# THE LEXICAL FREQUENCY OF LABIAL-VELAR STOPS AS A WINDOW ON THE LINGUISTIC PREHISTORY OF NORTHERN SUB-SAHARAN AFRICA

Dmitry Idiatov & Mark Van de Velde

LLACAN, CNRS, Sorbonne-Paris Cité, INALCO

dmitry.idiatov@cnrs.fr

mark.vandevelde@cnrs.fr

# INTRODUCTION

# INTRODUCTION

- Northern sub-Saharan Africa is obviously a spread zone with a marked areal distribution of various linguistic features
  - Macro-Sudan belt (Güldemann 2008)
  - Sudanic zone (Clements & Rialland 2008)
  - …
- LV are common in NSSA languages but typologically they are known to be rather rare (e.g., Cahill 2008, Maddieson 2011)

**$Q_1$:** What can the areality of LV tell us about the history of the languages of NSSA and of the populations speaking these languages?

# INTRODUCTION

We start with the observation that:

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

We proceed by looking into the following questions:

- Are LV "normal" phonemes in the languages of NSSA in general?
- Are the distributions of LV within the lexicons random?
- What is the spatial pattern of the LV frequency distribution?

We conclude by providing an interpretation of our findings:

- What can our findings tell us about the spatio-temporal dynamics of the languages of NSSA and the populations speaking these languages?

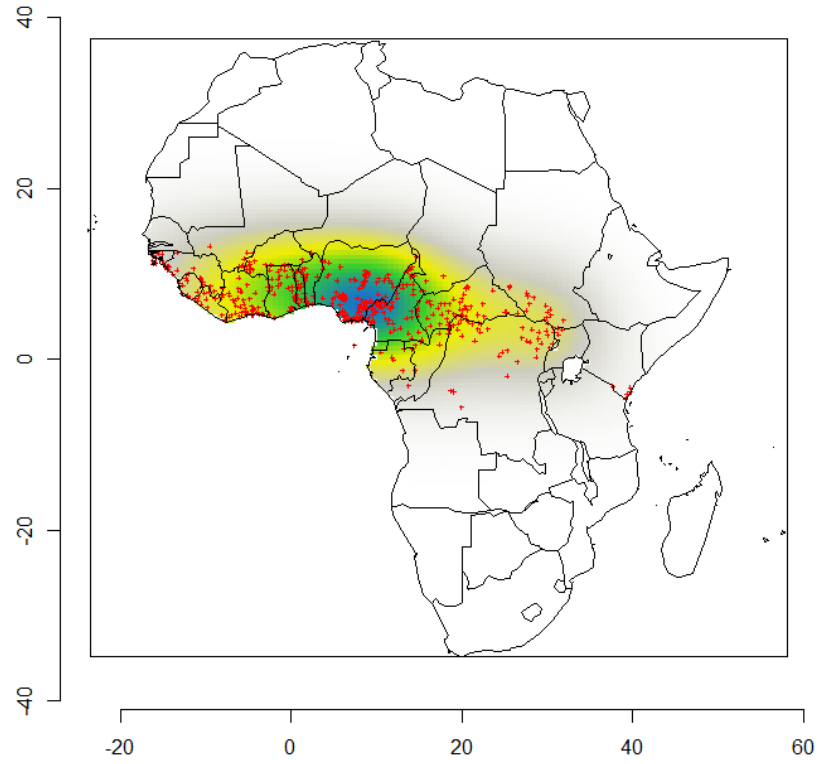
LV data sources:

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

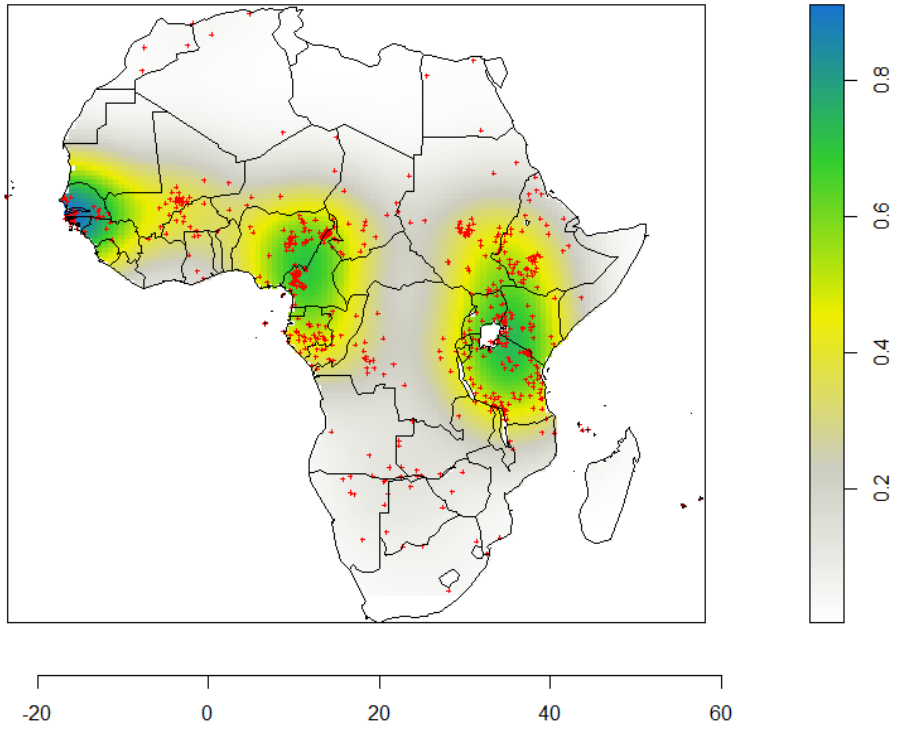
The composition of our sample (1304 languages):

- 336 languages with LV & LV frequency is known
- 230 languages with LV & LV frequency is not known
- 738 languages without LV

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)

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

$LV_O$ - observed LV count

$LV_E$ - expected LV count

$T_{LV}$ - LV token

$T_C$ - any C token

$P_{LV}$ - LV phoneme

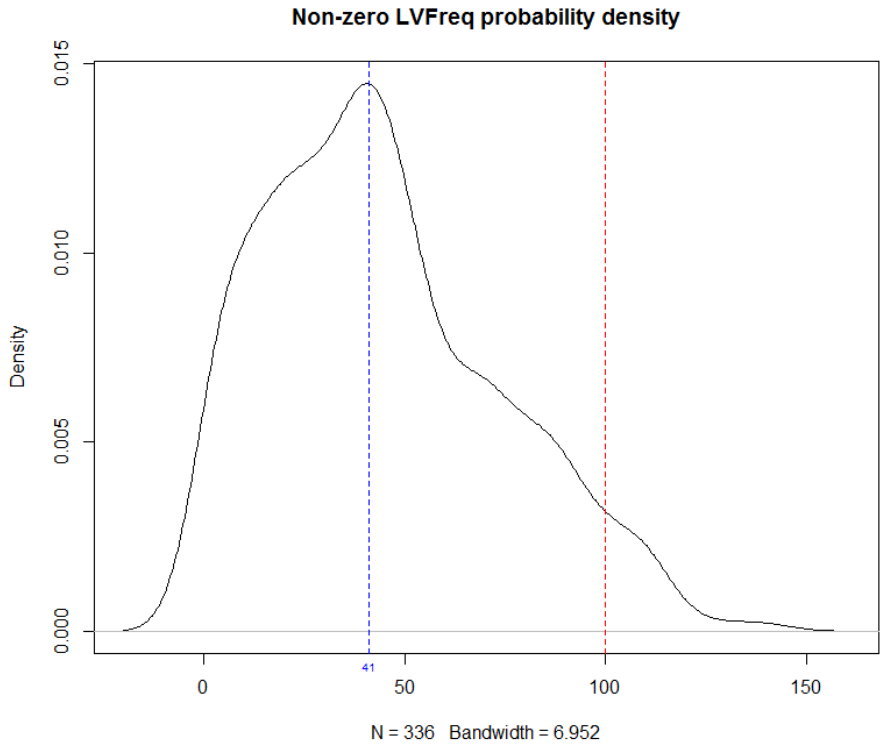
$P_C$ - any C phoneme

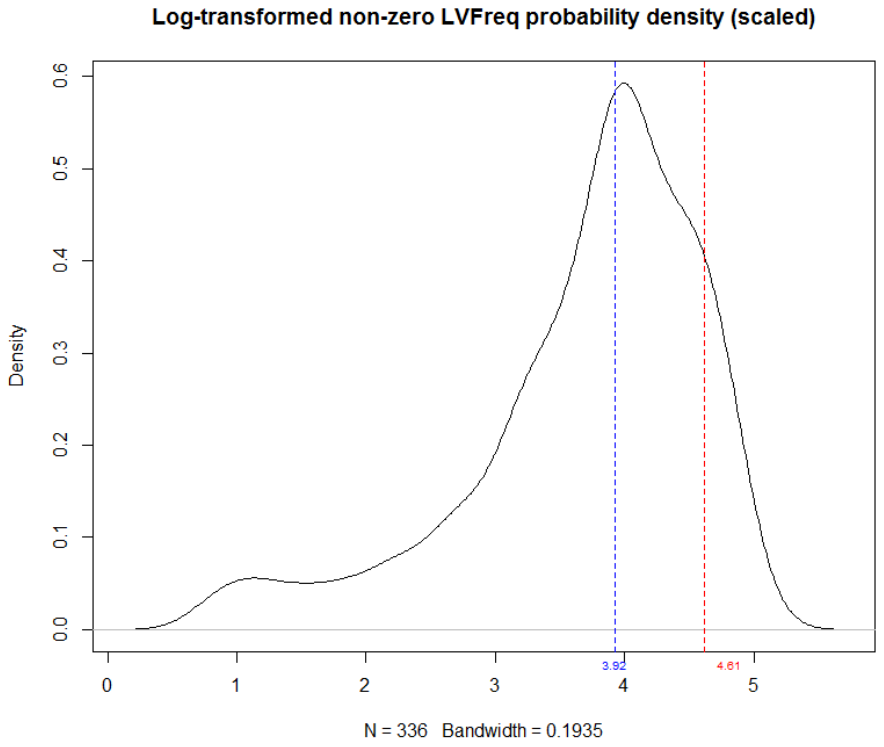
# LVFreq estimation

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 under the $H_0$

# LV FREQUENCY ESTIMATION




median reference LVFreq

- LV are relatively **rare phonemes** in most languages that have them, which is in accordance with their typological rarity

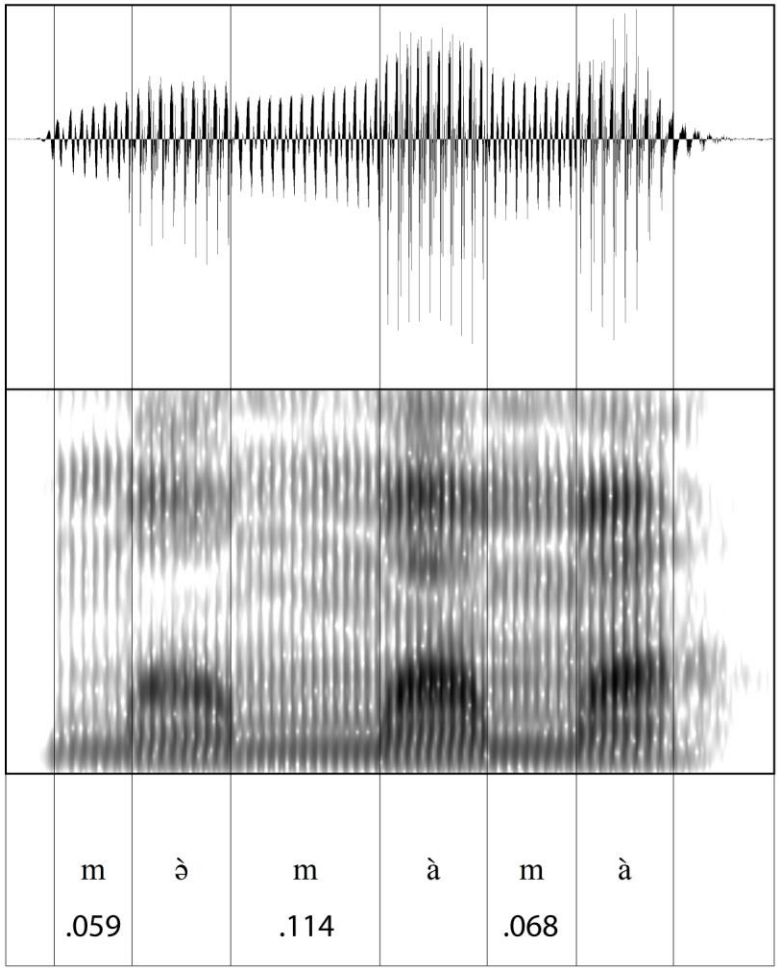
## Are the distributions of LV within the lexicons random?

- 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, SI C-ACCENT & C-EMPHASIS PROSODY

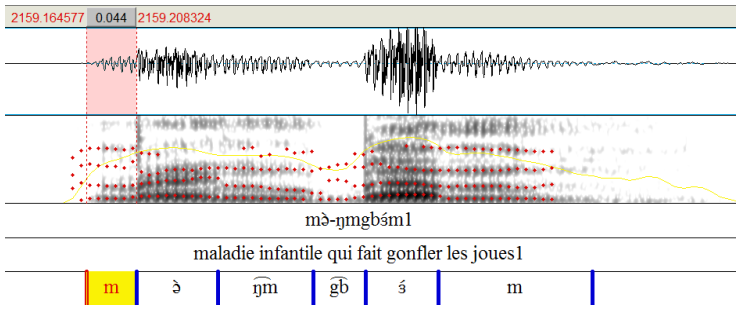
- The correlation [LV ~ "expressive" vocabulary] is not independent of the correlation [LV ~ stem-initial position]
- **SI C-accent** (as a manifestation of a more general phenomenon of **C-emphasis prosody**) is a very important factor behind the emergence of LV in NSSA

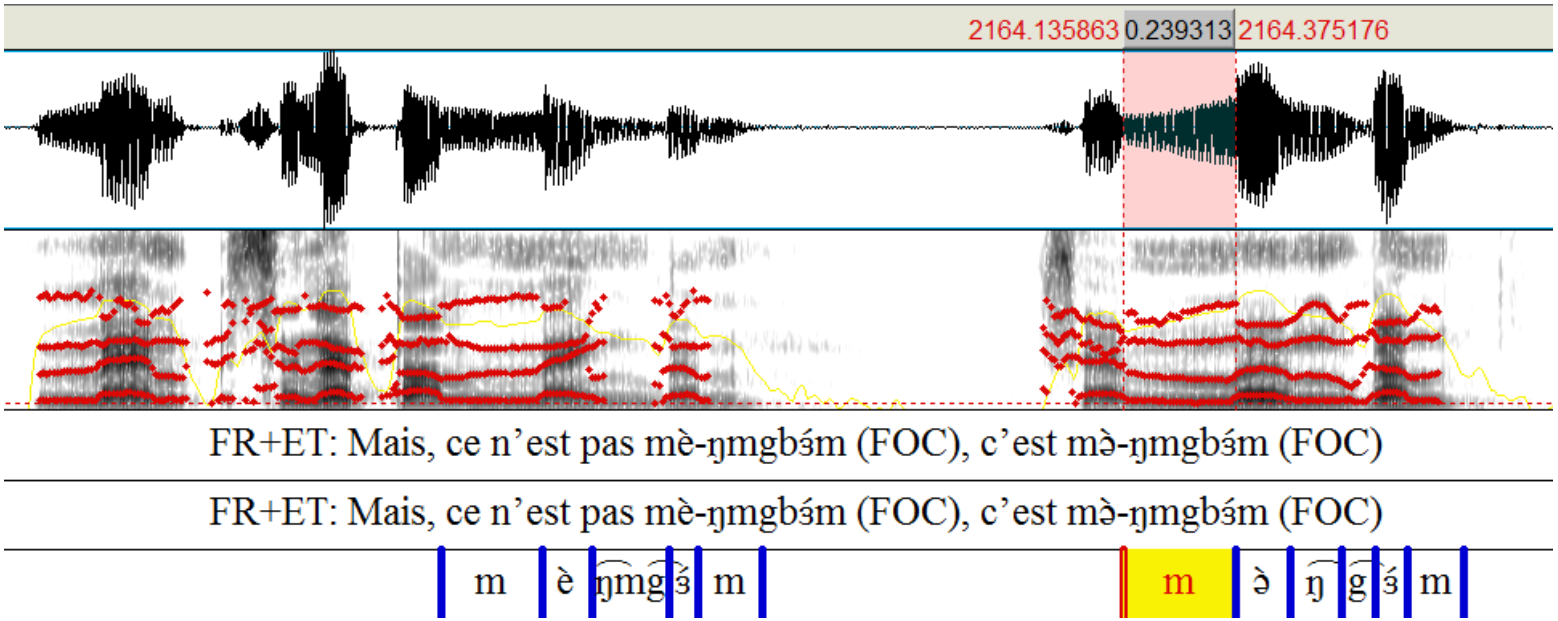
Consonant length in the nonsense word *mə-màmà* (Eton, Bantu A70)



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

Eton (A70)



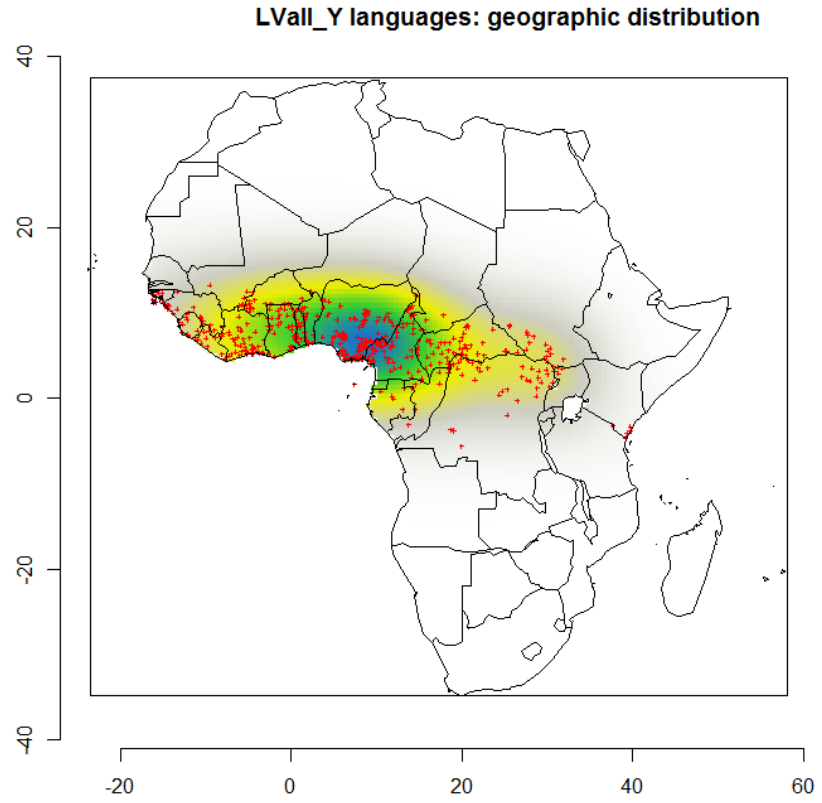



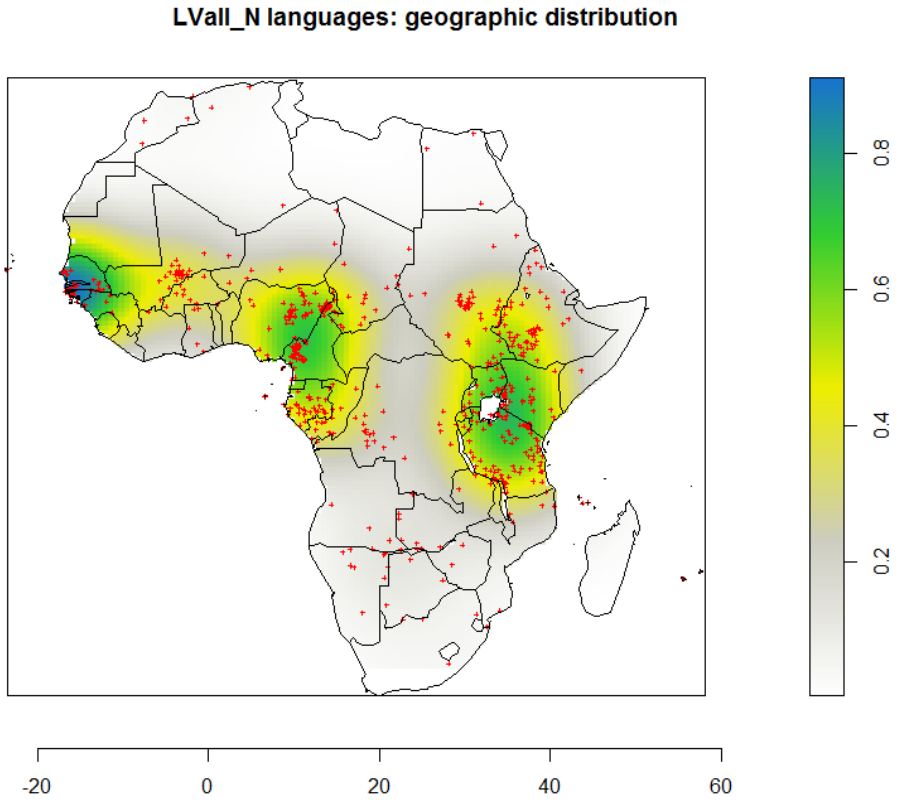
# LV, SIC-ACCENT & C-EMPHASIS PROSODY

- 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 clicks

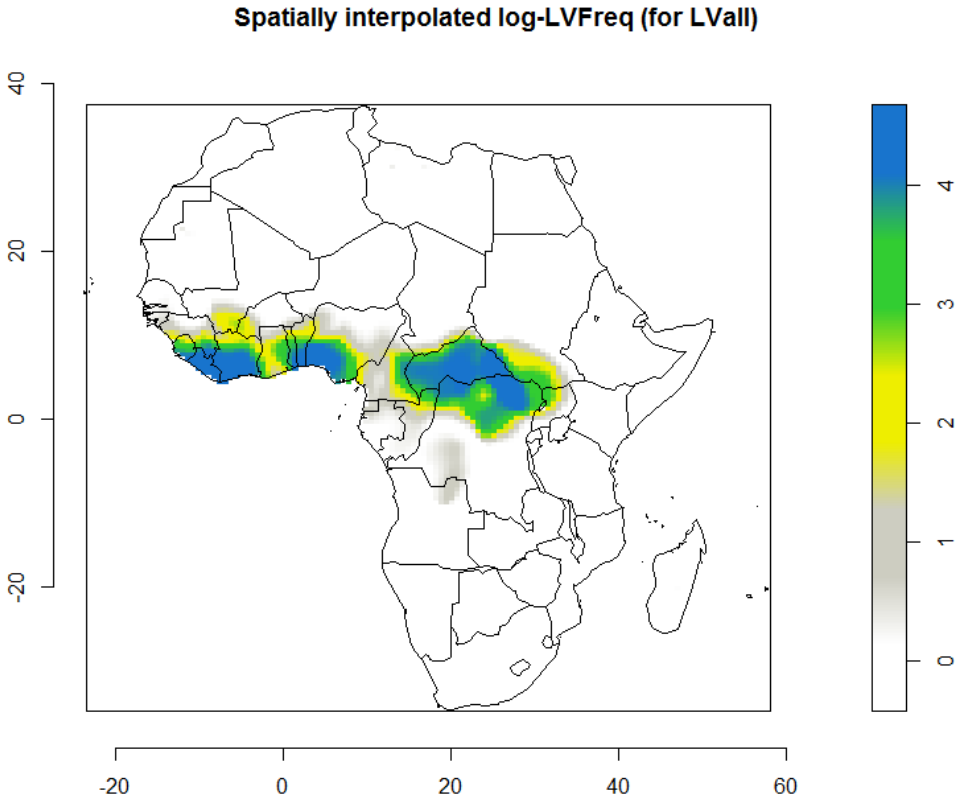
Langage, Langues et Cultures d'Afrique Noire
Llacan

# SPATIAL DISTRIBUTION



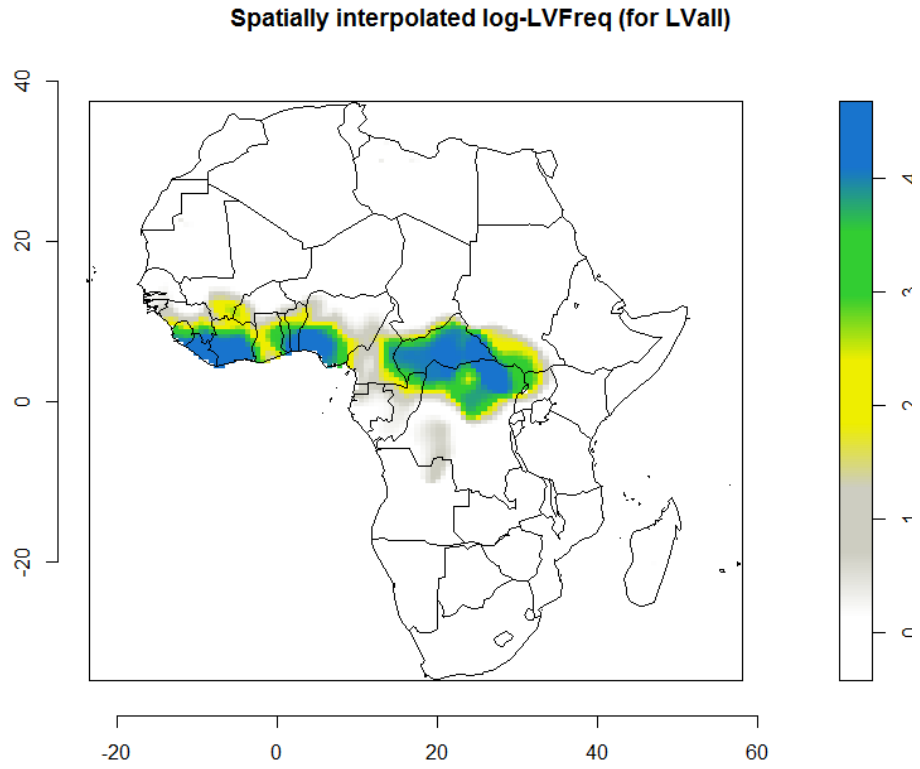



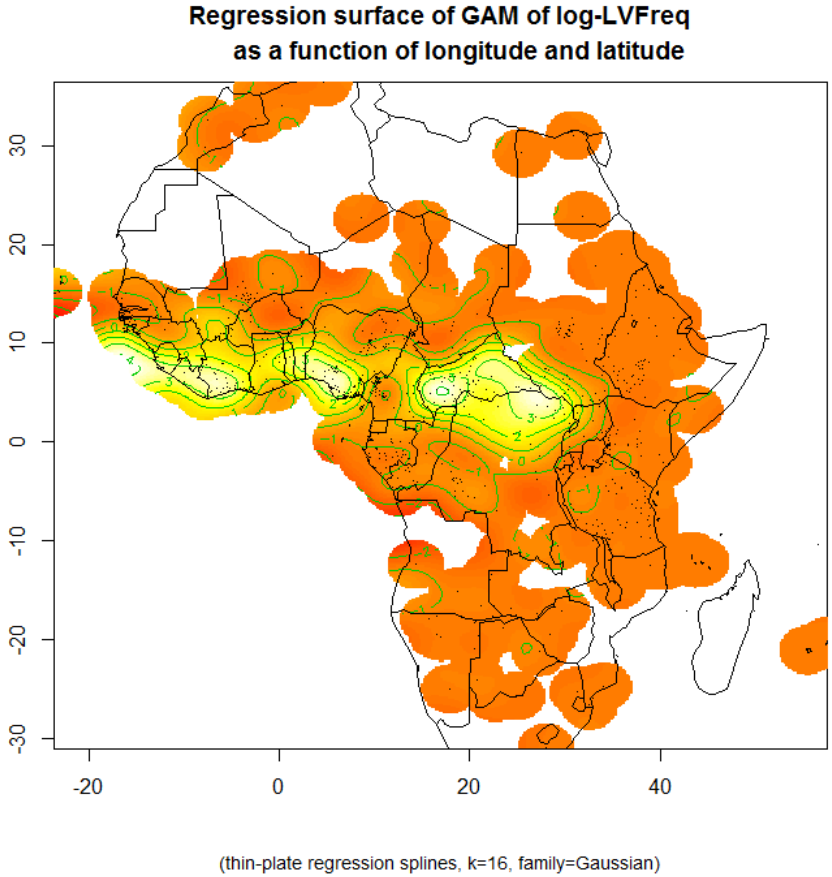
# SPATIAL DISTRIBUTION




- 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




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

# MODEL CROSS-VALIDATION

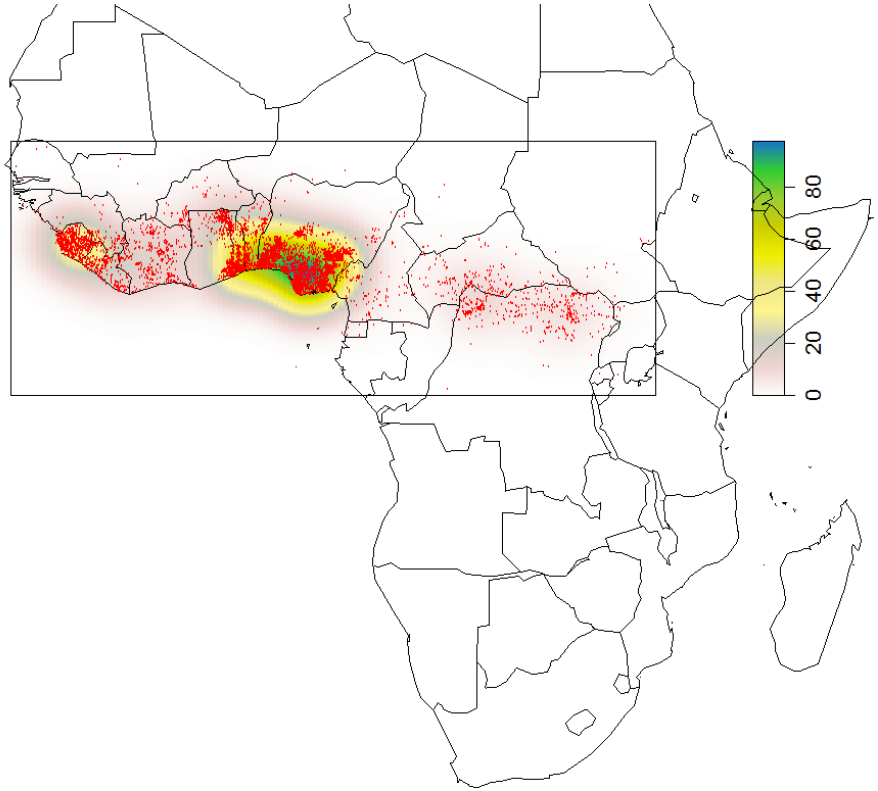
- How can we cross-validate our model?

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

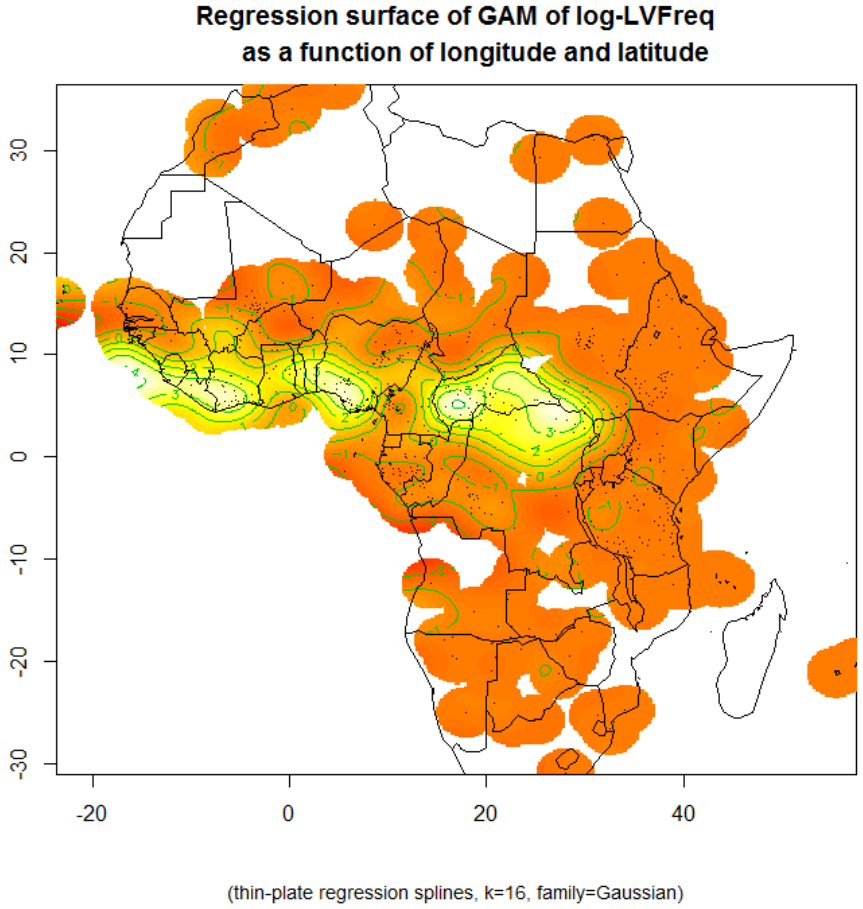
  $H_0$: Frequency of settlement names with LV in a given area should roughly correlate with (be representative of) lexical frequency of LV in the languages spoken in the area

- Big data approach: quantity compensates for quality
- Settlement names data source: GeoNames.org

# MODEL CROSS-VALIDATION

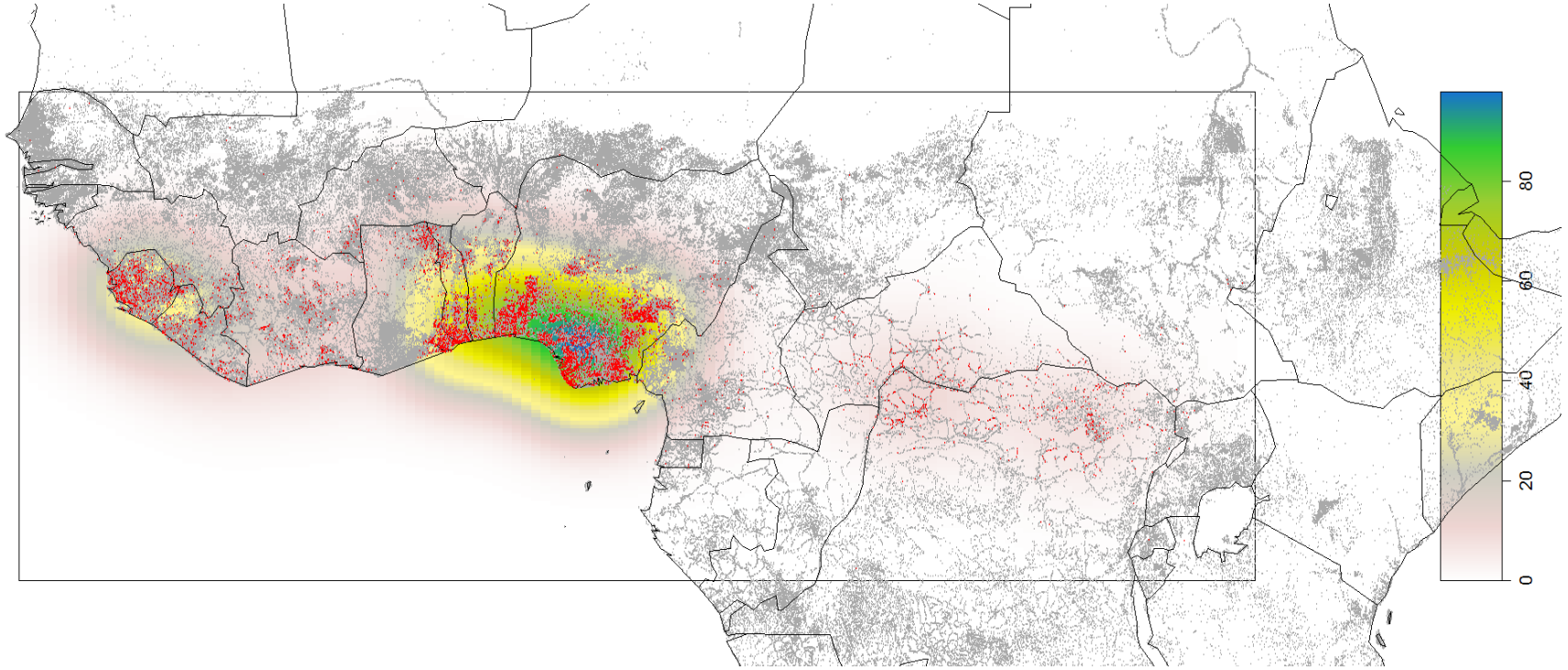



Spatial intensity of unique settlement names with a <LV>

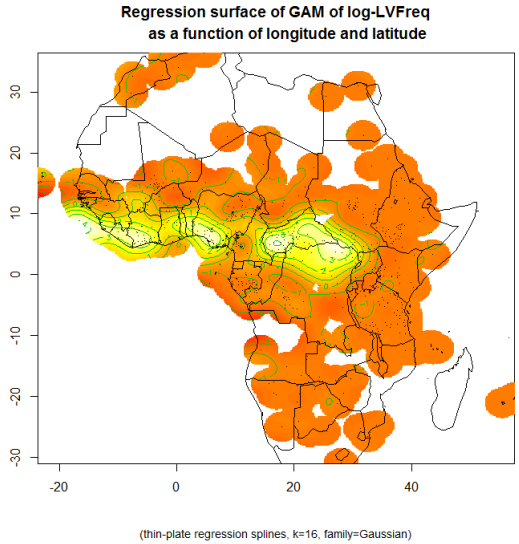


# MODEL CROSS-VALIDATION

- 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

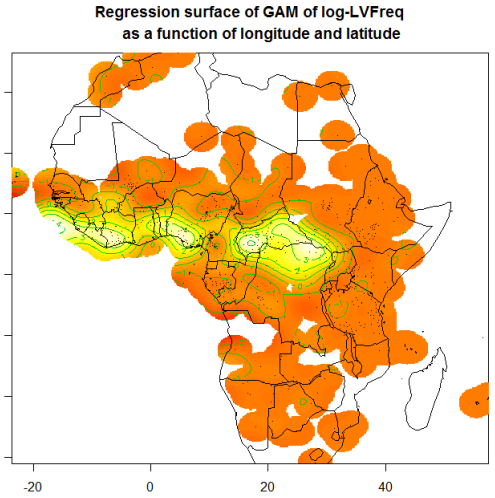


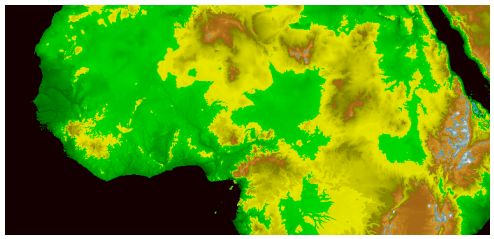
# HISTORICAL IMPLICATIONS: REFUGE ZONES

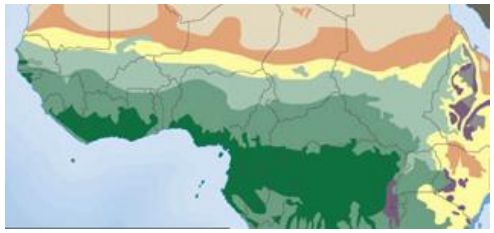


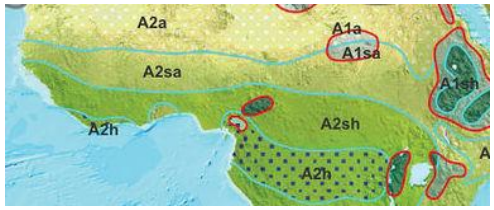

- Logically, the 3 major zones of high LVFreq (and the possible minor zone) are most likely to be **refuge zones**:
  - Typologically, LV are rare
  - Several emergent hotbeds of high LVFreq historically independent of each other are unlikely

# HISTORICAL IMPLICATIONS: REFUGE ZONES



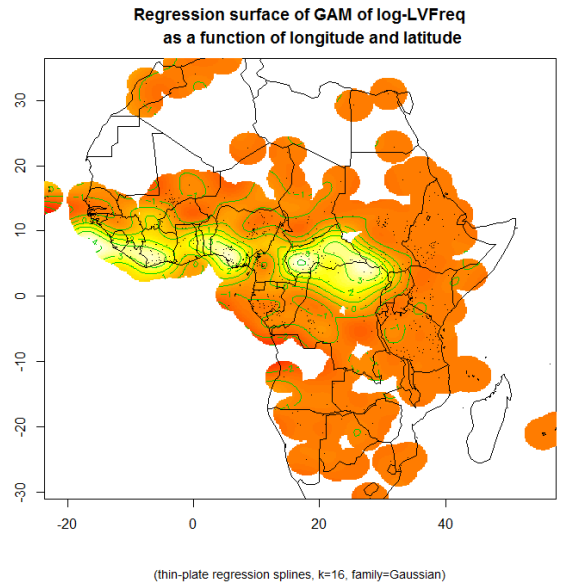







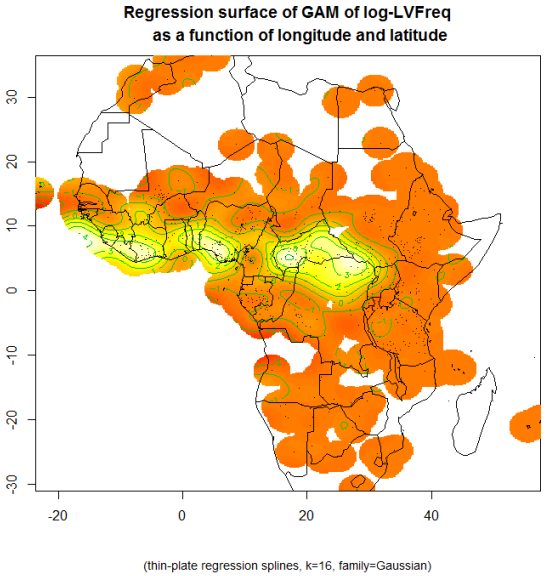

- Geographically, the 3 major zones of high LVFreq (and the possible minor zone) also look like **refuge zones**: mostly forests delimited by natural boundaries (sea, savanna, mountain ranges)
- Ghana discontinuity ≈ Dahomey forest gap
- NE Nigeria & Cameroon discontinuity ≈ Adamawa Plateau, Cameroon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

# HISTORICAL IMPLICATIONS: RETENTION




- "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 propagation but for **retention** of the feature C-emphasis and derived features, inlcuding SIC-accent & LV, present in the original population

# HISTORICAL IMPLICATIONS: SHIFT VS CONTACT



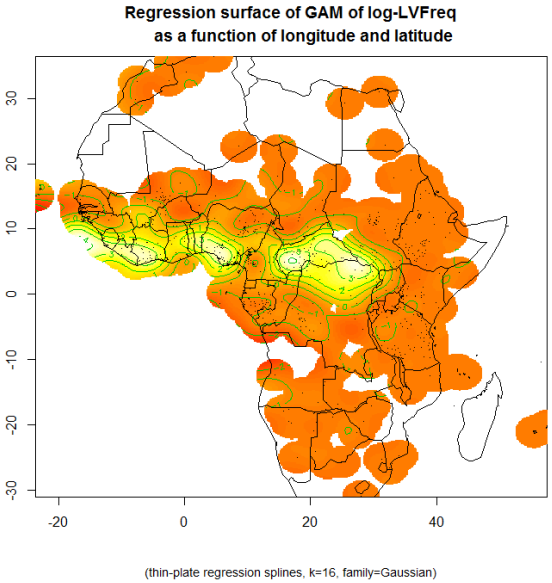

- Genetic make-up of the hotbeds:
  - W: mostly Niger-Congo
  - E: Gbaya, Ubangian, parts of Central Sudanic
- Genetic make-up of the periphery:
  - W: mostly Niger-Congo
  - E: Niger-Congo, parts of Central Sudanic

- Linguistically, the original population with C-emphasis/SIC-accent/LV may be almost any of these (unlikely Niger-Congo or Central Sudanic) or none

- Hotbeds as refuge zones & retention:

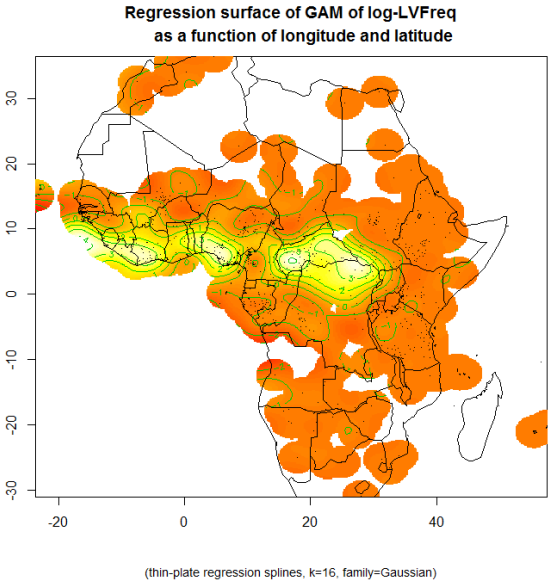
- hotbeds ‖ language shift
- periphery ‖ change in language contact situations

# HISTORICAL IMPLICATIONS: NIGER-CONGO




- 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: NIGER-CONGO

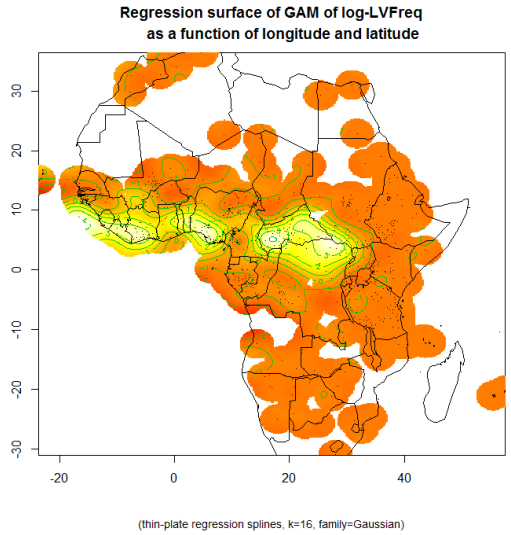



- A rather northern localization of the homelands of most major branches of Niger-Congo in grassland and savanna ecoregions

- 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

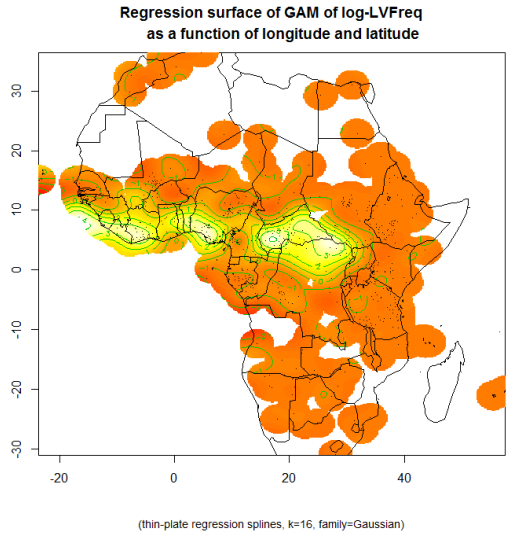
- Probably, somewhere in present-day Sahel or southern Sahara.

# HISTORICAL IMPLICATIONS: BANTOID




- Bantoid, Adamawa and Chadic are responsible for the major discontinuity around Cameroon & NE Nigeria
- The majority of Bantu languages (a major Bantoid subgroup) are spoken outside of the hotbeds of high LVFreq
- Bantoid & Adamawa are likely to have arrived in the discontinuity area relatively recently
- Bantoid may have passed it & then re-entered or just entered late

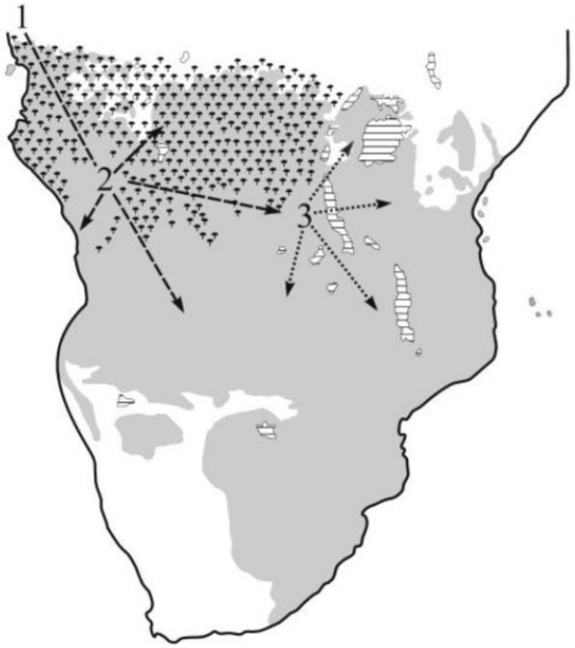
# HISTORICAL IMPLICATIONS: BANTU EXPANSION



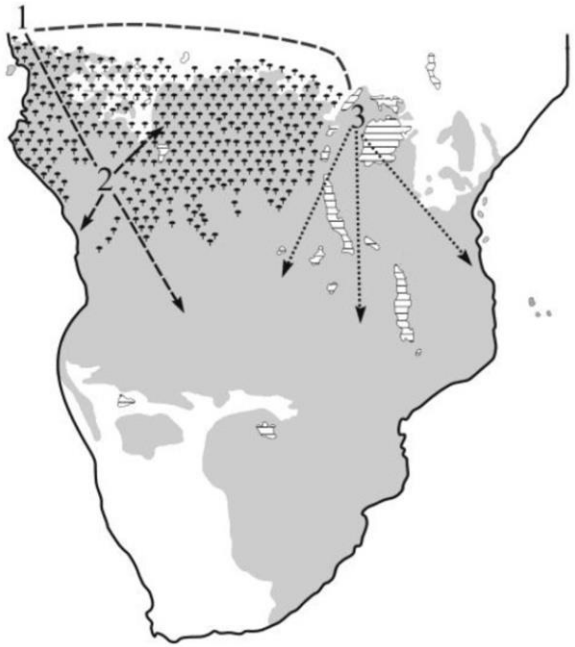

- Further spread of Bantoid, particularly Bantu expansion, must have happened without much language shift from any original "LV" populations involved
- The only Bantu group with a relatively high LVFreq are languages in the N of DRC which moved into the hotbed of high LVFreq from SW
- The spread of Bantu in the N or DRC is most likely to have been a spread of languages through language shift much more than a spread of Bantu speaking populations

# HISTORICAL IMPLICATIONS: BANTU EXPANSION

## Two principal models of Bantu expansion



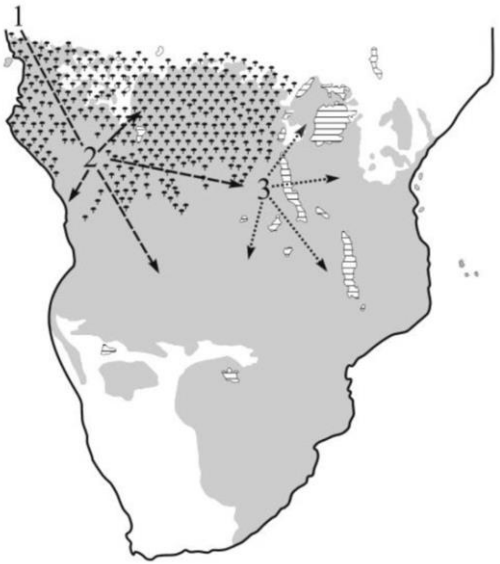

a. East out of West

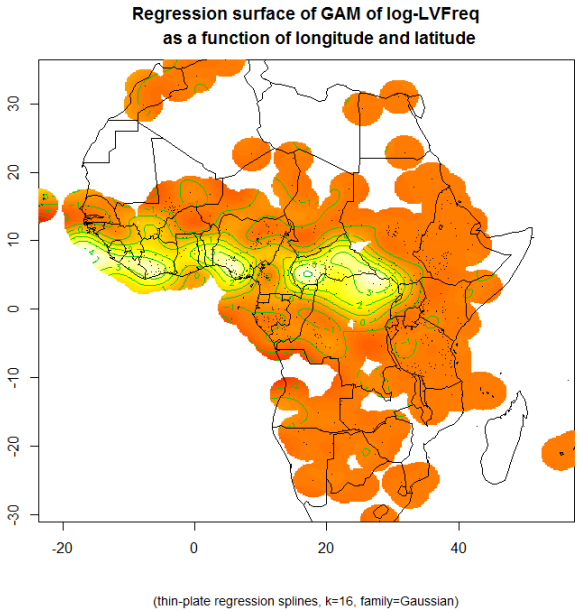


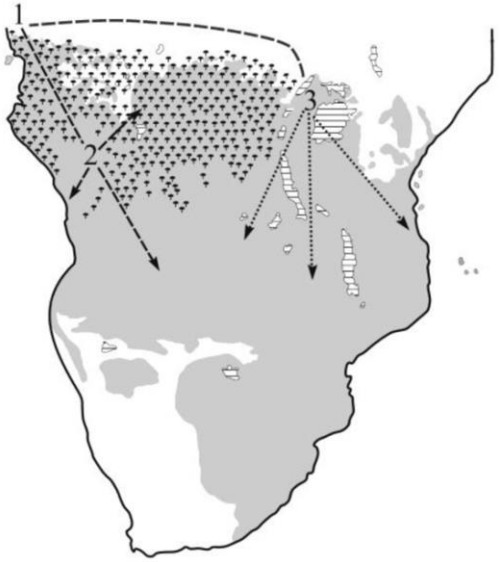

b. East separate from West

(adapted from Pakendorf et al. 2011 :8).

# HISTORICAL IMPLICATIONS: BANTU EXPANSION



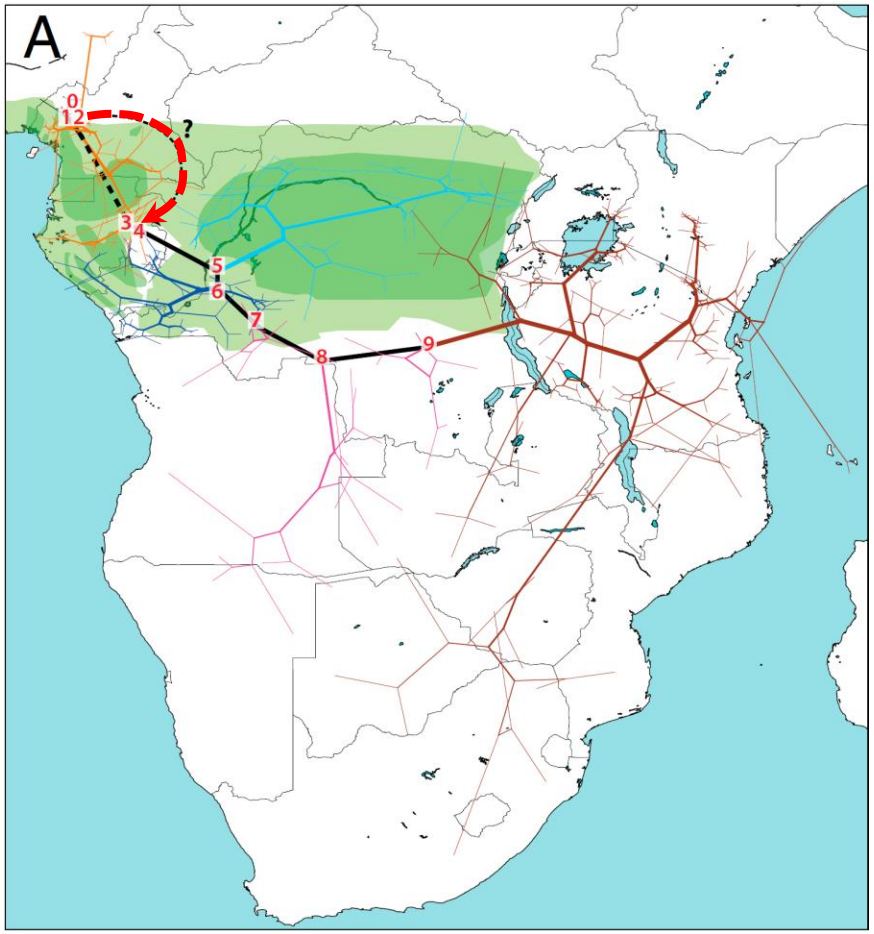




a. East out of West

b. East separate from West

- Our model 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: BANTU EXPANSION



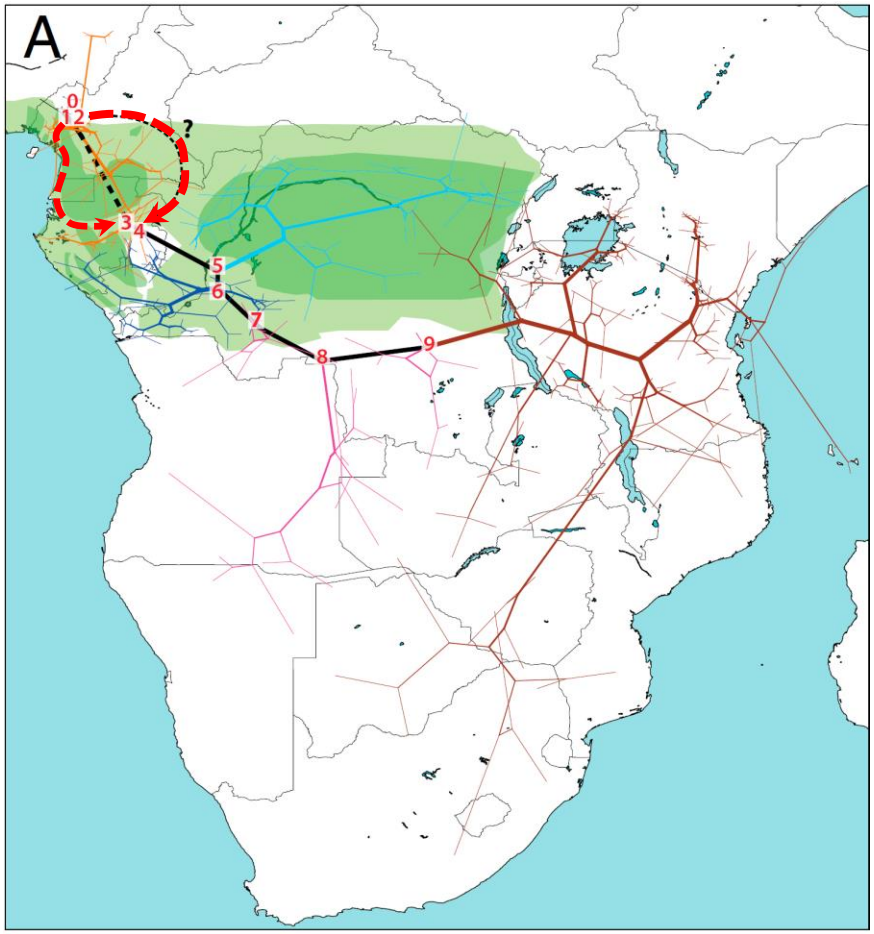

Bantu migration route reconstructed by Grollemund et al. (2015) on consensus tree by using geographical locations of contemporary languages and connecting ancestral locations by straight lines (true route will differ).

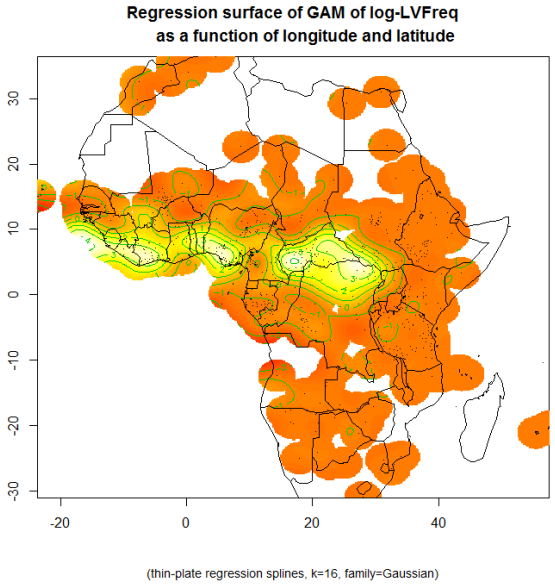
Numbered positions correspond to major diversification nodes on the consensus tree.

Curved dashed line indicates suggested migration route through savannah corridors (**Sangha River Interval**)

Lighter green shading corresponds to the delimitation of the rainforest at 5.000 B.P.; the darker green corresponds to the delimitation of the rainforest at 2.500 B.P.

# HISTORICAL IMPLICATIONS: BANTU EXPANSION




- Our model suggests that the migration between nodes 2 and 3 is more likely to have happened through a **coastal route** rather than the Sangha River Interval.