方言群の時空間解析にむけて: フィジー語を例に

京都大学
村脇 有吾

2019年2月21日　ヤポネシアゲノム言語班 2018年度第2回研究集会@石垣市

自己紹介: 村脇 有吾

- 京都大学 大学院情報学研究科
  知能情報学専攻 助教
  工学部電気電子工学科 兼担 (旧有線通信工学講座)
- 表の仕事は自然言語処理
  - Deep Learning, AI💰💰💰
- 裏テーマ: 歴史言語学、言語類型論等の問題への統計的取り組み (2014年-)
- 計算機科学の他の多くの分野と同じく国際会議指向

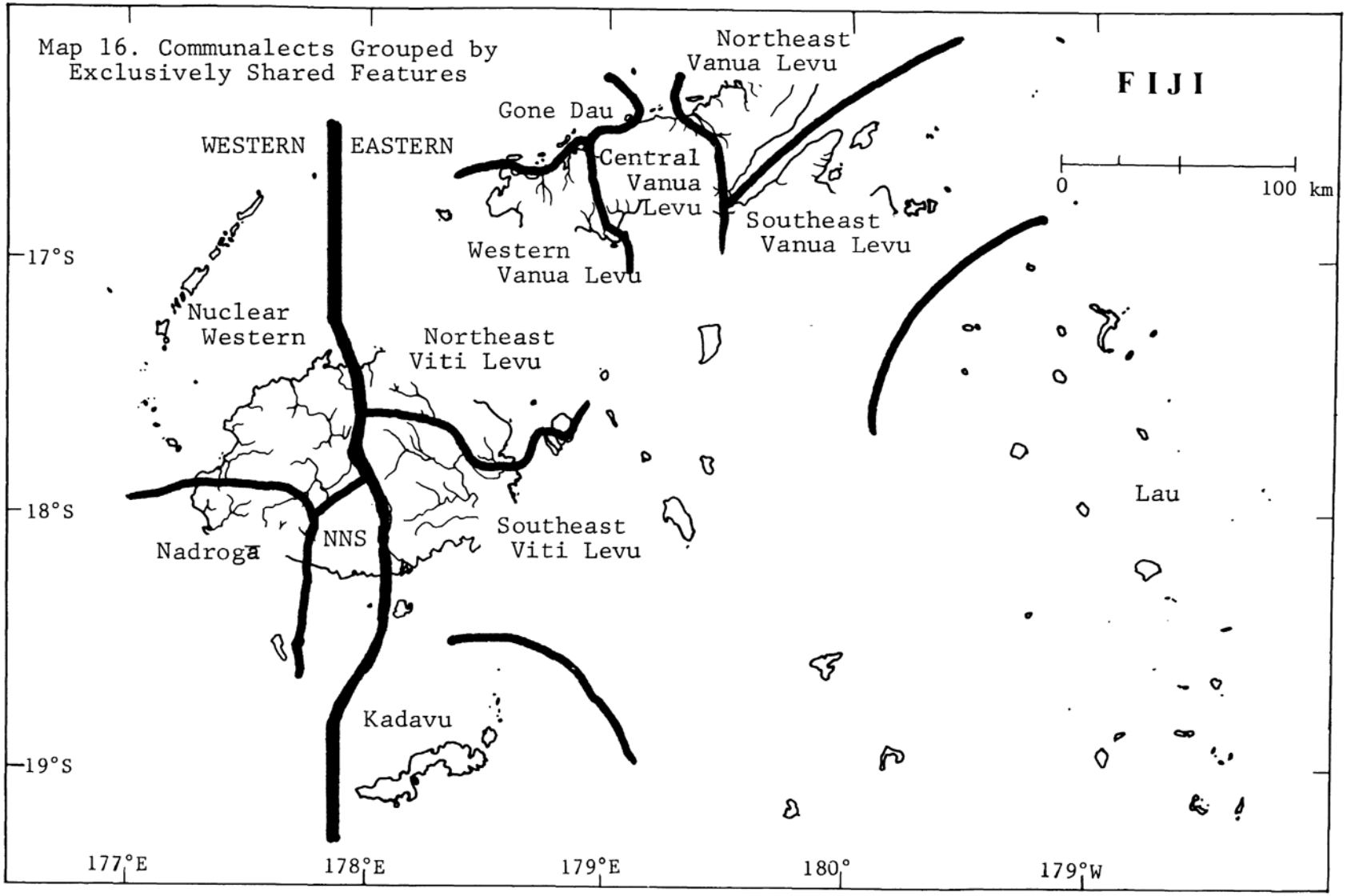
科研費国際共同研究強化 (B) (2018.10-2024.03)
時空間を融合する: GISと数理モデルを用いた新たな言語変化へのアプローチ

- 代表: 菊澤 律子 (民博)
- 分担者:
  - 村脇 有吾 (京大) 数理的解析
  - 吉岡 乾 (民博) 多言語展開?
  - 持橋 大地 (統数研) 数理的解析
- 共同研究者:
  - John Lowry (Massay) GIS専門家
  - Paul A. Geraghty (USP) 言語学者 (フィールド/歴史)
  - Apolonia Tamata (TTFB) 言語学者、母語話者

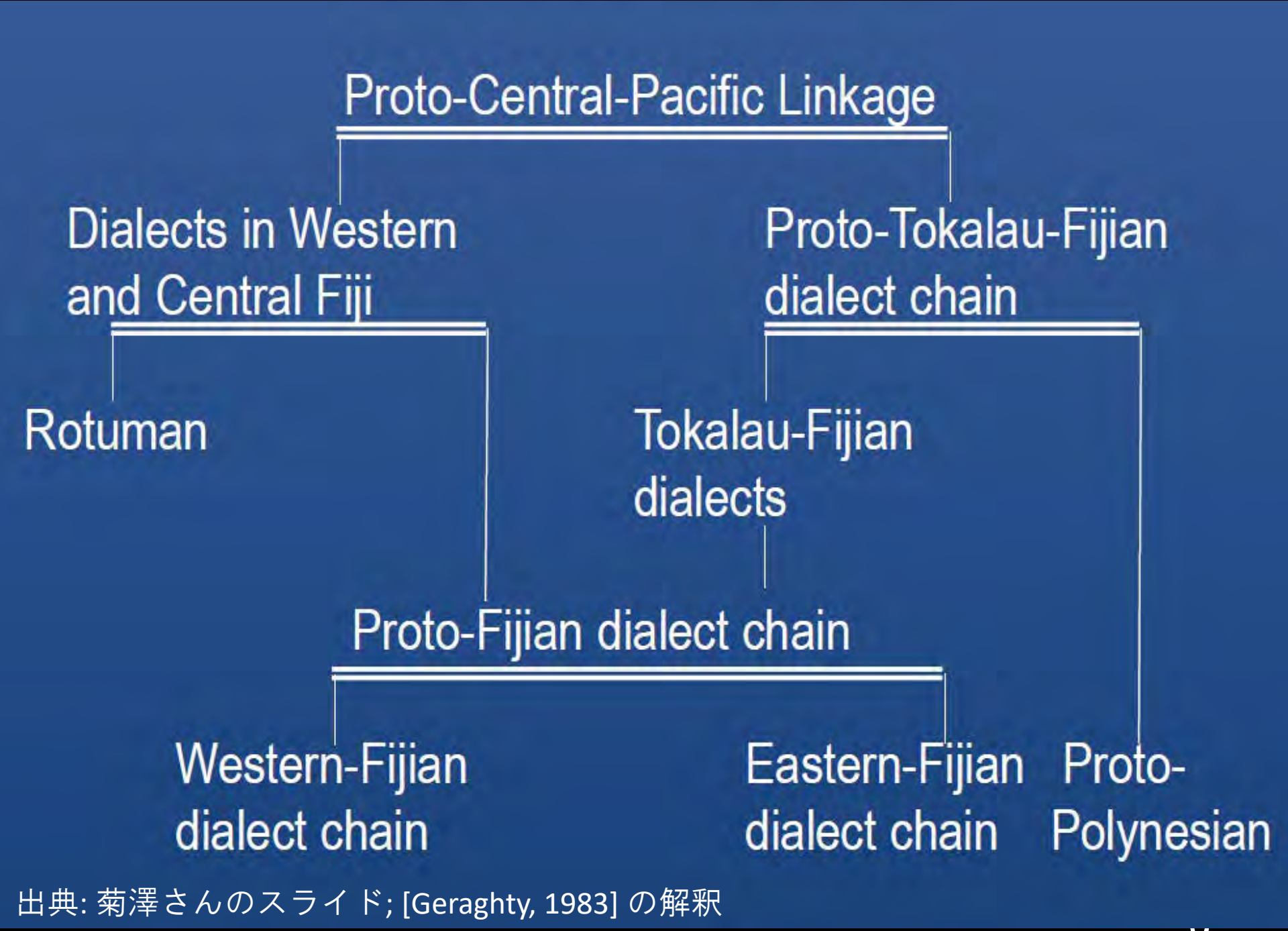
本研究では、フィジー語諸方言の大規模データ（約300方言、約5800語）をテストケースとし、言語の時間的な変化と空間的な伝播との関係を明らかにする。さらに、その結果と非言語情報（地形や地点間の移動距離など）との相関を検証し、言語変化をヒトの活動と関連付けて包括的にとらえることを目的とする。

$L = 106$

※座標データとの対応づけが不完全

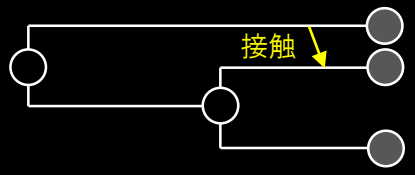


[Geraghty, 1983]



出典: 菊澤さんのスライド; [Geraghty, 1983] の解釈

系統樹モデル



- 分岐の連鎖と分岐後の独立性を仮定
- 仮定に従って生成されたと期待できないデータに適用してもgarbage in, garbage out
  - 単なる階層クラスタリングと割り切るならともかく
- ネットワークは木の重ね合わせに過ぎず、接触を直接モデル化しているわけではない
- 先行研究は仮定に従いそうなサブセットにデータを絞り込む [ローレンス, 2006][Pellard, 2009][五十嵐, 2016]
- データ全体が持つ性質を捉えたい

フィジー語諸方言基礎語彙データ

	conj (or)	1sg subj (I)	post. (up)	…
Standard Fijian	1	1	1	..
Nadi	1	2	1	…
Nalea	2	2	2	…
Naloto Wainibuka	1	1	1	…
Dakuinuku	1	1	1	…
…	…	…	…	…

- LingPyを用いて文字列比較に基づき同源語を自動認定
  - 精度面で人間に劣るが大規模化可能
- 後述の提案モデルが要求するのは特徴の離散化だけ
  - 音韻・形態統語的特徴も扱える

文字列比較に基づく同源語自動認定

[List+, JoLE 2018]

Input

numeric identifier, language name, concept label, value in source, phonetic transcription, space-segmented transc., variants of word, if given, cognate identifier

ID	DOCULECT	CONCEPT	VALUE	IPA	TOKENS	VARIANTS	COGID
188	Emae	Eight	βaru	βaru	β a r u		750
447	Rennell_Bellona	Eight	baŋgu	baŋgu	b a ⁿg u		750
703	Tuvalu	Eight	valu	valu	v a l u		750
927	Sikaiana	Eight	valu	valu	v a l u		750
1135	Penrhyn	Eight	varu	varu	v a r u		750
6114	Kapingamarangi	Eight	waru,walu	waru	w a r u	walu	750
6115	Kapingamarangi	Eight	waru,walu	walu	w a l u	waru	750

Output

			Ra'.	Haw.	Man.	Mao.	Rap.		Ra'.	Haw.	Man.	Mao.	Rap.
Ra'ivavae	ʔaɢapoʔa	1	0.00	0.75	0.88	0.88	0.75	1	0.00	0.49	0.59	0.83	0.62
Hawaiian	ʔaːʔiː	1	0.75	0.00	1.00	1.00	1.00	2	0.49	0.00	0.67	0.73	0.79
Mangareva	kaki	2	0.88	1.00	0.00	0.75	0.75	3	0.59	0.67	0.00	0.73	0.39
Maori	ua	3	0.88	1.00	0.75	0.00	0.67	4	0.83	0.73	0.73	0.00	0.47
Rapanui	ŋao	3	0.75	1.00	0.75	0.67	0.00	3	0.62	0.79	0.39	0.47	0.00

A *NED* B *SCA*

文字列比較に基づく同源語自動認定

[List+, JoLE 2018]

A Scoring Function

Determines how segments are compared with each other. Most generally represented as a symmetric matrix of transition probability scores. Scores between segments are usually given in logarithmic scale, with unexpected matches being smaller 0 and expected ones being greater 0. On the right, a short scoring matrix is shown, in which matches between vowels (a) and consonants (p, f, h) are not allowed and therefore assigned high negative values.

	p	f	h	a
p	10	6	2	-10
f	6	10	6	-10
h	2	6	10	-10
a	-10	-10	-10	10

B Sound Classes

To reduce the huge number of sounds in phonetic alignment analyses, sounds are clustered into classes, assuming that correspondences inside a class are more likely than outside a class. Scoring functions are defined for sound classes, and sound sequences are converted to sound-class sequences before aligning them. On the right, it is shown, how Austral "you" can be rendered in different sound-class alphabets.

IPA	ʔ	oː	ɢ	u	a
Dolgo	H	V	K	V	V
SCA	H	U	K	Y	A
ASJP	7	o	q	u	a

C Gap Function

Gaps are introduced in alignments if a given element cannot be matched with any other element. Gap penalties can be defined globally, by assigning the same penalty to any segment of a given sequence, or individually, based on the context in which the segment occurs. In phonetic alignment analyses, individual gap penalties can be derived from *prosodic profiles* which reflect the relative sonority of each segment in a sequence. On the right, it is shown, how individual gap penalties are derived for the Austral "you".

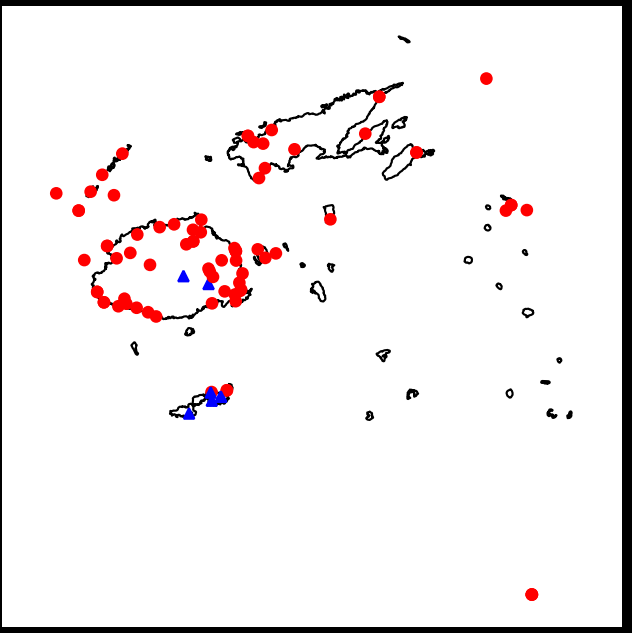
IPA	ʔ	oː	ɢ	u	a
Sonority	1	7	1	7	7
Prostring	#	V	C	V	>
Weights	2.0	1.5	1.75	1.3	0.8
Gap-Scores	-4	-3	-3.5	-2.6	-1.6

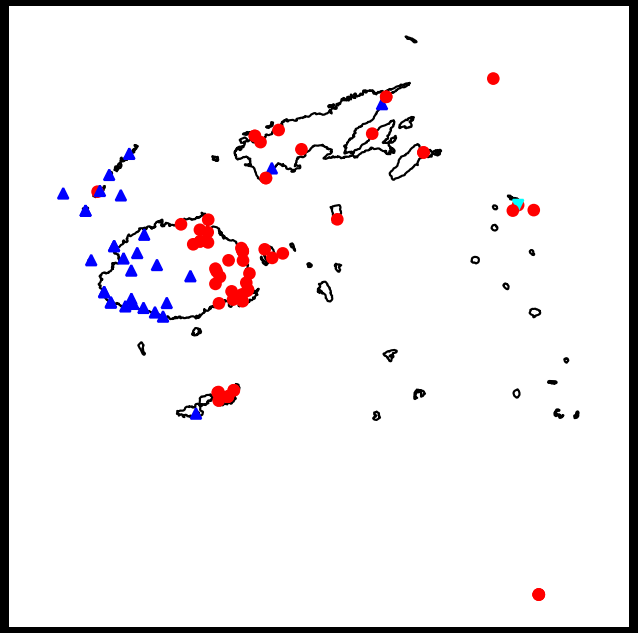
D Alignment Mode

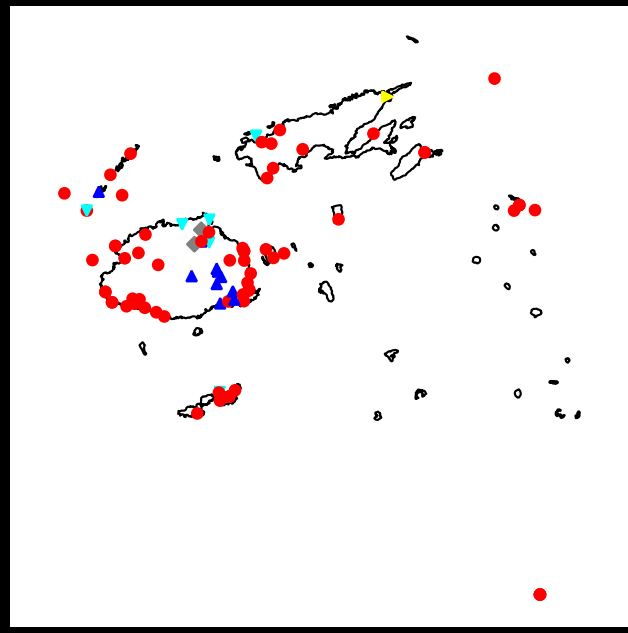
Alignment modes determine how sequences are compared. Global alignment compares the sequences entirely, assuming that they are indeed comparable in all their parts. Local alignment seeks to find the most similar subsequence of an alignment and deliberately strips off prefixes or suffixes. Semi-global alignment can only ignore prefixes or suffixes in one of the sequences. If the other sequence also has a suffix, it needs to strip it off. On the right, it is shown, how the different alignment modes account for the phonetic alignment of "you (dual)" in Ra'ivavae and Mangareva. While the local and the semi-global alignment fail to identify the regular sound correspondence of glottal stop and [k], the global alignment correctly matches words. This does not mean, however, that global alignment always yields the best solutions. Unaligned parts are shaded gray.

Global Alignment						
Ra'ivavae	ʔ	oː	ɢ	u	a	
Mangareva	k	oː	r	u	a	
Local Alignment						
Ra'ivavae	(ʔoː)	ɢ	-	-	u	a
Mangareva	-	k	oː	r	u	a
Semi-Global Alignment						
Ra'ivavae	ʔ	oː	ɢ	-	-	u	a
Mangareva	-	-	k	oː	r	u	a

各特徴の可視化





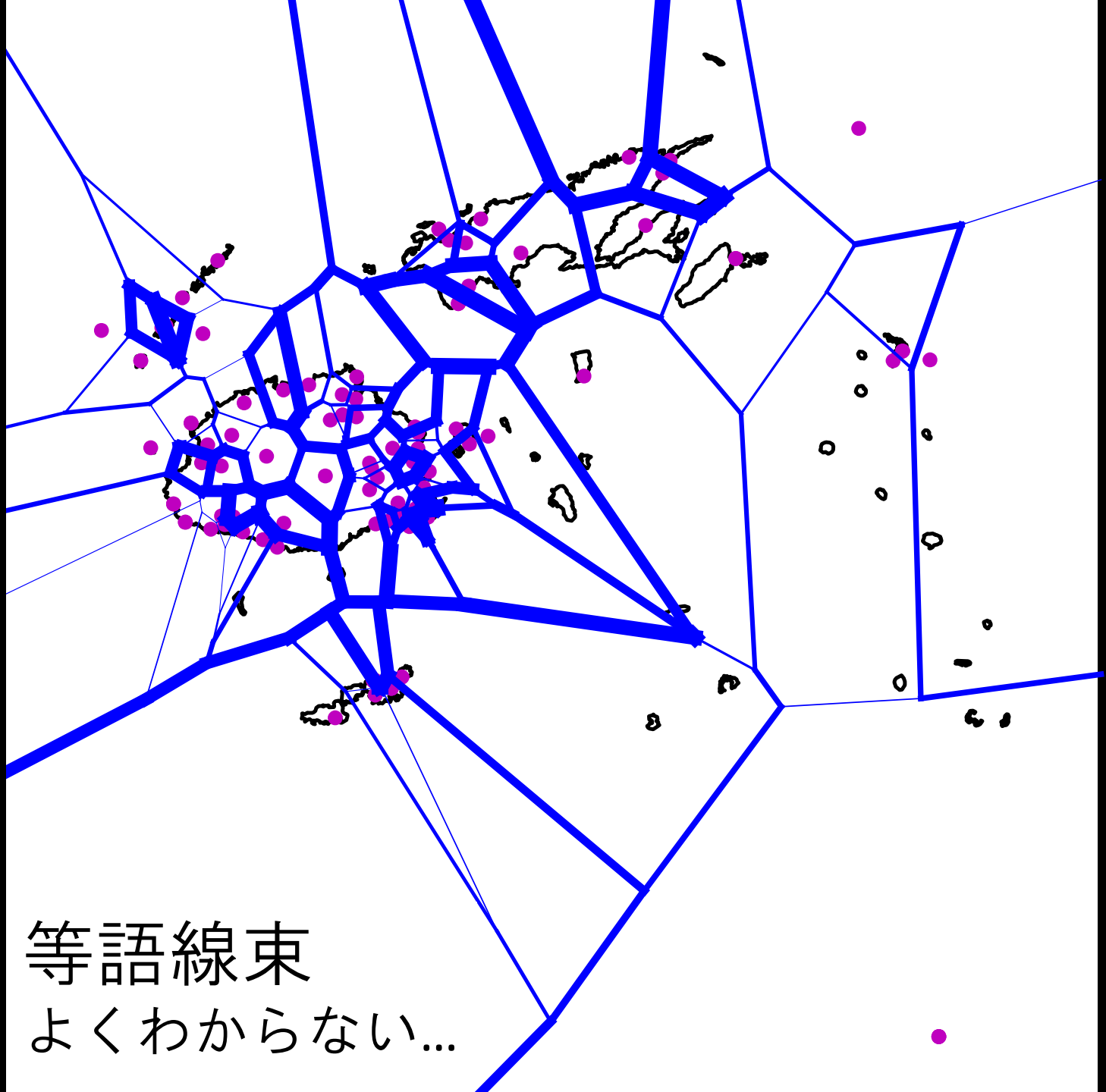


概念$N = 97$, 同源語グループ$M = 480$

※同源語グループが1個しかない概念は除外

データの全体構造を可視化したい

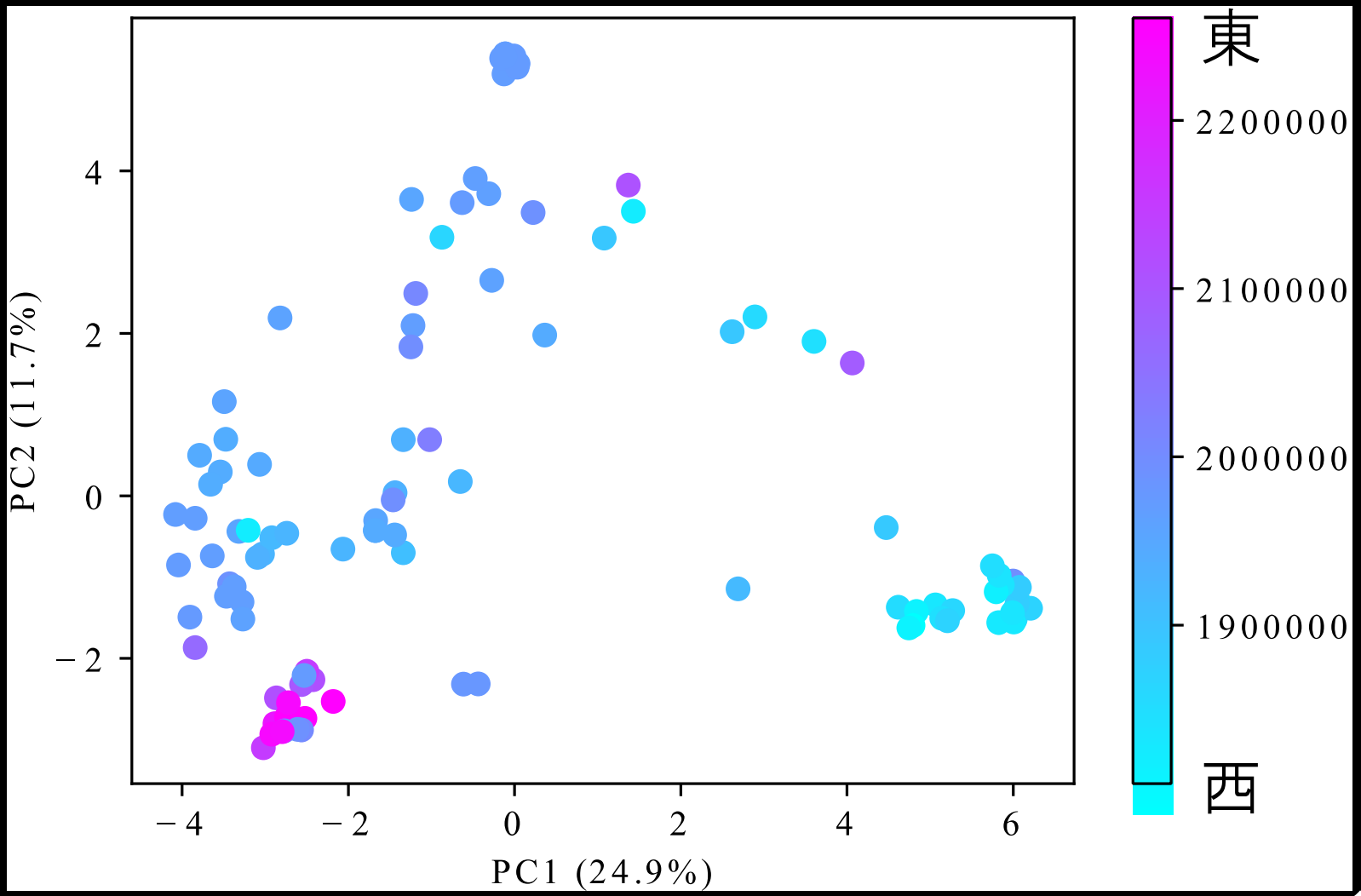
1. 等語線束
2. 主成分分析
3. ADMIXTURE解析
4. 提案手法: 潜在変数を用いた
   地理的分布の再構成



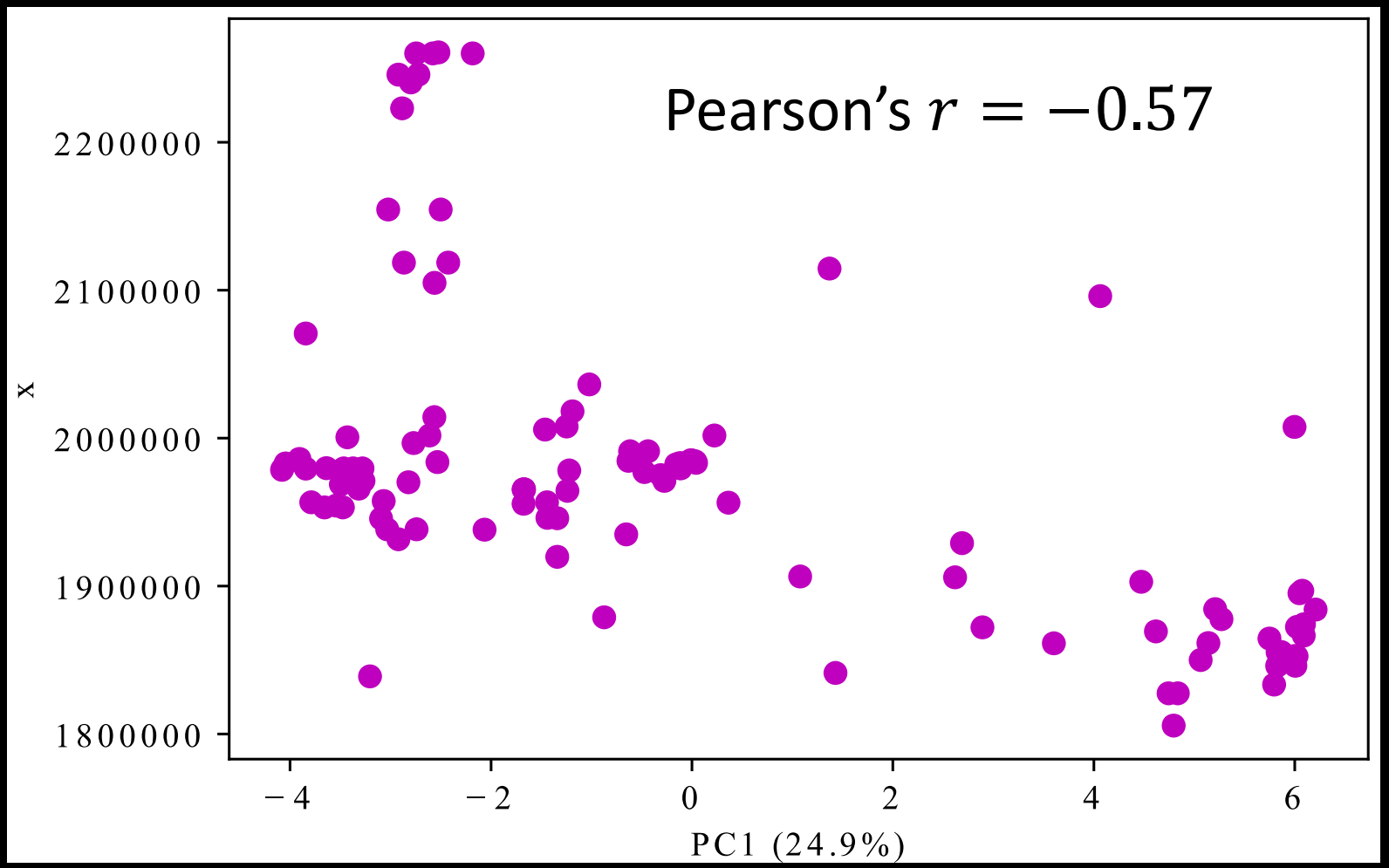
等語線束

よくわからない…

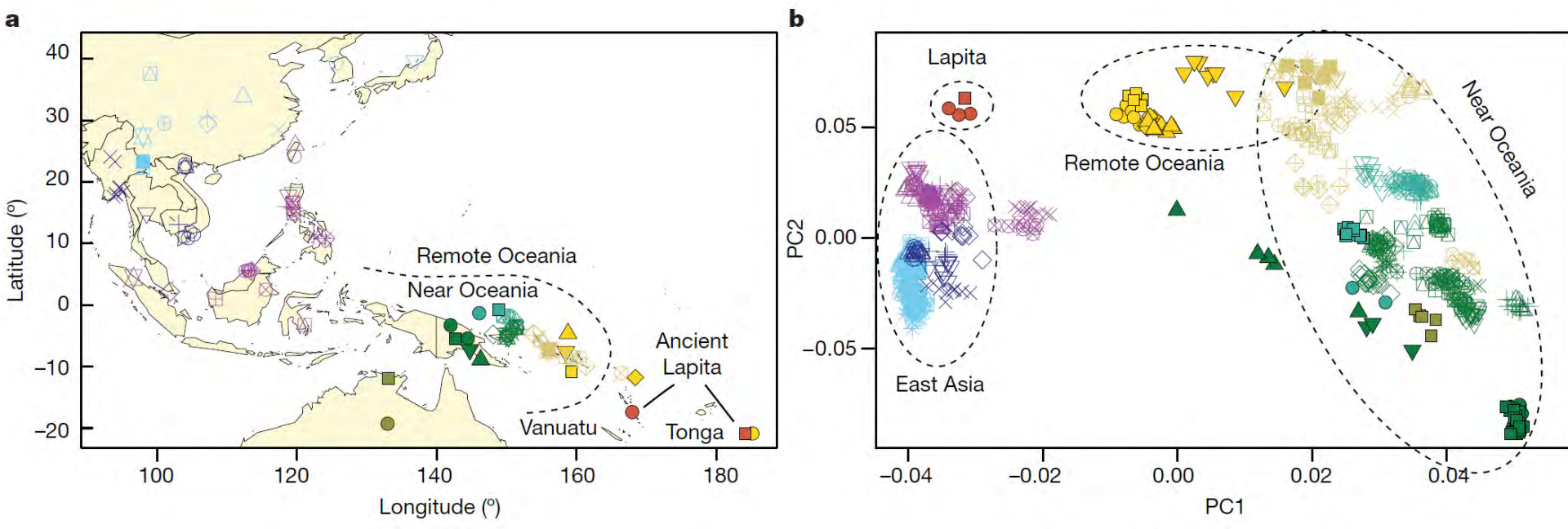
主成分分析は東西対立を可視化



主成分分析は東西対立を可視化



主成分分析は集団遺伝学でも多用される



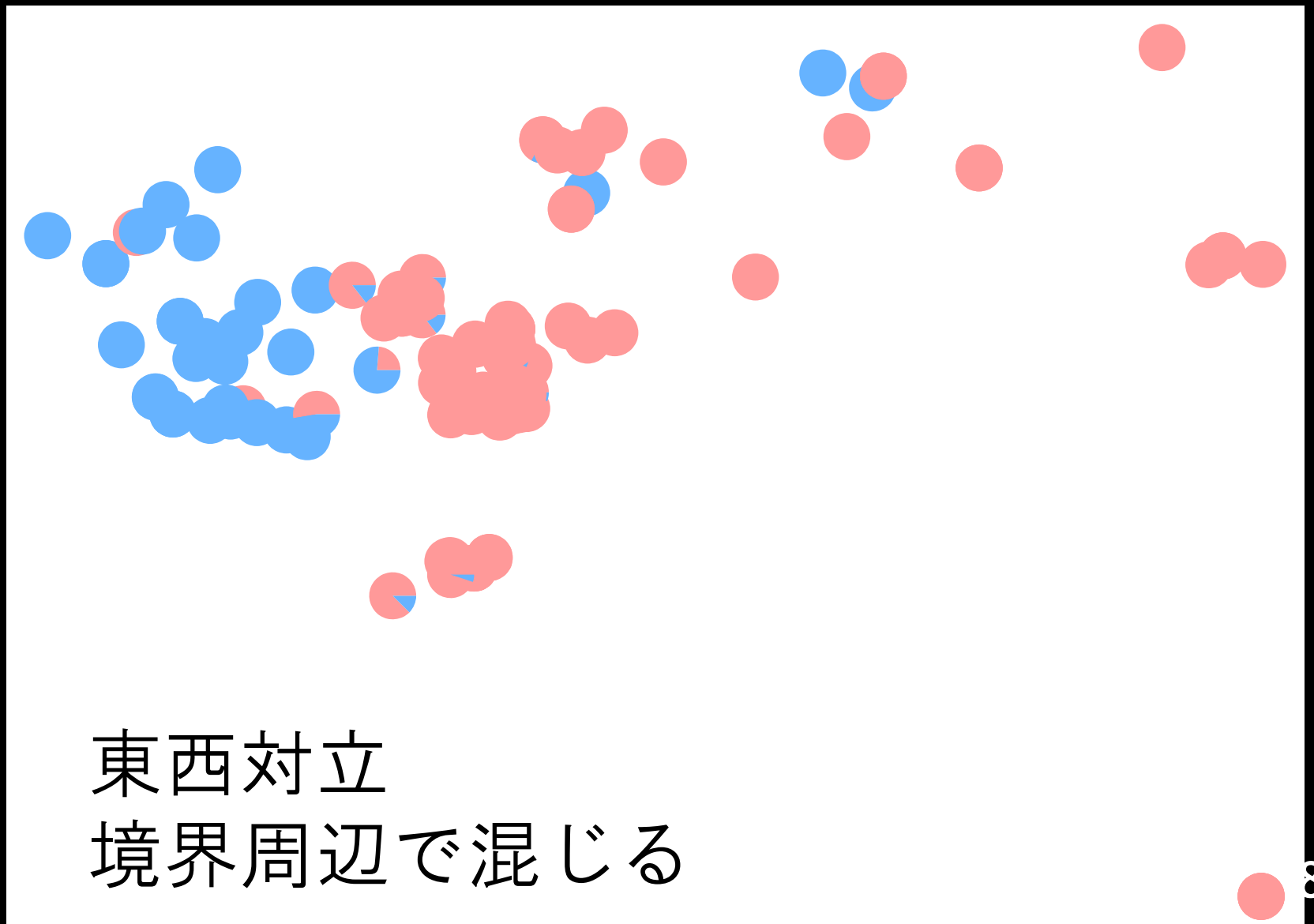
[Skoglund+, Nature 2016]

~~ADMIXTURE~~解析 [Pritchard+, 2000]

実際にはSTRUCTUREにより近い独自モデル

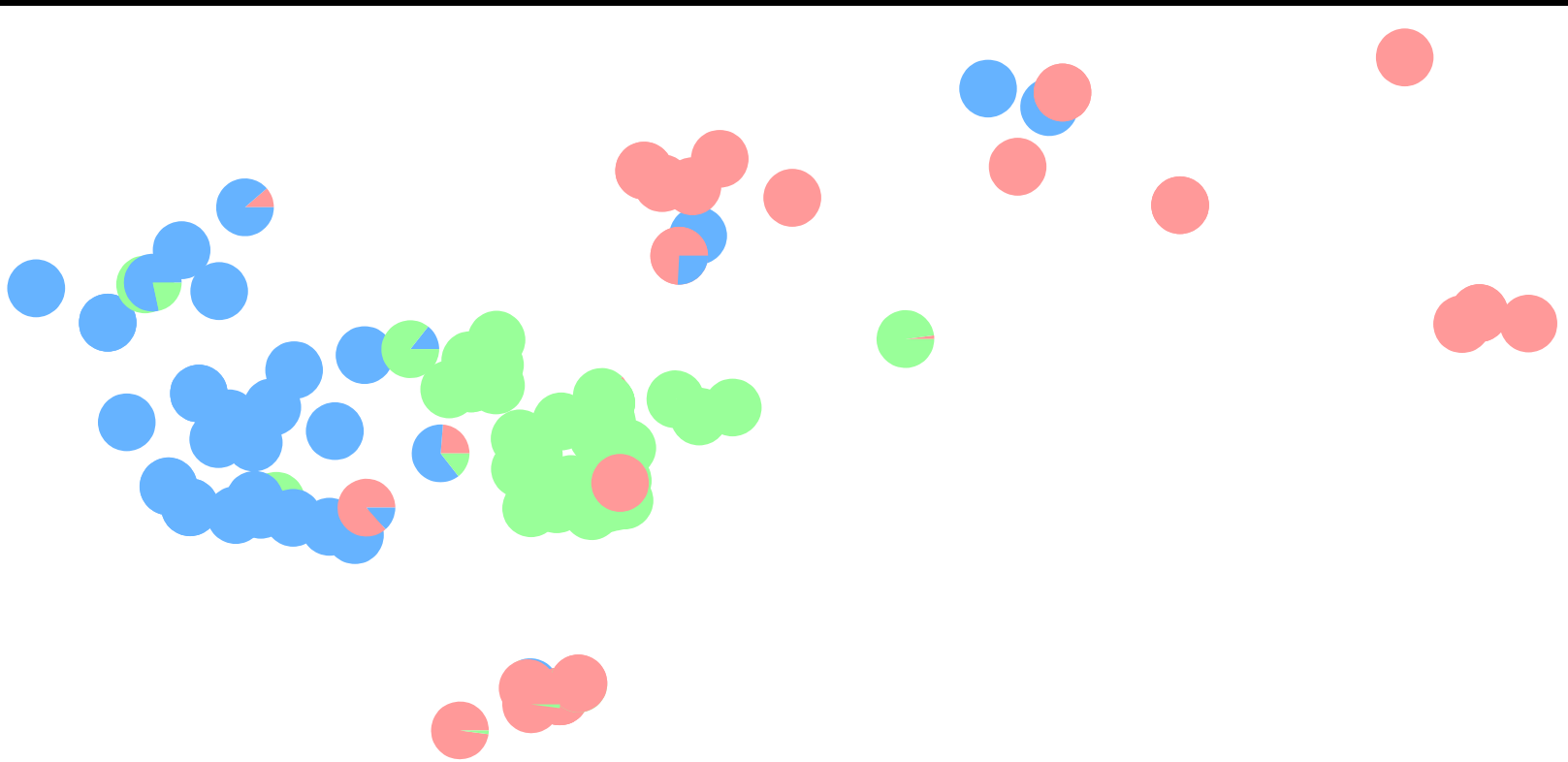
- 各言語がK個の祖先集団の混合であると仮定し、混合比を推定
- 集団遺伝学のモデル
  - 自然言語処理のLDA (Latent Dirichlet Allocation) に似ている
  - Reesink+ (2009) は言語類型論のデータに適用
- DNAを対象にするなら良いけど、言語の特徴を扱うには微妙な点がある

ADMIXTURE解析 $K = 2$



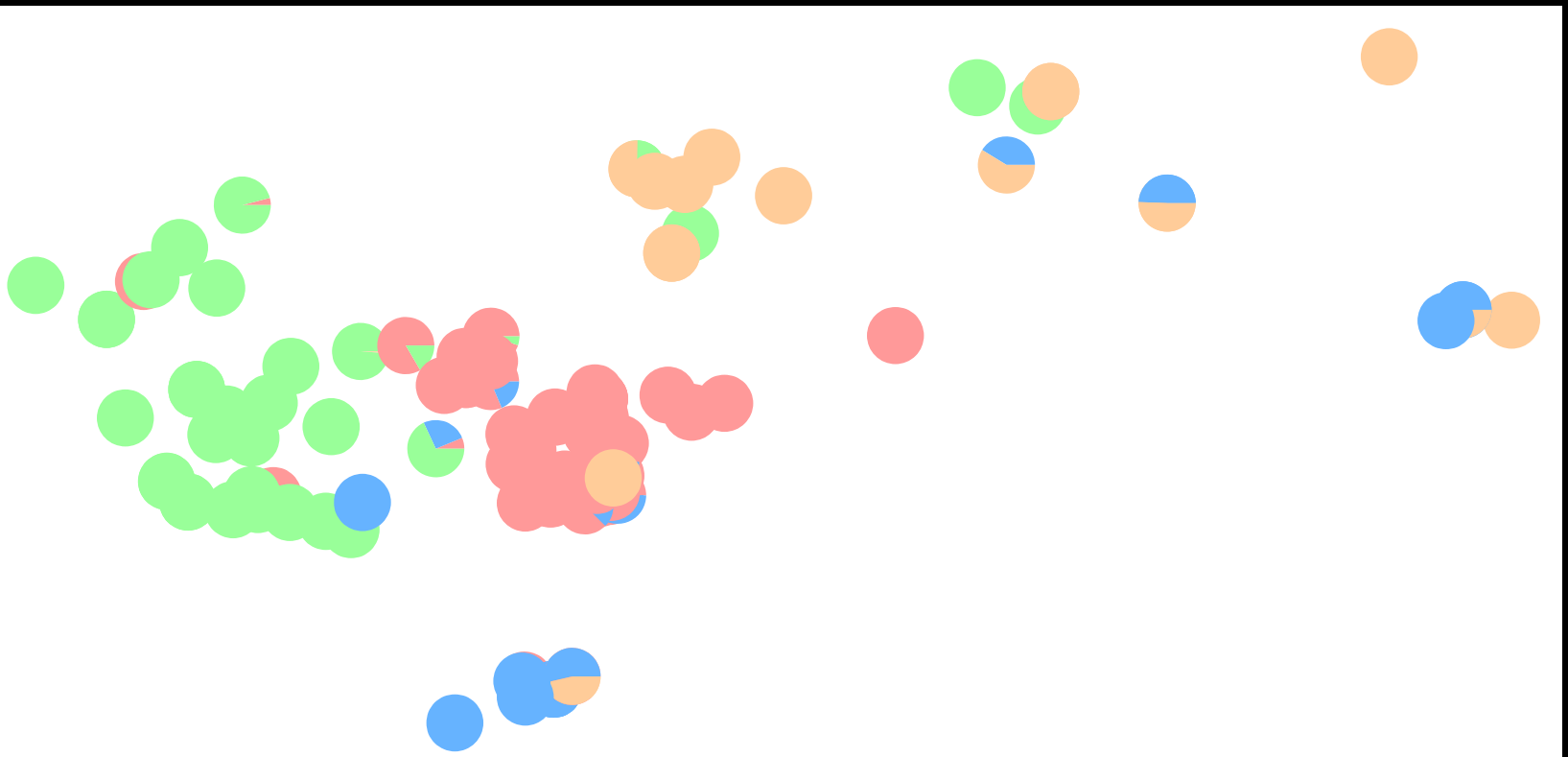
東西対立
境界周辺で混じる

ADMIXTURE解析 $K = 3$



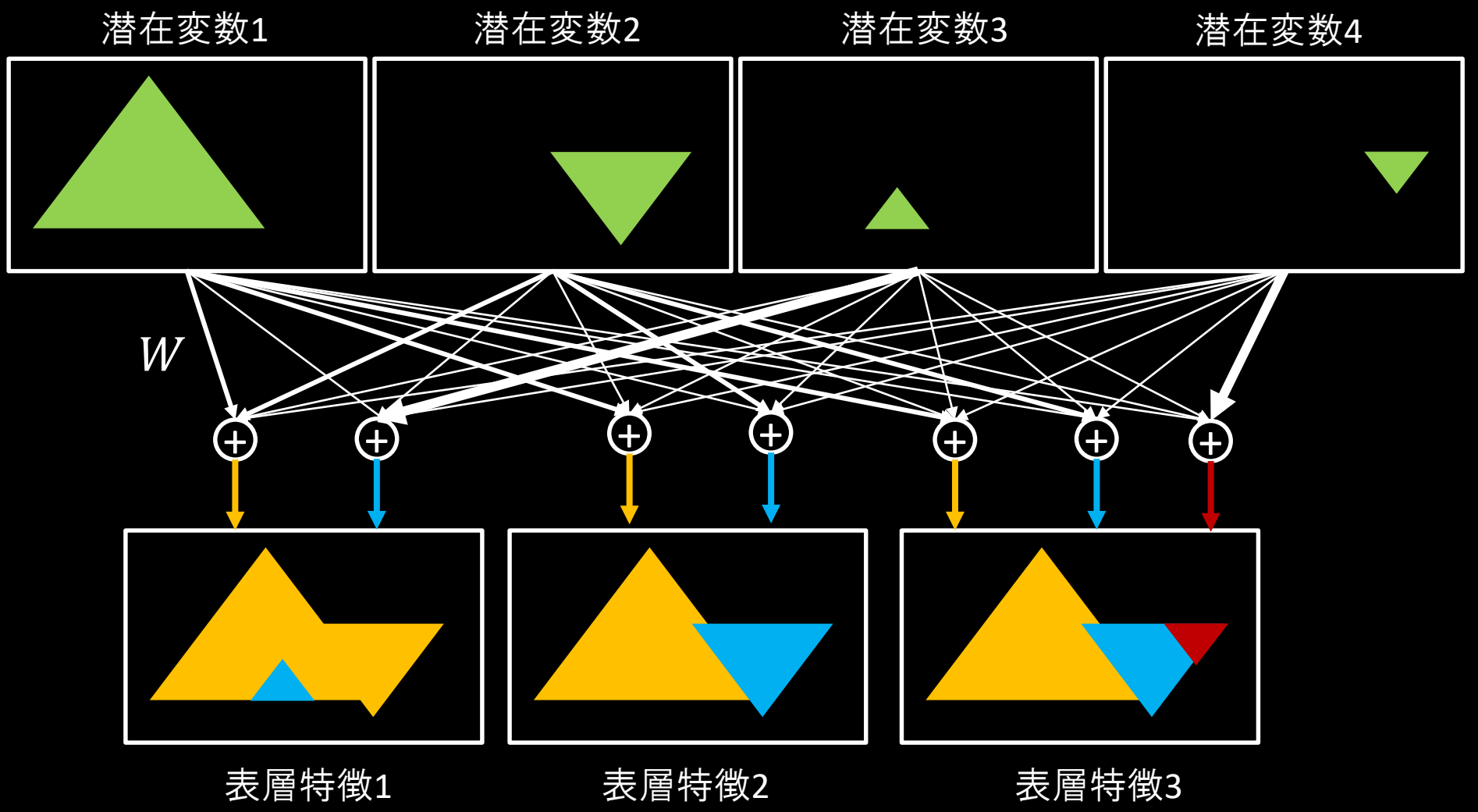
東方言が分割された

ADMIXTURE解析 $K = 4$



よくわからない…

提案手法: 地理的分布の再構成



提案手法: 地理的分布の再構成

1. 各特徴の地理的分布を一般に1個以上の分布に分解
2. 複数の表層特徴に共通する地理的分布をクラスタリング
3. 平滑化
   - 新しい変化の影響で元の分布の一部が観測できていない

実はMurawaki (2019) のマイナーチェンジ

Z

K binary parameters

L languages					
1	0	1	0	⋯	
1	1	0	0	⋯	
0	1	0	1	⋯	
⋮	⋮	⋮	⋮	⋱	
1	0	1	0	⋯	

×

W

K binary parameters	M binarized features				
2.9	0.4	0.3	⋯	0.2	
6.3	0.2	0.1	⋯	5.9	
8.2	0.2	0.5	⋯	0.3	
0.2	0.3	1.2	⋯	0.1	
⋮	⋮	⋮	⋮	⋮	

=

$\tilde{\Theta}$

M binarized features

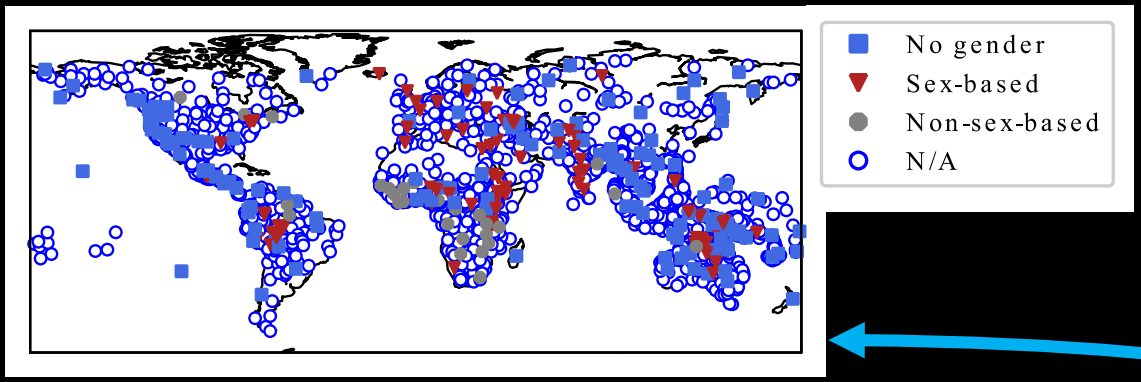
L languages					
10.2	9.8	0.9	⋯	2.9	
2.1	8.9	0.9	⋯	9.4	
8.4	2.3	0.3	⋯	2.5	
⋮	⋮	⋮	⋮	⋮	
3.2	1.8	3.9	⋯	0.3	

Draw from softmax distributions ⇓

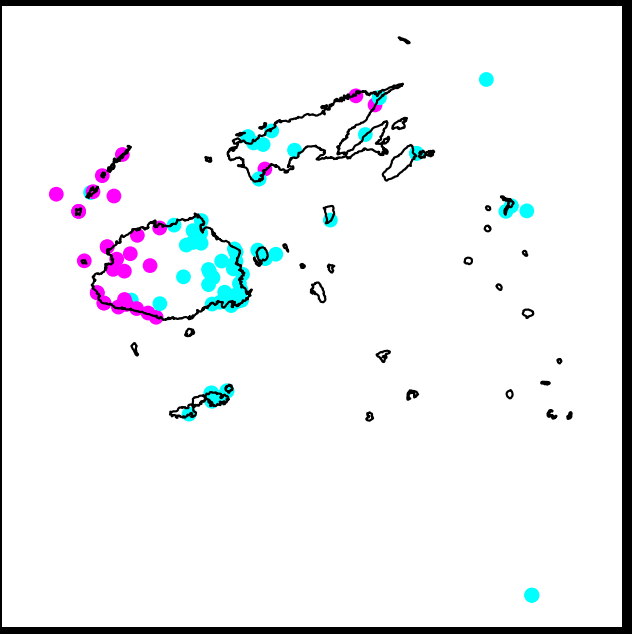
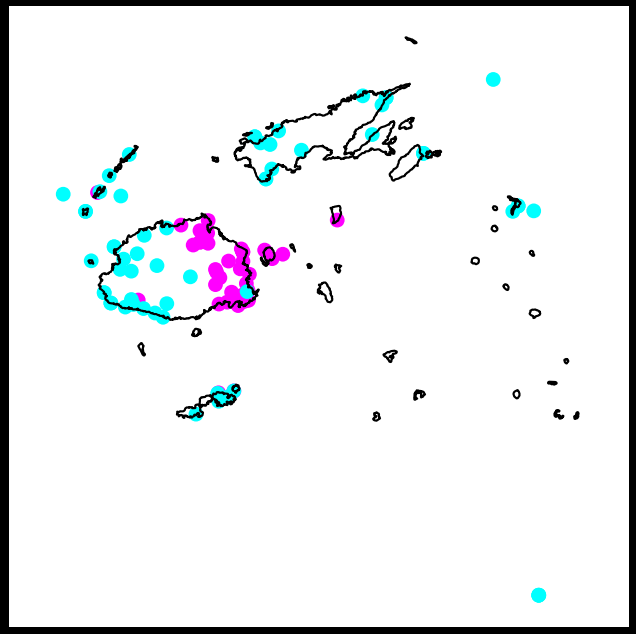
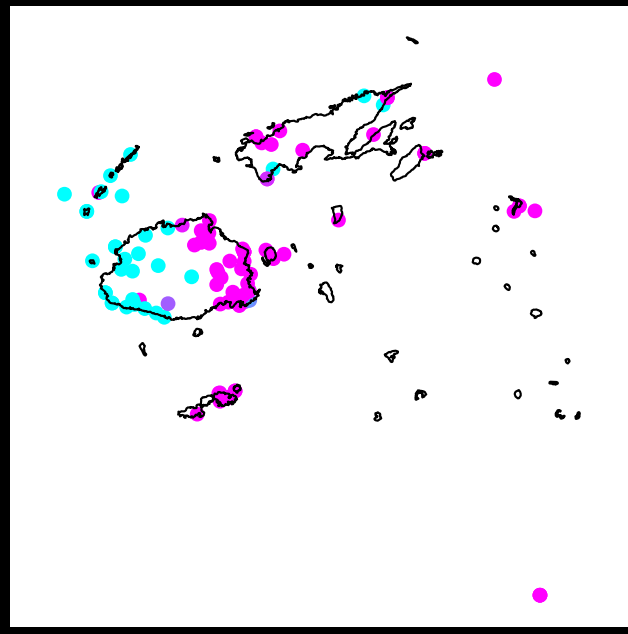
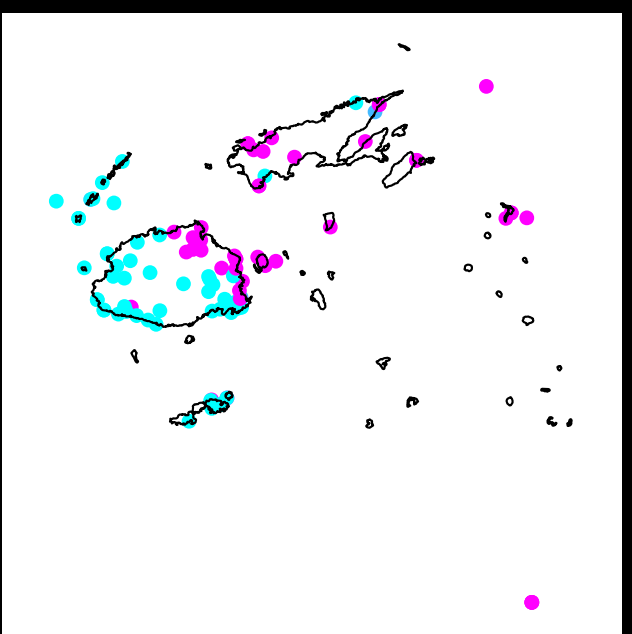
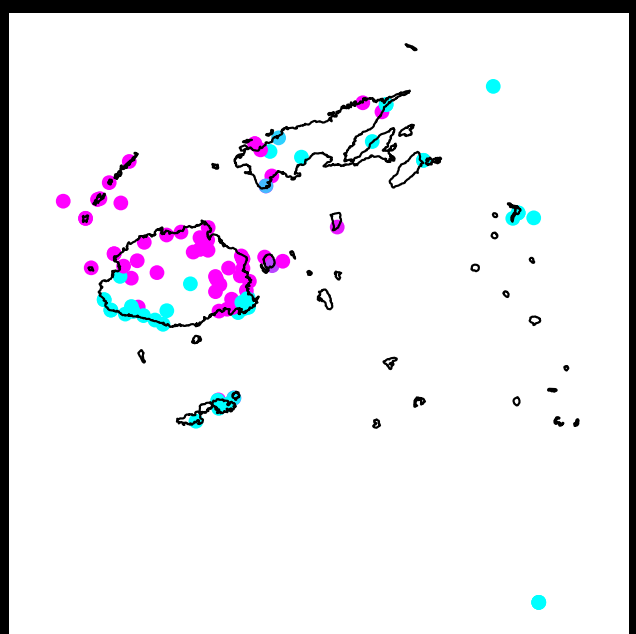
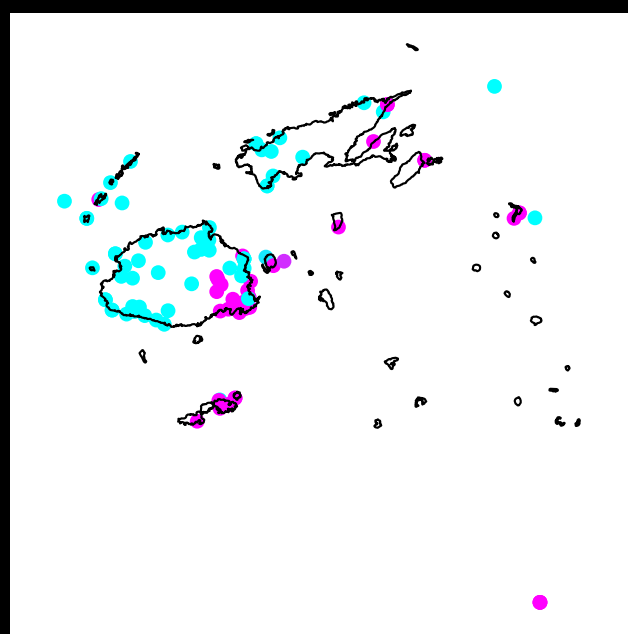
N categorical features

L languages				
1	1	⋯	2	
2	4	⋯	3	
1	1	⋯	3	
⋮	⋮	⋮	⋮	
3	1	⋯	1	

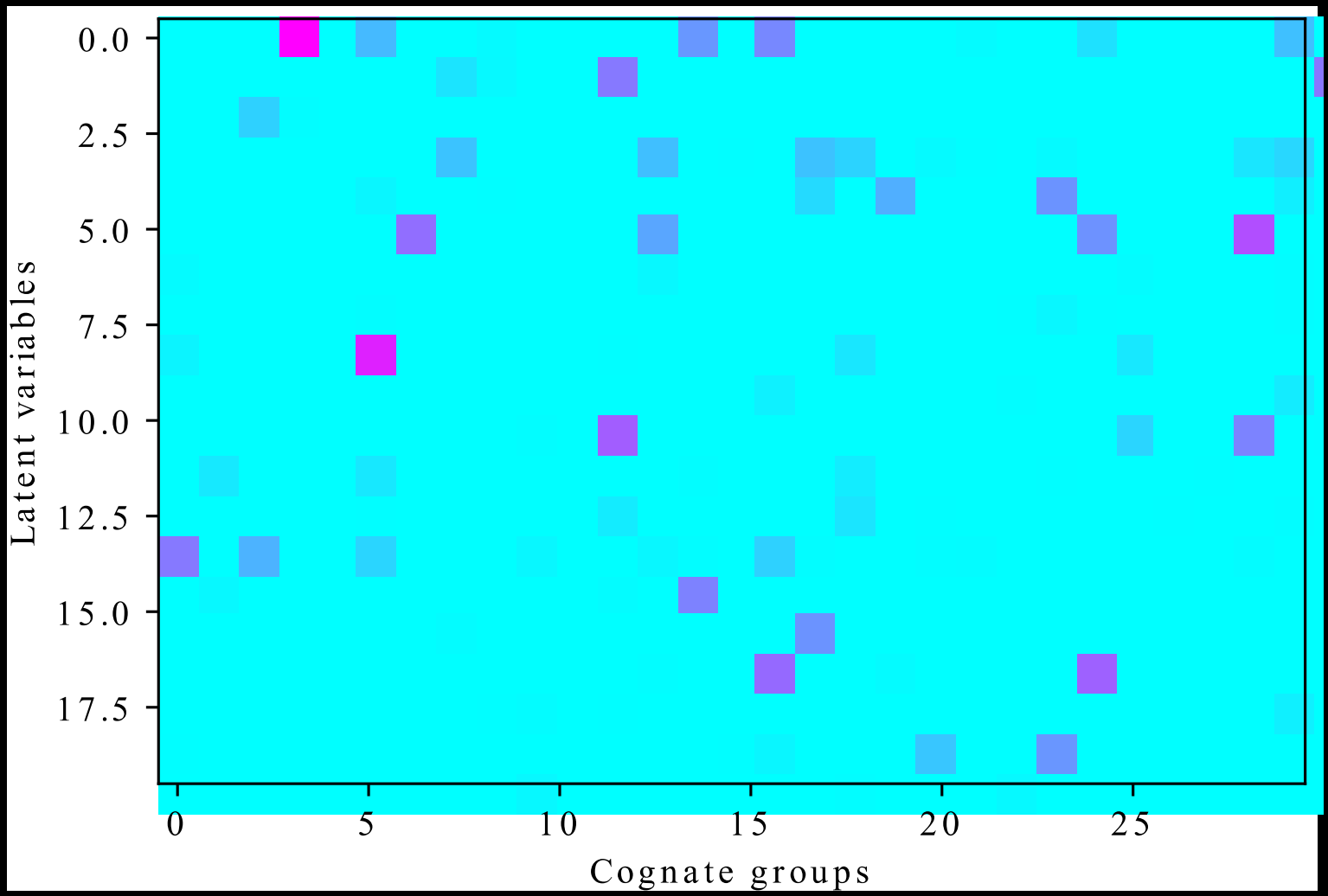
X



地理的分布の再構成 $K = 20$

W


Latent variables
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0
5
10
15
20
25
Cognate groups

まとめと今後の課題

- 潜在変数を用いて諸特徴の地理的分布を再編成するモデルを提案
- 他の手法ではデータ全体を支配する要素を上位1-2個程度しか把握できないのに対し、提案手法はより多くの要素を可視化できる
- 今後の課題
  - データセット、前処理の精緻化
  - 時間に関する推論
  - 人文地理学的要素の組み込み