Exploring typological diversity and its areal and genealogical basis in South America

Emily Clem and Lev Michael
University of California, Berkeley

eclem@berkeley.edu, levmichael@berkeley.edu

Arealphonologie
5 July 2016

Introduction

- How can we grasp large scale patterns in areal datasets?
- How can we identify areality?
- How can we identify typological tendencies?

Research questions

- How is the phonological diversity of South America structured?
  - What are the major typological parameters of differentiation?
- What areal patterns are detectable?
- What genetic patterns do we find?
- What are the major types of phonological inventories and what are "extreme" languages?

A tool: PCA

- One promising tool to identify large-scale patterns is PCA (principal component analysis)
- An advantage to PCA is that it does not rely on the analysts' presuppositions about the data
- This tool allows us to identify broad patterns in datasets in a less biased manner

Preview of results

- We will present recurring evidence for a strong areal signal in the Andean and Circum-Andean region, separating it from Amazonia
- We will also show evidence for a smaller linguistic area in Northwest Amazonia
- We will argue that languages in South America differ as to whether the locus of phonological contrast is in their consonant system or vowel system
- We will demonstrate that contrasts in nasality and length on vowels are both significant dimensions of differentiation for inventories in South America

Roadmap

1. Introduction
2. Data and methods
3. Profiles of PCs 1 through 5
4. Phonological "types" in South America
5. Conclusion

The data: SAPhon

- SAPhon (South American Phonological Inventory Database) is an online database of phonological inventories of languages of South America (Michael et al., 2016)
  - `http://linguistics.berkeley.edu/~saphon/en/`
- It contains phonological inventories for 363 languages
  - 56% of the total number of languages and dialects in South America identified on Glottolog (Hammarström et al., 2015)
  - 104 more than the number of languages with ISO codes in South America
- The data is compiled from published sources and unpublished field notes

Why SAPhon?

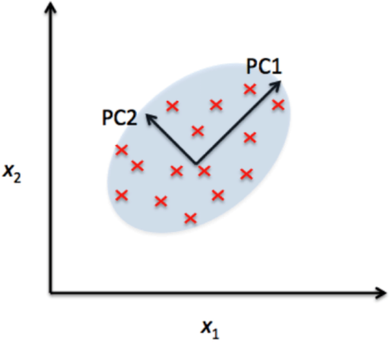
- Why construct datasets like this?
    - It provides a systematic and comprehensive exploration of phenomena
    - It allows researchers to avoid cherry-picking
    - It provides a starting point for quantitative analysis to avoid eye-balling
- Areally focused datasets like this would be useful to have for other regions of the world, and SAPhon is one model

A Question

How does one identify and comprehend patterns in large datasets?

The method: PCA

- PCA is a statistical method useful for dimension reduction
- It is a transformation of data along a new set of axes
- The axes along which there is the most variation in the data form the principal components (PC1, PC2, PC3...)



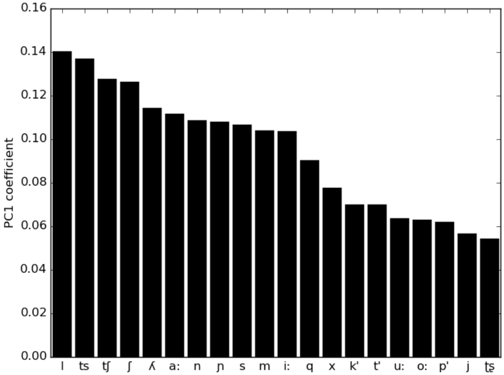
Raschka, S. Retrieved from www.quora.com

The method: PCA

- PCA is a statistical method useful for dimension reduction
- It is a transformation of data along a new set of axes
- The axes along which there is the most variation in the data form the principal components (PC1, PC2, PC3...)
- By examining only the PCs that explain the largest amount of variance, we can reduce the number of dimensions in our dataset
  - We began with 301-dimensional dataset (the total number of unique phonemes in our data)
  - We will be examining the first 5 PCs, which account for over 35% of the variance in the data
- Because PCA is linguistically-naive it does not pick out linguistic features as such but rather identifies segments which covary in their presence or absence in this dataset

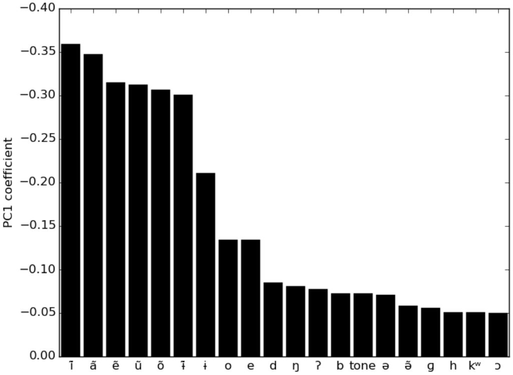
PC1: positive coefficients

- PC1 explains 14.2% of the variance in the data
- The segments with the largest positive coefficients include affricates, palatals, and laterals



PC1: negative coefficients

- The segments with the largest negative coefficients include nasal vowels, ɨ, and mid vowels

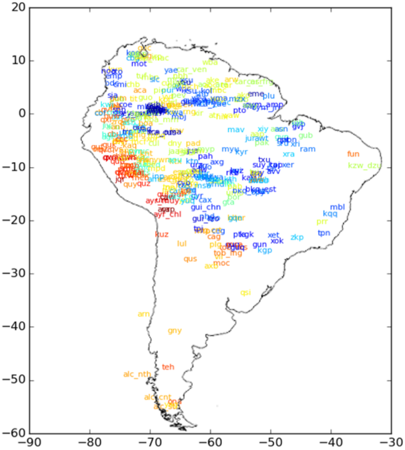


PC1: summary of coefficients

- Negative coefficients are slightly larger than positive coefficients
- Nasal vowels are the most strongly negative segments
- ɨ is also strongly negative
- Alveolar and palatal laterals, affricates, fricatives, and nasals are the most strongly positive segments
- The mid vowels e and o are moderately negative

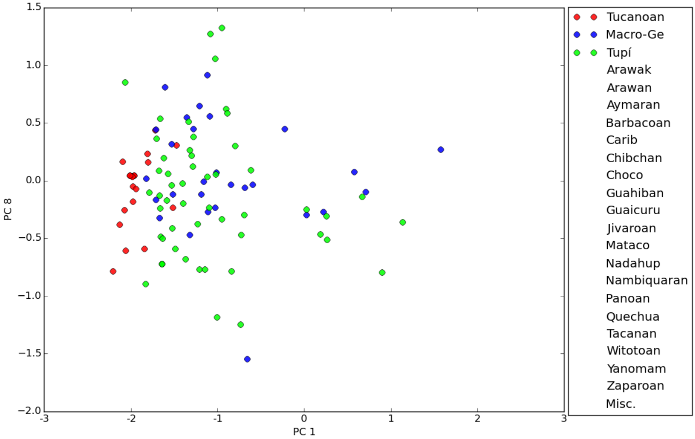
PC1: areal signal

- PC1 yields a strong positive signal in the Andean and Circum-Andean region (Michael et al., 2014)



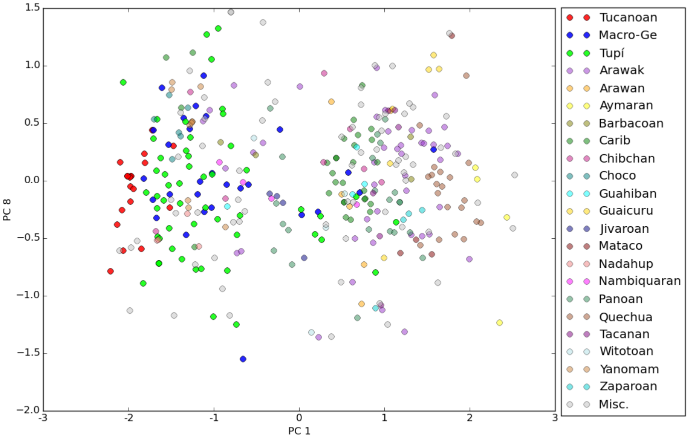
PC1: genetic signal

- PC1 shows a negative genetic signal including the Tucanoan, Macro-Ge, and Tupí families



PC1: genetic signal

- PC1 shows a negative genetic signal including the Tucanoan, Macro-Ge, and Tupí families

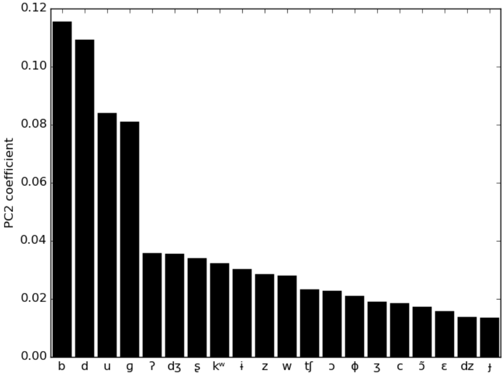


PC1: summary

- Positive segments: alveolar and palatal laterals, affricates, fricatives, and nasals
- Negative segments: nasal vowels, ɨ
- Positive component yields a strong areal signal in the Andes and Circum-Andean area, including Patagonia
- Negative component shows a genetic signal from Tucanoan, Macro-Ge, and Tupí

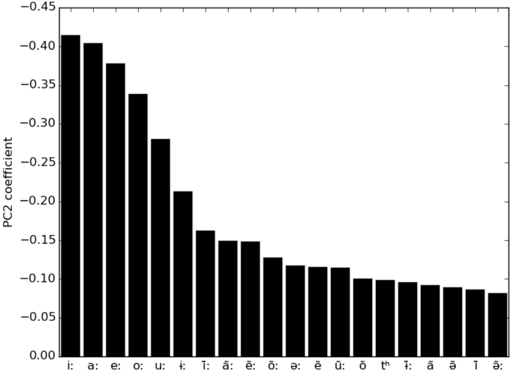
PC2: positive coefficients

- PC2 explains 7.6% of the variance in the data
- The segments with the largest positive coefficients include the voiced stops



PC2: negative coefficients

- All of the segments with the largest negative coefficients are long vowels

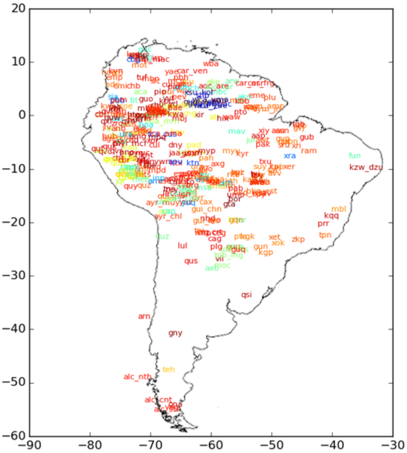


PC2: summary of coefficients

- Negative coefficients are larger than positive coefficients
- All strongly negative segments are long vowels
- The most strongly positive segments are voiced stops

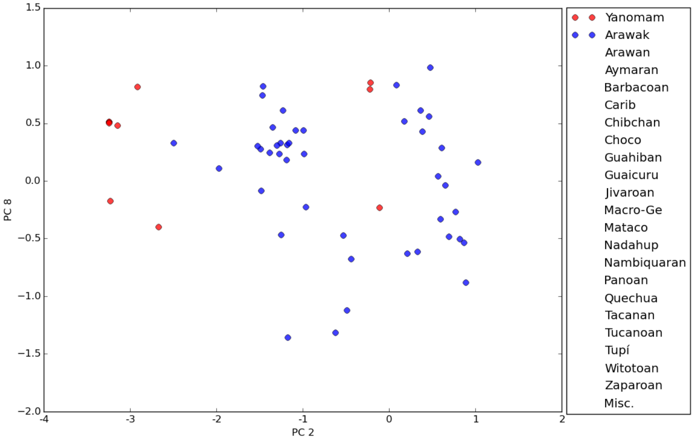
PC2: areal signal

- PC2 does not display a strong areal signal distinct from genetic relationships



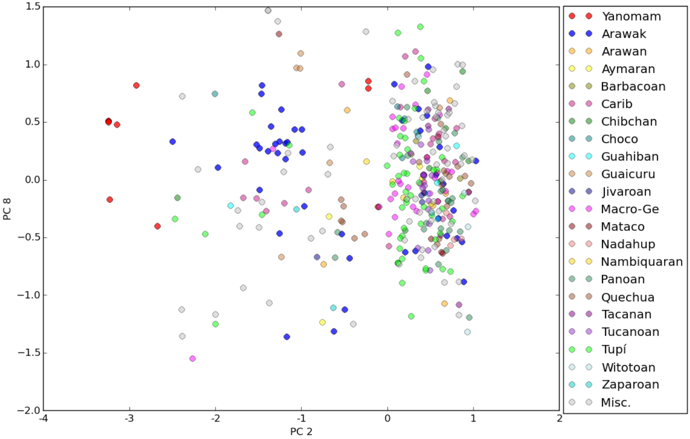
PC2: genetic signal

- PC2 shows a strongly negative genetic signal from the Yanomam family and a moderate signal from Arawak



PC2: genetic signal

- PC2 shows a strongly negative genetic signal from the Yanomam family and a moderate signal from Arawak

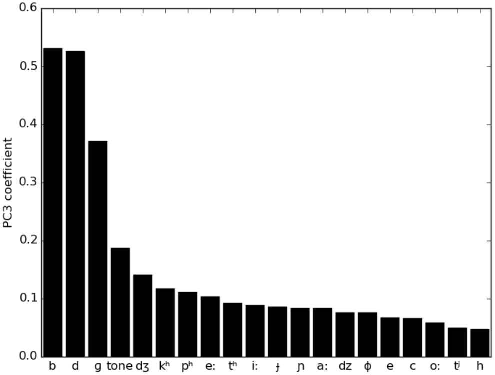


PC2: summary

- Positive segments: voiced stops
- Negative segments: long vowels
- Negative component shows a strong genetic signal associated with Yanomam, and other families also cluster together
- We see a large negative dispersion with the most strongly negative languages displaying vowel length contrasts for many vowels

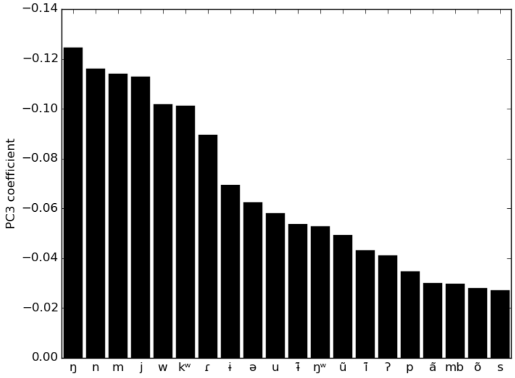
PC3: positive coefficients

- PC3 explains 5.4% of the variance in the data
- The segments with the largest positive coefficients include the voiced stops, tone, and aspirated stops



PC3: negative coefficients

- The segments with the largest negative coefficients include the nasal stops and approximants

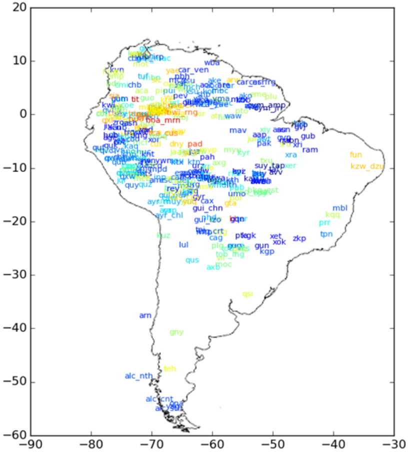


PC3: summary of coefficients

- Positive coefficients are much larger than negative coefficients
- The most strongly positive segments are voiced stops and aspirated stops
- Tone is strongly positive
- The most strongly negative segments are nasal stops
- Strongly positive languages are those which display nasal harmony

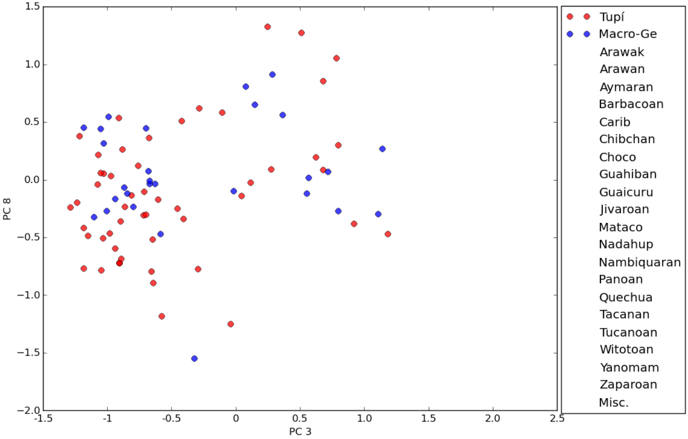
PC3: areal signal

- PC3 shows a strong positive signal in Northwest Amazonia



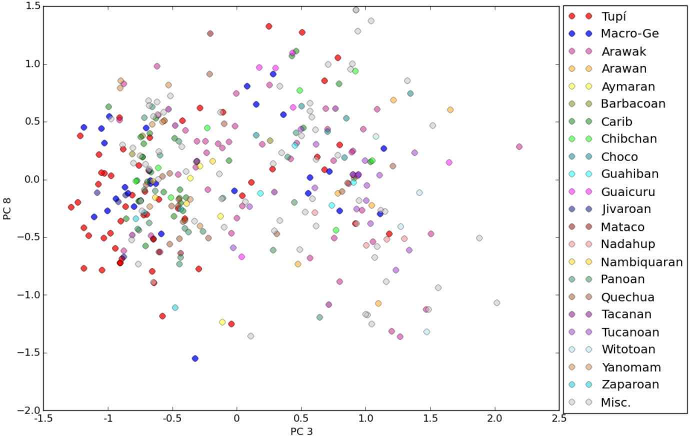
PC3: genetic signal

- PC3 shows a strong negative genetic signal associated with the Tupí and Macro-Ge families



PC3: genetic signal

- PC3 shows a strong negative genetic signal associated with the Tupí and Macro-Ge families

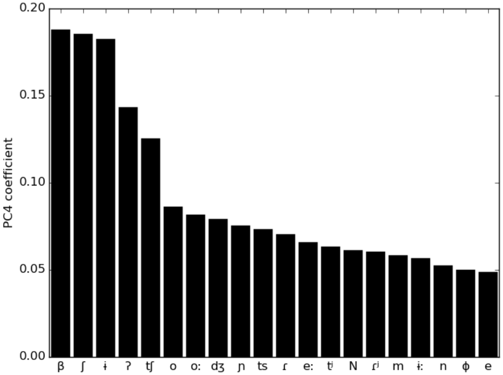


PC3: summary

- Positive segments: voiced stops, tone
- Negative segments: nasal stops, approximants
- Positive component yields a strong areal signal in Northwest Amazonia
- This positive signal reflects languages that have processes of nasal harmony rather than underlying nasal stops
- Negative component shows a genetic signal from Tupí and Macro-Ge

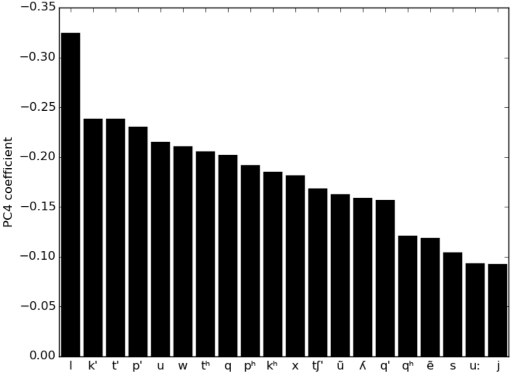
PC4: positive coefficients

- PC4 explains 4.6% of the variance in the data
- The segments with the largest positive coefficients include β, ɨ, palatals, and mid vowels



PC4: negative coefficients

- The segments with the largest negative coefficients are ejectives, aspirated stops, laterals, and uvulars

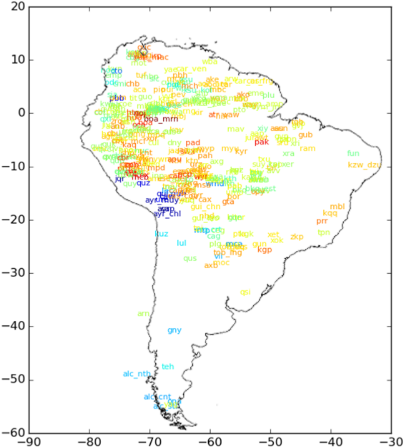


PC4: summary of coefficients

- Negative coefficients are larger than positive coefficients
- The lateral l is the most strongly negative segment
- Ejectives, aspirated stops, and uvulars are strongly negative
- β, ɨ, and palatals are the most strongly positive segments

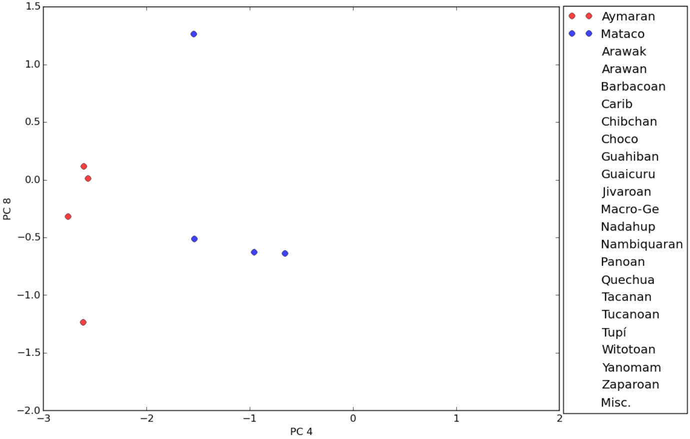
PC4: areal signal

- PC4 shows a strong negative signal in the Southern Andean and Circum-Andean region, including Patagonia



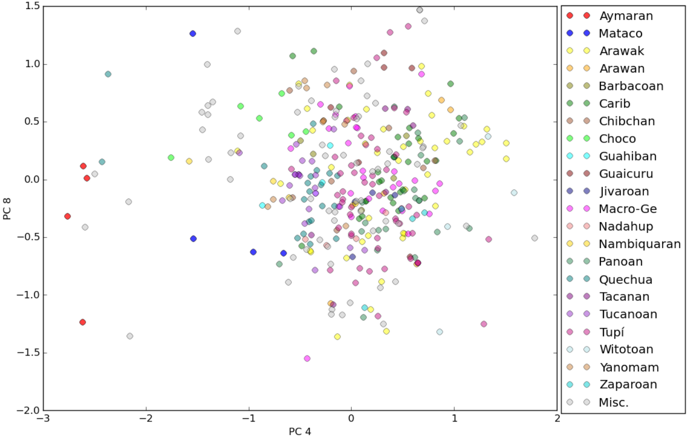
PC4: genetic signal

- PC4 does not show a strong genetic signal distinct from an areal signal



PC4: genetic signal

- PC4 does not show a strong genetic signal distinct from an areal signal

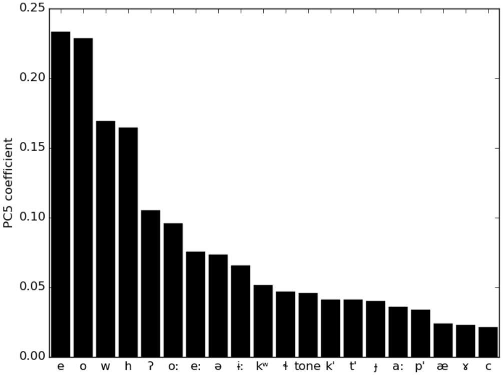


PC4: summary

- Positive segments: β, ɨ, palatals
- Negative segments: ejectives, aspirated stops, laterals, uvulars
- Negative component shows further support for a strong areal signal in Southern Andean and Circum-Andean region

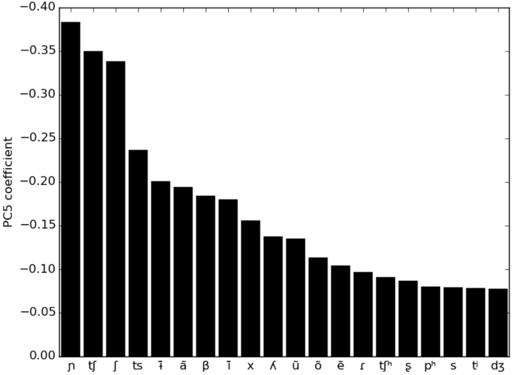
PC5: positive coefficients

- PC5 explains 3.9% of the variance in the data
- The segments with the largest positive coefficients are mid vowels, w, and glottals



PC5: negative coefficients

- The segments with the largest negative cofficients are palatals, affricates, and ɨ

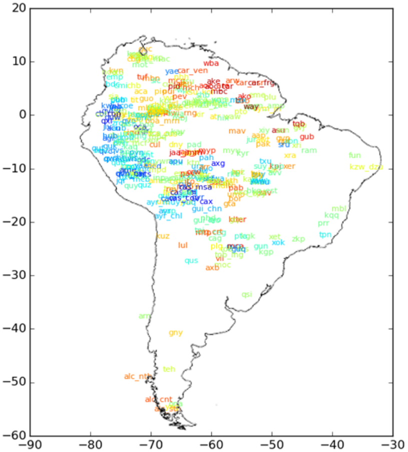


PC5: summary of coefficients

- Negative coefficients are slightly larger than positive coefficients
- Palatals and affricates are the most strongly negative segments
- Mid vowels are the most strongly positive segments
- ɨ is strongly negative

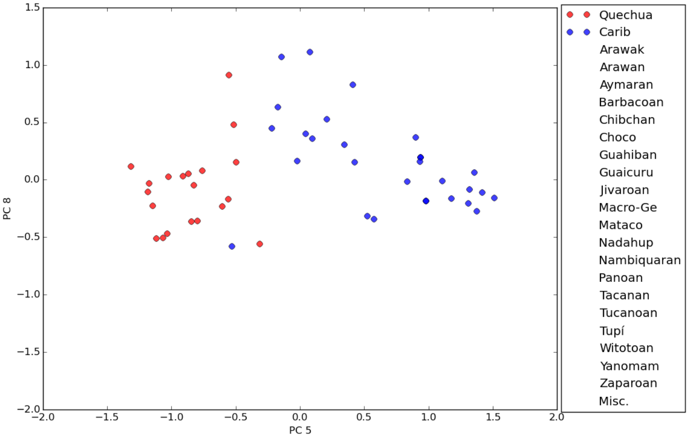
PC5: areal signal

- PC5 shows a negative signal in the Central-Northern Andean region and a positive signal in the east



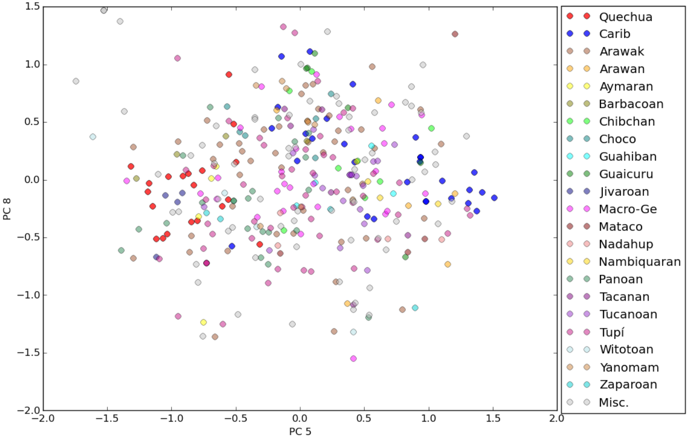
PC5: genetic signal

- PC5 shows a west/east divide illustrated by the Quechua and Carib families and neighboring languages



PC5: genetic signal

- PC5 shows a west/east divide illustrated by the Quechua and Carib families and neighboring languages



PC5: summary

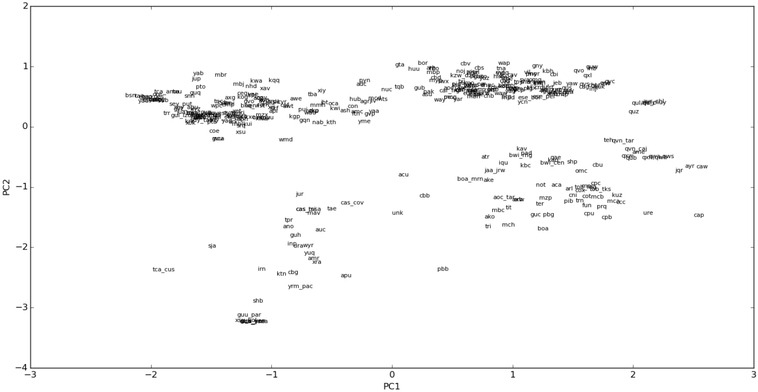
- Positive segments: mid vowels
- Negative segments: palatals, affricates, ɨ
- There is a general divide, both genetic and areal, between a negative signal in the west and a positive signal in the east
- Overall, the signal is becoming weak by this point

Identifying phonological "types"

- Having identified the major dimensions of variation in South America, we can develop a "continental typology" of phonological inventories
- Major types can be identified by examining how inventories cluster with respect to PC1 and PC2
- Sampling languages in each cluster allows us to identify their major features

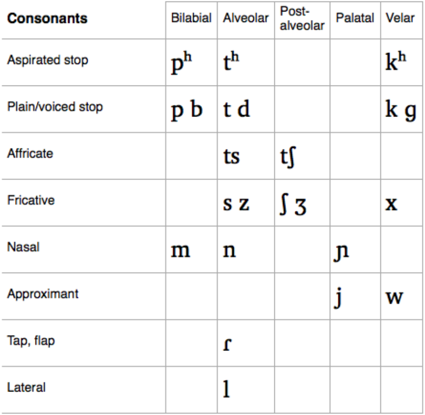
Phonological types in South America

- What major types can we extrapolate from PCA?
- What are the clusters and what are their general profiles?

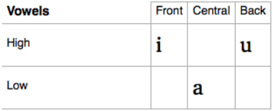


Quadrant 1: +PC1, +PC2

- Features: large consonant inventories (laterals, affricates, voiced stops) and small vowel inventories
- Example language: Salasca Quechua (Quechua)



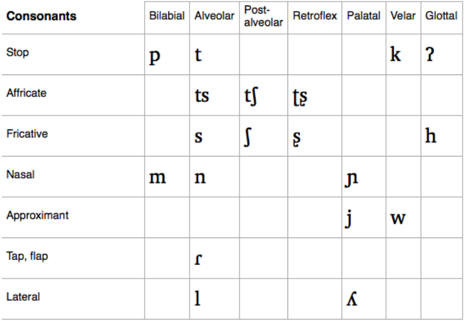
Consonants	Bilabial	Alveolar	Post-alveolar	Palatal	Velar
Aspirated stop	p^h	t^h			k^h
Plain/voiced stop	p b	t d			k g
Affricate		ts	tʃ		
Fricative		s z	ʃ ʒ		x
Nasal	m	n		ɲ	
Approximant				j	w
Tap, flap		ɾ			
Lateral		l			



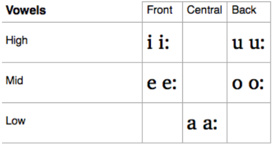
Vowels	Front	Central	Back
High	i		u
Low		a	

Quadrant 2: +PC1, -PC2

- Features: large consonant inventories (laterals, affricates), moderate vowel inventories (long vowels)
- Example language: Chamicuro (Arawak)



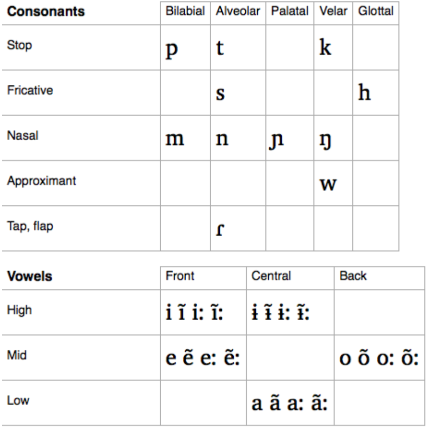
Consonants	Bilabial	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Glottal
Stop	p	t				k	ʔ
Affricate		ts	tʃ	ʈʂ			
Fricative		s	ʃ	ʂ			h
Nasal	m	n			ɲ		
Approximant					j	w	
Tap, flap		ɾ					
Lateral		l			ʎ		



Vowels	Front	Central	Back
High	i iː		u uː
Mid	e eː		o oː
Low		a aː	

Quadrant 3: -PC1, -PC2

- Features: small consonant inventories, large vowel inventories (nasal and long vowels)
- Example language: Karitiâna (Tupí)

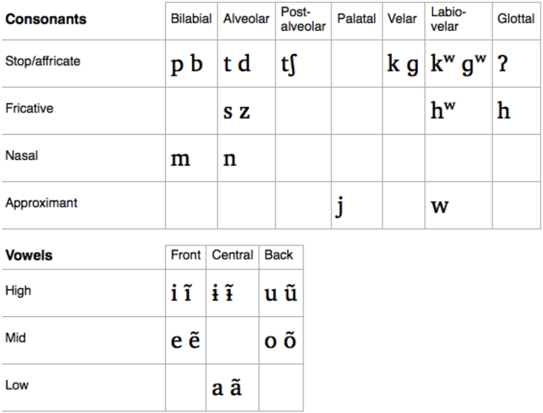


Consonants	Bilabial	Alveolar	Palatal	Velar	Glottal
Stop	p	t		k	
Fricative		s			h
Nasal	m	n	ɲ	ŋ	
Approximant				w	
Tap, flap		ɾ			

Vowels	Front	Central	Back
High	i ĩ iː ĩː	ɨ ɨ̃ ɨː ɨ̃ː	
Mid	e ẽ eː ẽː		o õ oː õː
Low		a ã aː ãː	

Quadrant 4: -PC1, +PC2

- Features: moderate consonant inventories (voiced stops), moderate vowel inventories (nasal vowels)
- Example language: Siona (Tucanoan)



Consonants	Bilabial	Alveolar	Post-alveolar	Palatal	Velar	Labio-velar	Glottal
Stop/affricate	p b	t d	tʃ		k g	kʷ gʷ	ʔ
Fricative		s z				hʷ	h
Nasal	m	n					
Approximant				j		w	

Vowels	Front	Central	Back
High	i ĩ	ɨ ɨ̃	u ũ
Mid	e ẽ		o õ
Low		a ã	

Phonological types, areality, and families

- Many languages with a small number of vowel contrasts (Quadrant 1) comprise the linguistic area of the Circum-Andean region
- Some families cluster together within one of the predominant types
  - Yanomam is very tightly clustered in Quadrant 3

Conclusion

- In South America, whether languages make a large number of contrasts in their vowels is very significant in "typing" languages
  - Nasal vs. oral is one of the most significant dimensions of variation
  - Length contrasts are also an important parameter of differentiation
- This continental profile of South America will look very different from the profiles of other continents
- This work provides a starting point for more quantitatively rigorous analyses of areality, such as NBC (naive Bayes classifier)

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2015). Glottolog 2.4.

Michael, L., Chang, W., and Stark, T. (2014). Exploring phonological areality in the cirum-Andean region using a naive Bayes classifier. *Language Dynamics and Change*, 4(1):27–86.

Michael, L., Stark, T., Clem, E., and Chang, W. (2016). *South American Phonological Inventory Database*. University of California, Berkeley.