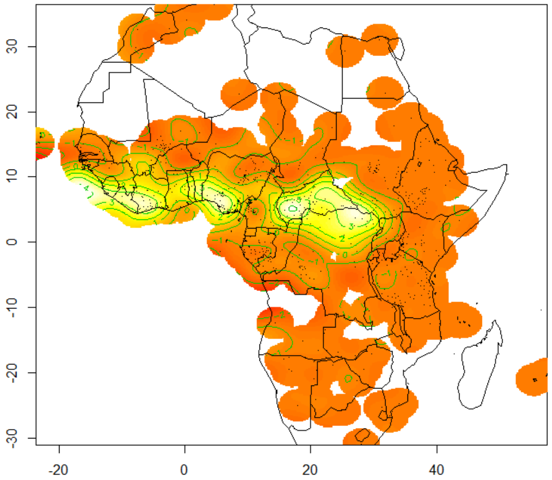
HISTORICAL IMPLICATIONS


Regression surface of GAM of log-LVFreq
as a function of longitude and latitude
30
20
10
0
-10
-20
-30
-20
0
20
40
(thin-plate regression splines, k=16, family=Gaussian)

- A rather northern localization of the homelands of most major branches of Niger-Congo
- In grassland and savanna ecoregions
- The homeland of Proto Niger-Congo is then likely to have been located in the northern part of the former extent of grassland and savanna ecoregions
- Probably, somewhere in present-day Sahel or southern Sahara.

HISTORICAL IMPLICATIONS

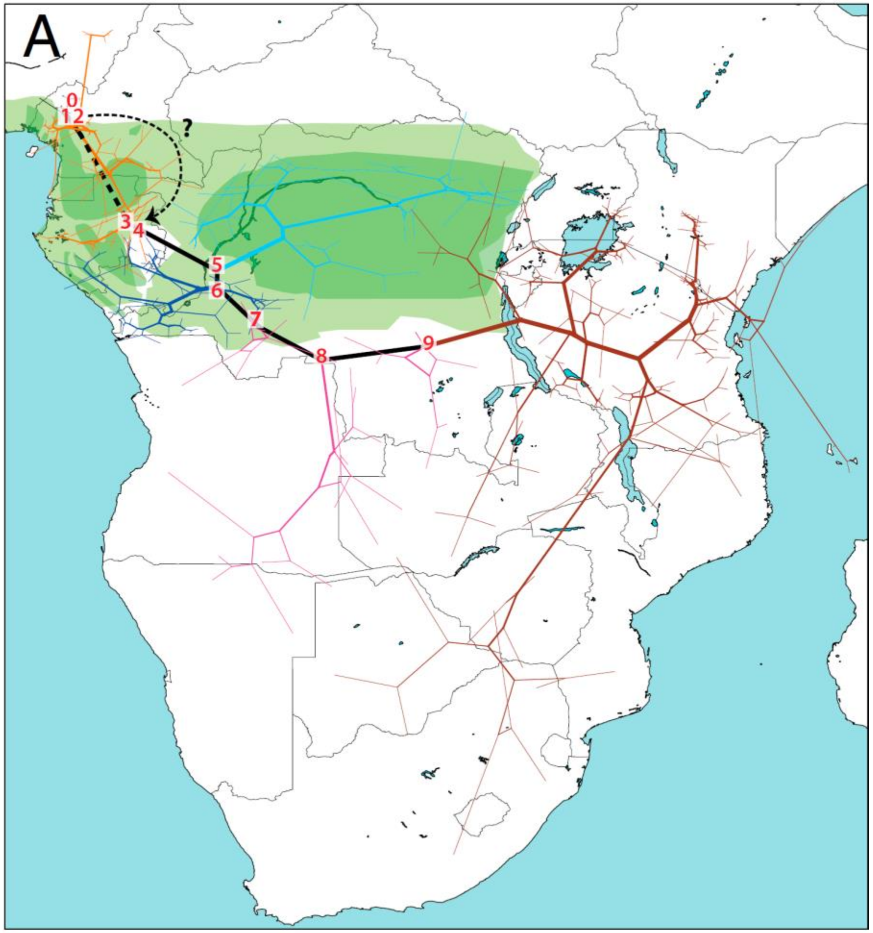
Regression surface of GAM of log-LVFreq as a function of longitude and latitude

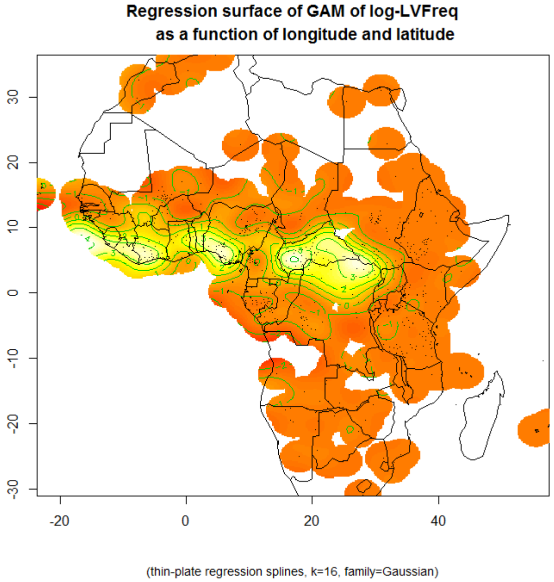


(thin-plate regression splines, k=16, family=Gaussian)

- Bantoid & Adamawa appear to have arrived in the area relatively recently
- Bantoid may have passed it & then re-entered or just entered late
- The spread of Bantoid must have also been rather quick without much language shift involved (except in the N of Congos)
- This model also supports the "East-out-of-West" hypothesis of the E Bantu emergence with the E Bantu break-off point somewhere south of the rainforest

HISTORICAL IMPLICATIONS




Regression surface of GAM of log-LVFreq
as a function of longitude and latitude
30
20
10
0
-10
-20
-30
-20
0
20
40
(thin-plate regression splines, k=16, family=Gaussian)

- This model also supports the "East-out-of-West" hypothesis of the E Bantu emergence with the E Bantu break-off point somewhere south of the rainforest

Grollemund et al. (2015:3)