方言解析のための
空間的インド料理店過程

持橋大地
統計数理研究所 / 国立国語研究所
daichi@ism.ac.jp

次世代言語科学研究センター 開所式
2025-3-3(月)

本研究は、菅澤翔之助氏(慶応大学経済学部)との共同研究です。

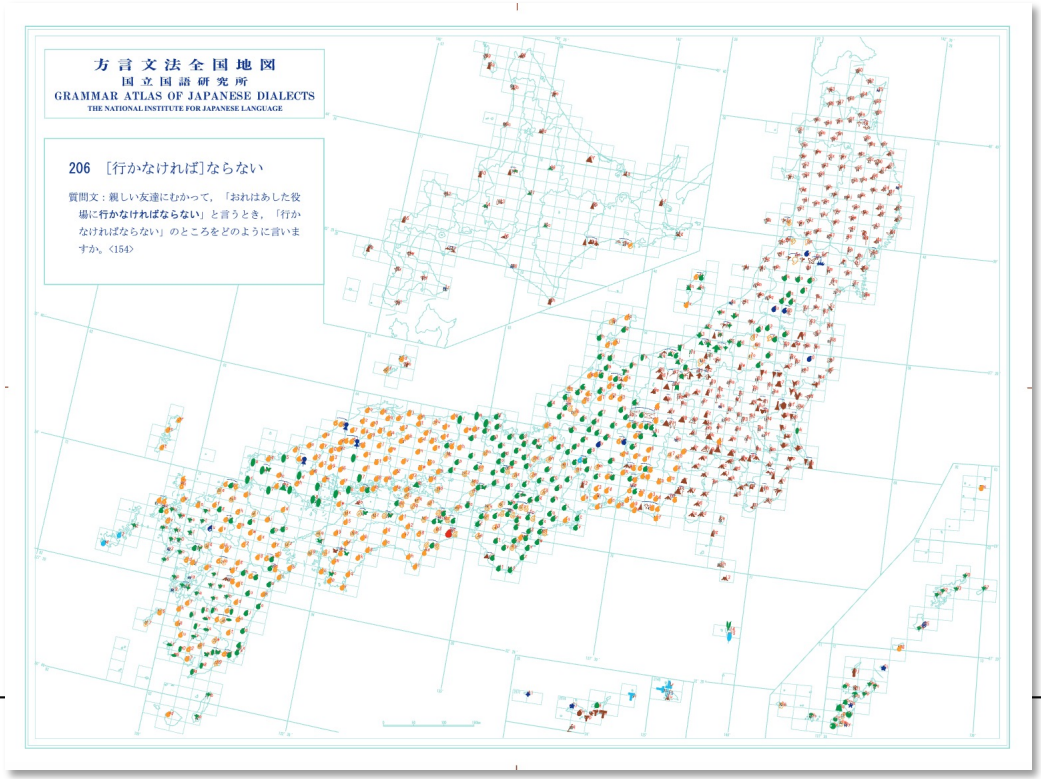
自己紹介

- 持橋大地
  統計数理研究所 統計基盤数理研究系 教授
  国立国語研究所 次世代言語科学研究センター 教授
  (クロスアポイントメント)
- 専門: 自然言語処理、統計的機械学習

統数研:
唯一の統計学の
国立研究所
(立川市、国語研の隣)

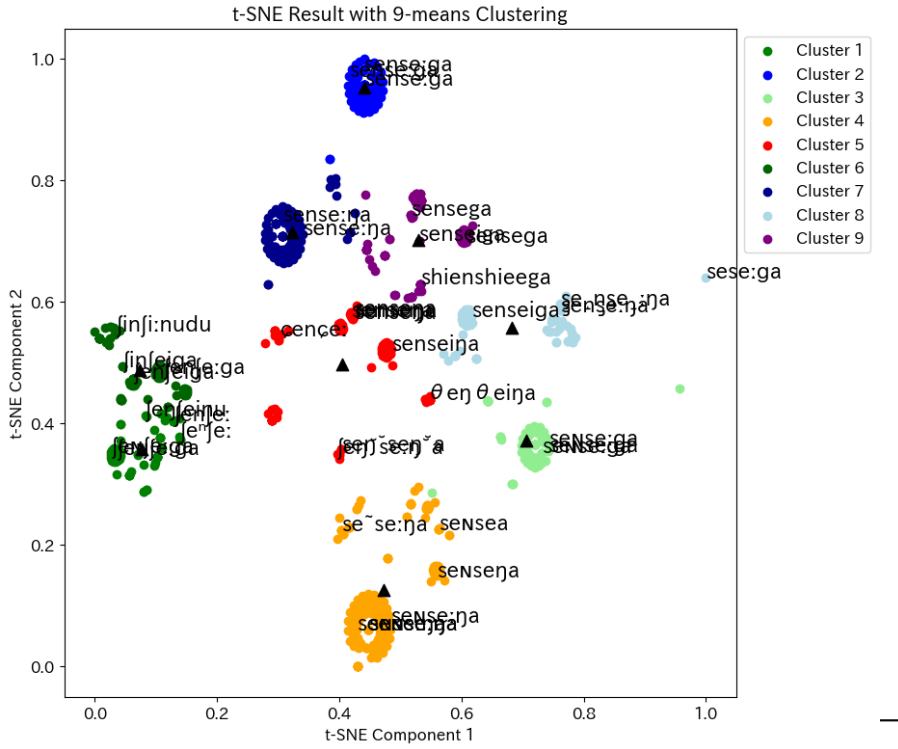
方言のデータ化と分析

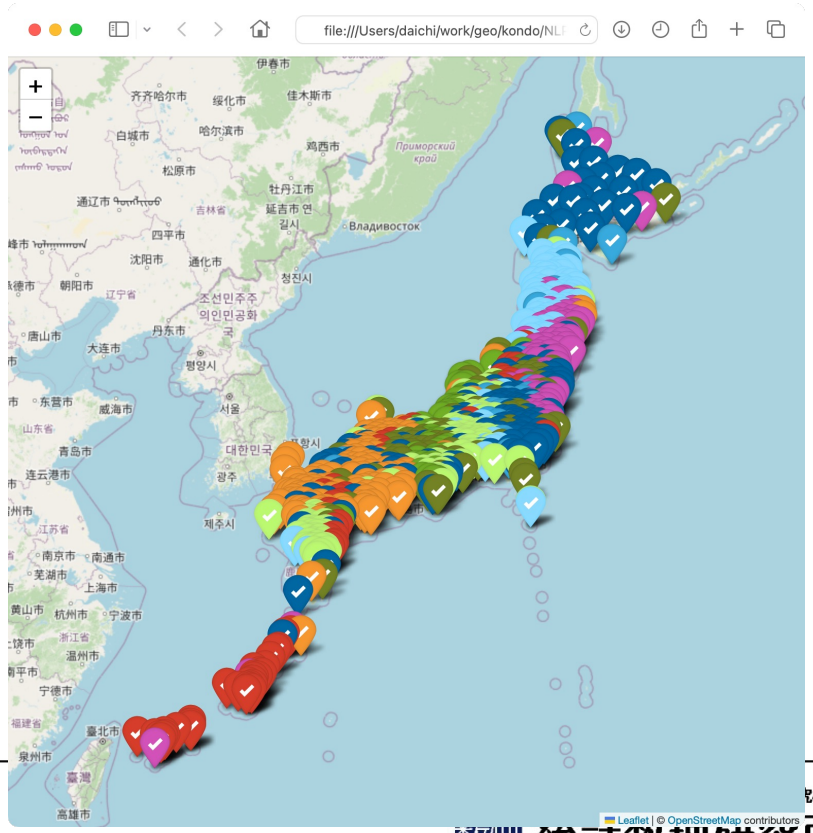
- 「日本言語地図」(1966年～),「新日本言語地図」(2016年)、「方言文法全国地図」(1989年～)など、国立国語研究所が古くからデータ化と研究を行ってきた



近藤先生との共同研究 (2025)

- 単語の綴りのnグラム特徴ベクトルをクラスタ化し、カーネル整列法で最適な色を用いて描画
  - 言語処理学会2025で発表予定





言語地理学の研究を行うきっかけ

- 国立民族学博物館 菊澤さんリーダーの国際科研費プロジェクト (2018〜, 2024からも新しく採択)

フィジーの場所


フィジー
ロシア
モンゴル
中華人民
共和国
大韓民国
日本
ネパール
インド
ミャンマー
(ビルマ)
タイ
ベトナム
フィリピン
マレーシア
インドネシア
パプア
ニューギニア
オーストラリア
ニュージーランド
カナダ
アメリカ合衆
メキシコ
ログイン
レイヤ
Google
画像 ©2024 NASA、TerraMetrics、地図データ ©2024 Google、INEGI
日本
利用規約
プライバシー
サービスに関するフィードバックを送信
2000 km

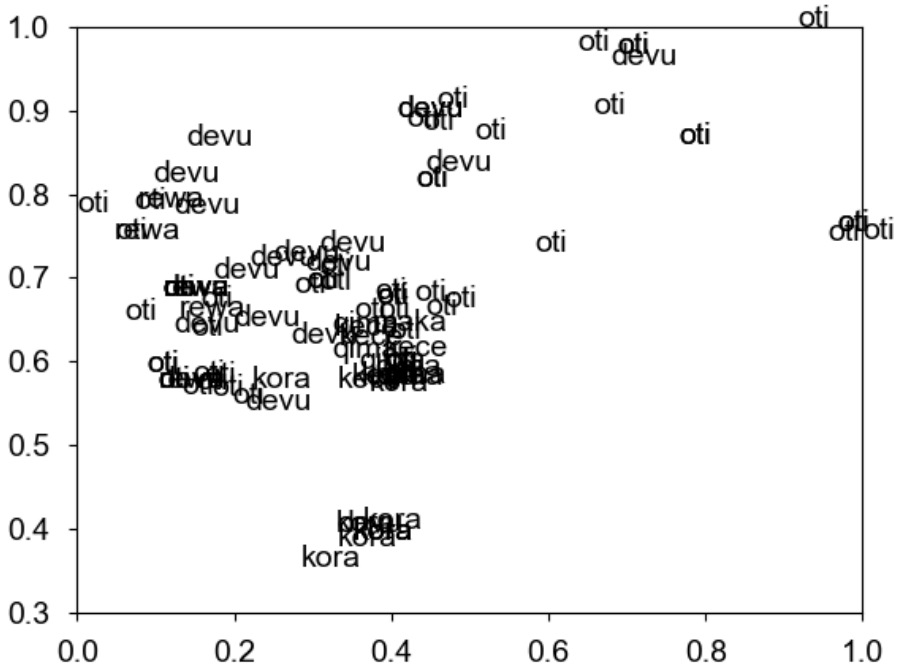
フィジー全体の地図


Google マップを検索する
あ
D
Nalele
Vatutandova
ラバサ
Labasa
Nasealevu
バトア島
Vanua Levu
サブサブ
Savusavu
Somosomo
カナセア
Kanacea
Nabouwalu
ナンディ
国際空港
ロマロマ
Lomaloma
Thikombia
ラキラキ
Rakiraki
ラウトカ
Lautoka
レブカ
Levuka
デナラウ島
Denarau Island
ナンディ
Nadi
フィジー
ロマジ
Lomaji
Liku
シンガトカ
Sigatoka
スバ
Suva
Cuvu
ナブア
Navua
首都スバ &
南太平洋大学
Tubou
Vanuavatu
ビッティ・レビュー・グループ
Viti Levu Group
ダクイロア
Dakuiloa
ナソー
Nasau
モアラ・アイランズ
Moala Islands
ナマック
Namaku
カンダブ
Kadavu Island
ヤロイ
Yaroi
レビュケイダック
Levukaidaku
Monothaki
Ongea
レイヤ
Google
画像 ©2024 TerraMetrics、地図データ ©2024 Google 日本 利用規約 プライバシー サービスに関するフィードバックを送信 50 km

フィジー語の方言

180.58	39.84	oti devu
182.76	39.67	oti rewa
183.34	39.17	oti
183.90	39.86	oti
184.13	39.87	rewa devu
184.61	38.85	oti devu
185.01	40.03	devu
185.27	38.75	oti devu rewa
185.54	39.32	oti rewa devu
186.13	39.09	devu
186.16	39.83	devu
186.46	39.19	rewa oti
186.66	38.71	oti devu
186.94	40.25	devu
187.21	39.07	oti
187.24	38.79	oti
187.42	38.74	oti

- どうやって、こうした空間的変異を扱うか?



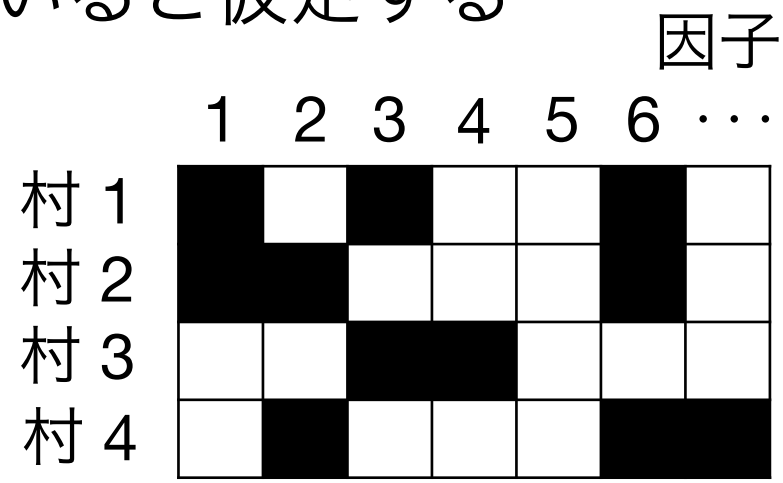
プロジェクトで作成したデータ

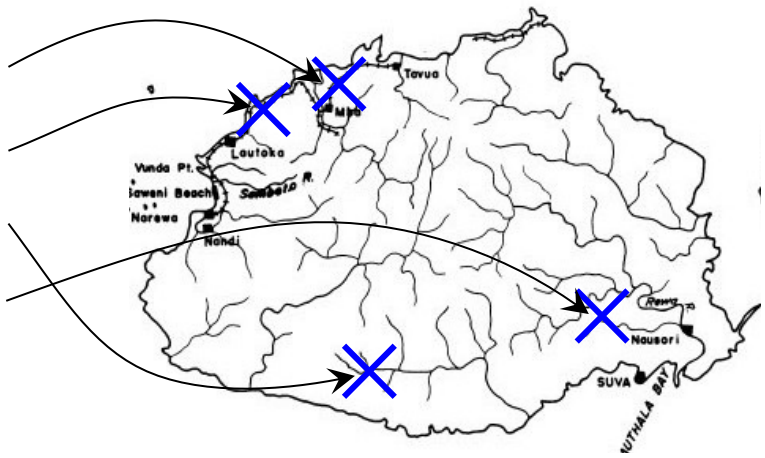
- 100 Word Lists : 100個の意味 x 村の方言の行列
  - 空間的な離散データ、列ごとに語彙が異なる

	A	B	C	D	E	F	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1	OBJECTID_	VillagesComCodes_	VillagesComCo	VillagesCor	ComCodeGrp	ComGrou	e__at__in	na__the_	nō__gen	N_oqō__t	oqori__tha	oyā__that	iko__you_	koya__hin	keda__us_	keimami_
2	1	Tuvana village			Ono	Lau	i	a	we	kaiīiei	xxx	xxx	iko	koikoyaikoy	kīketaketa	keimami
3	2	Matavuralevu Settlement			Ono	Lau	i	a	we	kaiīiei	xxx	xxx	iko	koikoyaikoy	kīketaketa	keimami
4	3	Matokana village			Ono	Lau	i	a	we	kaiīiei	xxx	xxx	iko	koikoyaikoy	kīketaketa	keimami
5	4	Nukuni village			Ono	Lau	i	a	we	kaiīiei	xxx	xxx	iko	koikoyaikoy	kīketaketa	keimami
6	5	Lovoni village			Ono	Lau	i	a	we	kaiīiei	xxx	xxx	iko	koikoyaikoy	kīketaketa	keimami
7	6	Doi village			Ono	Lau	i	a	we	kaiīiei	xxx	xxx	iko	koikoyaikoy	kīketaketa	keimami
8	7	Vatoa village		LAU	Lauicake	Lau	i	a	o−	iei, ī	iqore, iqoya	maiei, maiī	iko	koikoya	keiketa, ke	keimami
9	8	Ogea Village		LAU	Lauicake	Lau	i	a	o−	iei, ī	iqore, iqoya	maiei, maiī	iko	koikoya	keiketa, ke	keimami
10	9	Muanaicake Village			Lauicake	Lau	i	a	o−	iei, ī	iqore, iqoya	maiei, maiī	iko	koikoya	keiketa, ke	keimami
11	10	Muanaicake Village			Lauicake	Lau	i	a	o−	iei, ī	iqore, iqoya	maiei, maiī	iko	koikoya	keiketa, ke	keimami
12	11	Matanuku Village	RAVITAKI	KADAVU	Ravitaki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
13	12	Burelevu Village	RAVITAKI	KADAVU	Ravitaki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
14	13	Muani Village		KADAVU	Ravitaki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
15	14	Levuka Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
16	15	Kabariki Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
17	16	Nasau Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
18	17	Qaliira Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
19	18	Nadaviqele Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
20	19	Naividamu Village			Lauicake	Lau	i	a	o−	iei, ī	iqore, iqoya	maiei, maiī	iko	koikoya	keiketa, ke	keimami
21	20	Cevai Village	TAVUKI	KADAVU	Tavuki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
22	21	Namanusa Village	RAVITAKI	KADAVU	Ravitaki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
23	22	Nabukelevuira Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
24	23	Lomati Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
25	24	Talaulia Village		KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
26	25	Mokoisa Village		KADAVU	Ravitaki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
27	26	Dagai Village		KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
28	27	Wailevu Village	RAVITAKI	KADAVU	Ravitaki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
29	28	Baidamudamu Village	TAVUKI	KADAVU	Tavuki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
30	29	Waisomo Village		KADAVU	Tavuki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
31	30	Natumua Village	TAVUKI	KADAVU	Tavuki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī

問題

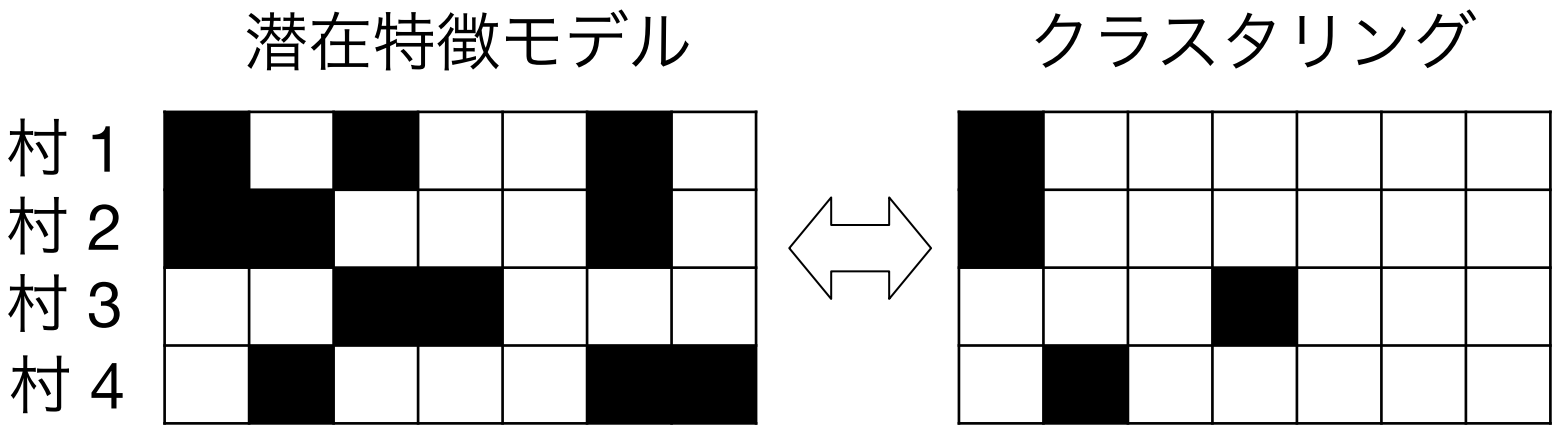
- 村ごとに100個の概念に対する方言のデータから、潜在的なパターンを見つけたい
- 全体を単にグループ化するのではなく、各村は、幾つかの未知の潜在的な因子を持っていると仮定する





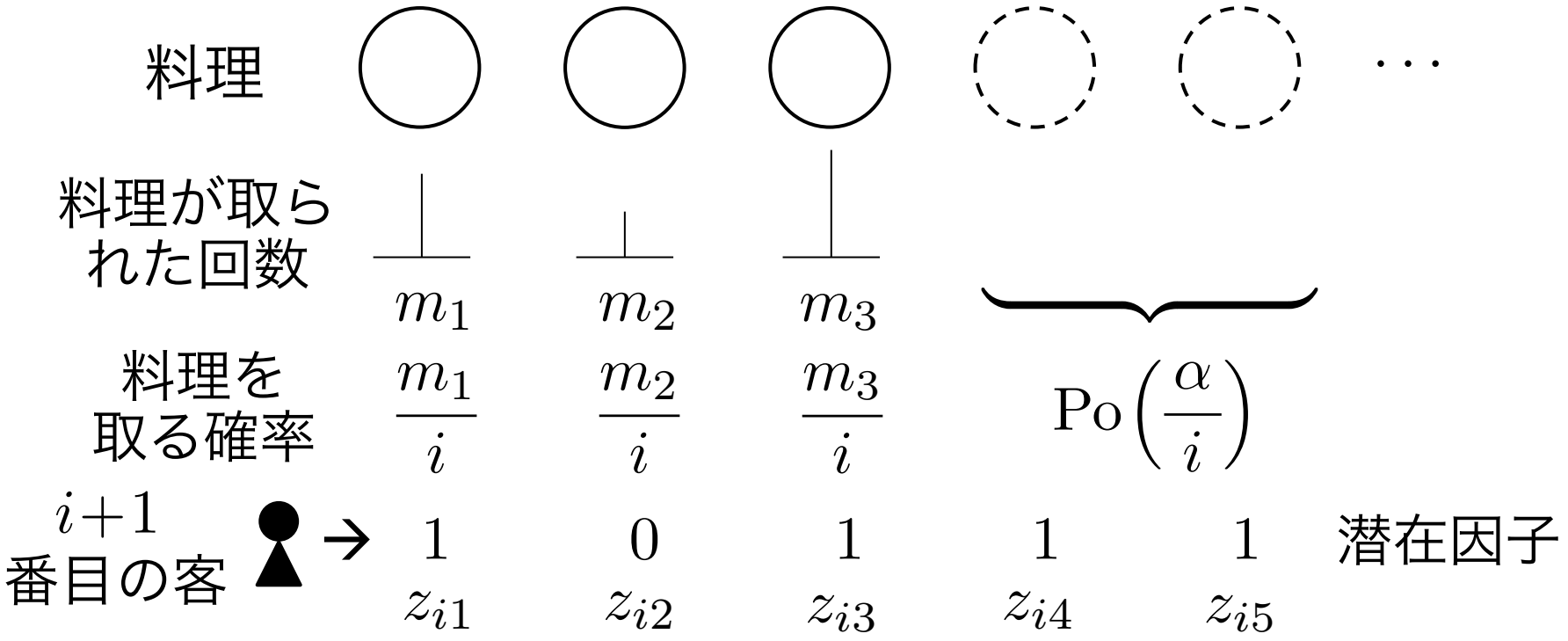
潜在特徴モデルとクラスタリング

- こうしたモデルは、潜在特徴モデルとよばれる
  - クラスタリングでは、各村は1つのグループにだけ所属
  - 潜在特徴モデルでは、各村は複数の因子が「オン」になっている (組み合わせは2^K通り)

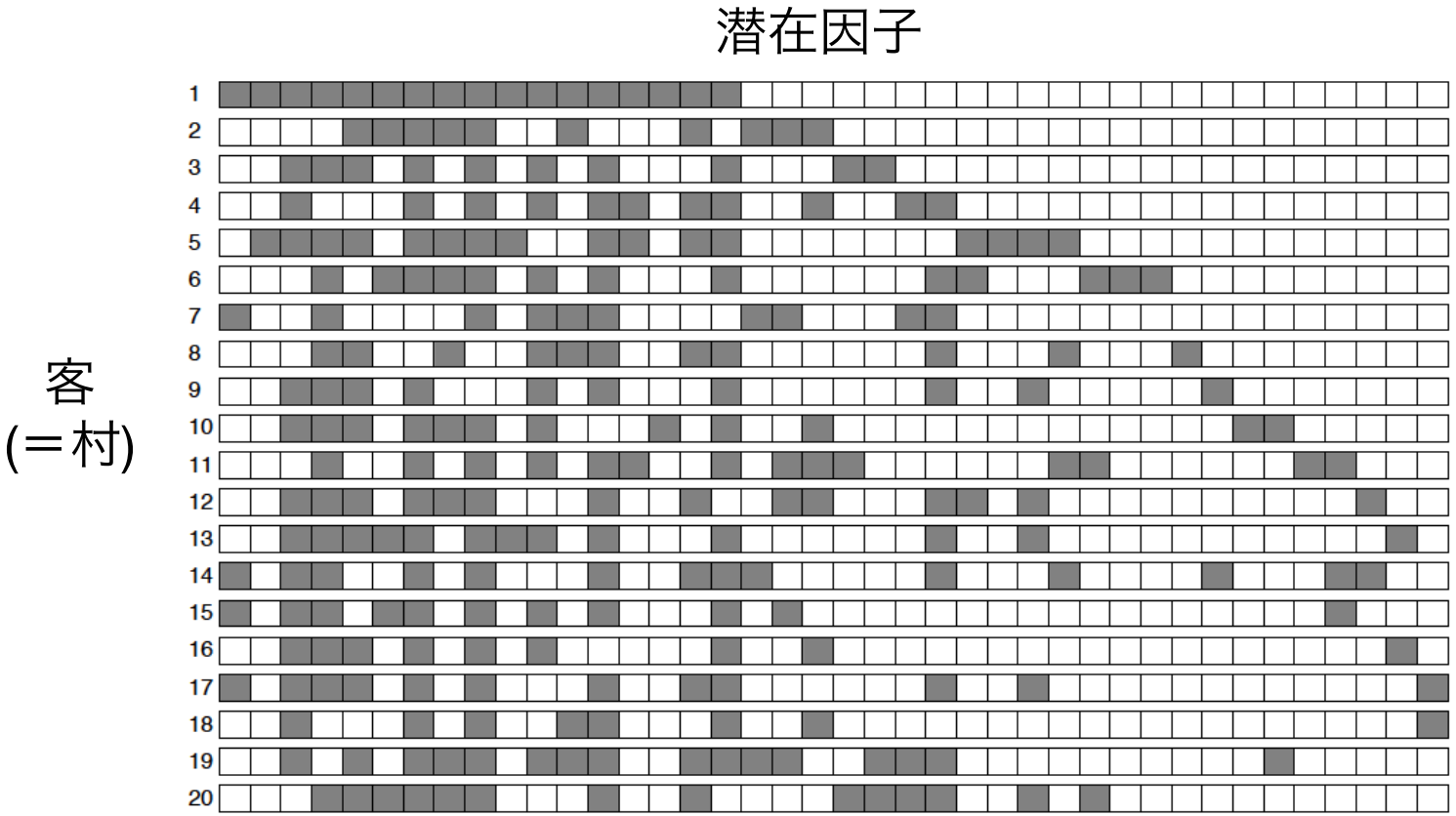


インド料理店過程(IBP) (Griffiths&Ghahramani 2005)

- こうした潜在的因子の生成モデル
  → インド料理店過程 (Indian buffet process, IBP)
- 無限個の特徴を生成できる



IBPで生成された潜在因子の例



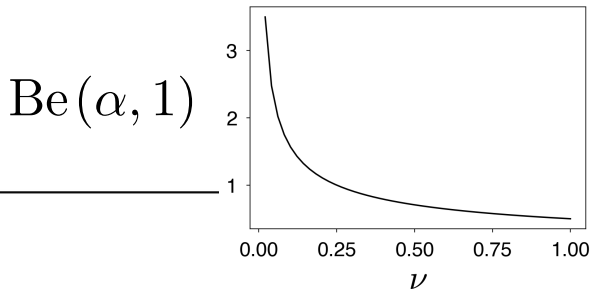
- $\alpha = 10$ のIBPから生成された因子行列
  - この場合は、村は独立→地理的相関を入れたい

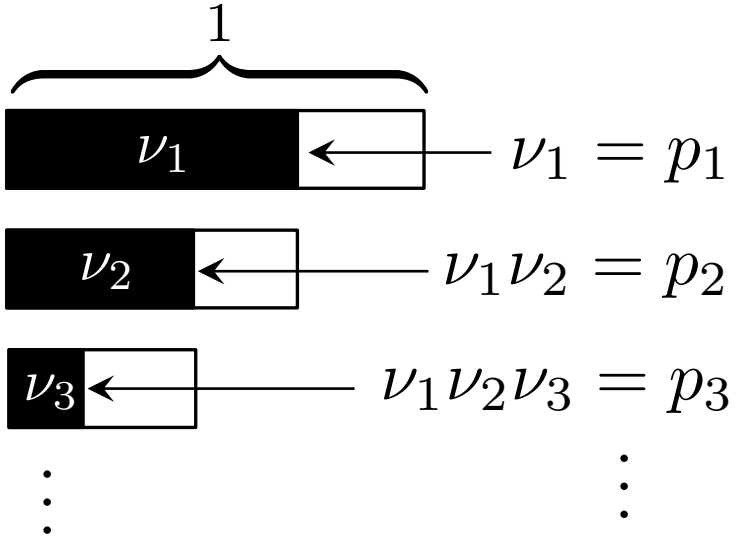
空間的インド料理店過程

- どうやって、地理的に近い客(この場合は村)の間の相関を入れるか?
  → IBPのstick-breaking表現 (Teh+ 2007)
- 長さ1の棒を確率的に折ることで、離散的な確率を生成する (元々はベータ分布を使用)

$$p(z_{ik}=1) = \prod_{j=1}^{k} \nu_j \,,$$

$$\nu_j \sim \mathrm{Be}(\alpha, 1)$$



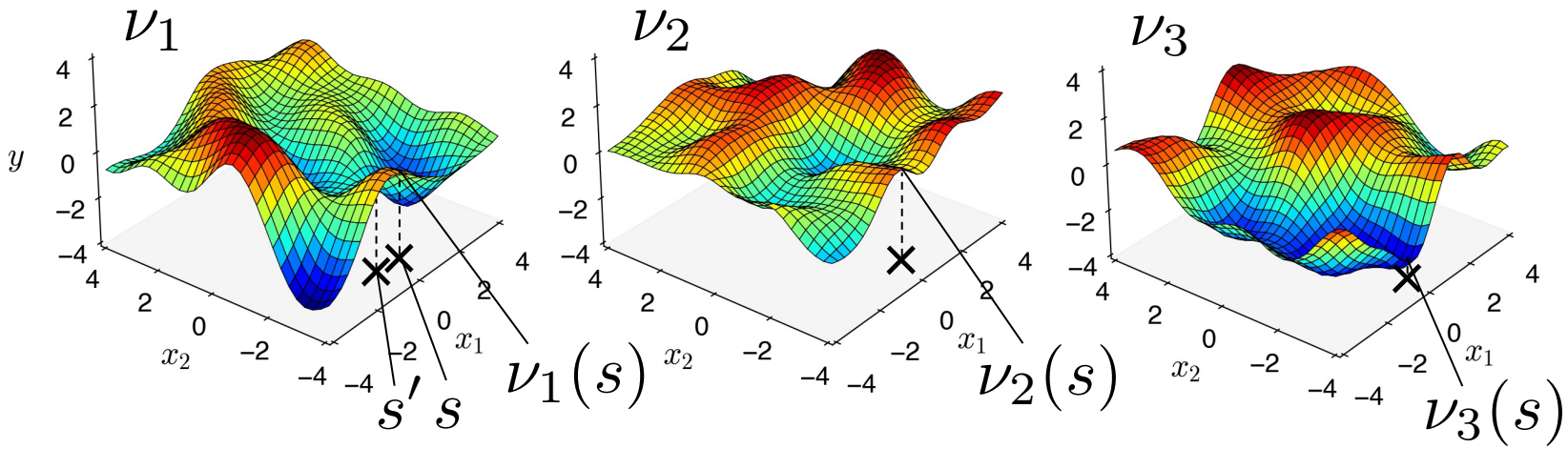


空間的インド料理店過程 (2)

- ロジスティック棒折り過程 (Ren+, JMLR2011)の方法を使うことで、ガウス過程で棒を折る場所を決定 (＝空間相関)

$$p(z_{ik}(s_i)=1) = \prod_{j=1}^{k} \sigma(\nu_j(s_i)),\ \sigma(x)=\frac{1}{1+e^{-x}},$$

$$(\nu_j(s_1),\cdots,\nu_j(s_N)) \sim \mathrm{GP}(s_1,\cdots,s_N)$$



プロジェクトで作成したデータ

- 100 Word Lists : 100個の意味 x 村の方言の行列
  - 空間的な離散データ、列ごとに語彙が異なる

	A	B	C	D	E	F	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1	OBJECTID_	VillagesComCodes_	VillagesComCo	VillagesCor	ComCodeGrp	ComGrou	e__at__in	na__the_	nō__gen	N_oqō__t	oqori__tha	oyā__that	iko__you_	koya__hin	keda__us_	keimami_
2	1	Tuvana village			Ono	Lau	i	a	we	kaiīiei	xxx	xxx	iko	koikoyaikoy	kīketaketa	keimami
3	2	Matavuralevu Settlement			Ono	Lau	i	a	we	kaiīiei	xxx	xxx	iko	koikoyaikoy	kīketaketa	keimami
4	3	Matokana village			Ono	Lau	i	a	we	kaiīiei	xxx	xxx	iko	koikoyaikoy	kīketaketa	keimami
5	4	Nukuni village			Ono	Lau	i	a	we	kaiīiei	xxx	xxx	iko	koikoyaikoy	kīketaketa	keimami
6	5	Lovoni village			Ono	Lau	i	a	we	kaiīiei	xxx	xxx	iko	koikoyaikoy	kīketaketa	keimami
7	6	Doi village			Ono	Lau	i	a	we	kaiīiei	xxx	xxx	iko	koikoyaikoy	kīketaketa	keimami
8	7	Vatoa village		LAU	Lauicake	Lau	i	a	o−	iei, ī	iqore, iqoya	maiei, maiī	iko	koikoya	keiketa, ke	keimami
9	8	Ogea Village		LAU	Lauicake	Lau	i	a	o−	iei, ī	iqore, iqoya	maiei, maiī	iko	koikoya	keiketa, ke	keimami
10	9	Muanaicake Village			Lauicake	Lau	i	a	o−	iei, ī	iqore, iqoya	maiei, maiī	iko	koikoya	keiketa, ke	keimami
11	10	Muanaicake Village			Lauicake	Lau	i	a	o−	iei, ī	iqore, iqoya	maiei, maiī	iko	koikoya	keiketa, ke	keimami
12	11	Matanuku Village	RAVITAKI	KADAVU	Ravitaki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
13	12	Burelevu Village	RAVITAKI	KADAVU	Ravitaki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
14	13	Muani Village		KADAVU	Ravitaki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
15	14	Levuka Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
16	15	Kabariki Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
17	16	Nasau Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
18	17	Qaliira Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
19	18	Nadaviqele Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
20	19	Naividamu Village			Lauicake	Lau	i	a	o−	iei, ī	iqore, iqoya	maiei, maiī	iko	koikoya	keiketa, ke	keimami
21	20	Cevai Village	TAVUKI	KADAVU	Tavuki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
22	21	Namanusa Village	RAVITAKI	KADAVU	Ravitaki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
23	22	Nabukelevuira Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
24	23	Lomati Village	NAKUKELEVU	KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
25	24	Talaulia Village		KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
26	25	Mokoisa Village		KADAVU	Ravitaki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
27	26	Dagai Village		KADAVU	Nabukelevu	Kadavu				kea	meri	mē	iko	kaia	kēdā	
28	27	Wailevu Village	RAVITAKI	KADAVU	Ravitaki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
29	28	Baidamudamu Village	TAVUKI	KADAVU	Tavuki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
30	29	Waisomo Village		KADAVU	Tavuki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī
31	30	Natumua Village	TAVUKI	KADAVU	Tavuki	Kadavu	i	na	nō−	kā	kari	kacei	iko	kia	keda	kēmī

方言解析の目的関数

- $\pi_{im\ell}(s)$: 意味 m を表すのに、i 番目の村 (座標 s)で方言 ℓ が使われる確率
- 村の潜在的特徴

$$\pi_{im\ell}(s) = \frac{\exp\left(\eta_{m\ell} + \sum_{k=1}^{K} \theta_{km\ell} z_{ik}(s)\right)}{\sum_{\ell=1}^{c_m} \exp(\cdots)}$$

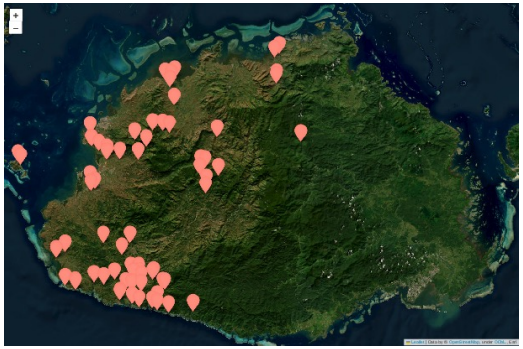
  - 例: Naividamu村 (南緯18.1267°, 東経177.5200°) で、“him/her”を表すのに ‘koikoya’ を使う確率 (他の候補: ‘kia’, ‘kaia’, ‘koya’, ‘kua’)
  - K は最大の潜在特徴数 (ここではK=10) で、自動学習
  - 観測データ全体の対数尤度を最大化:

$$L = \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{\ell=1}^{c_m} \log \pi_{im\ell}(s_i)$$

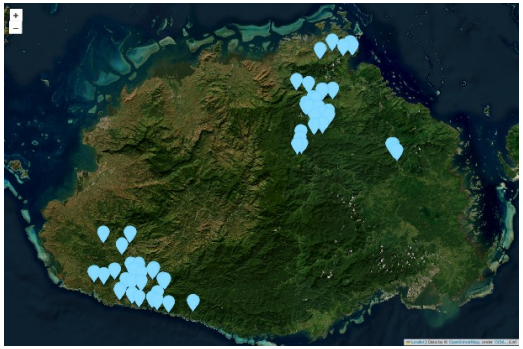
推論

- 目的関数にガウス過程とそこからのロジスティック回帰の両方が含まれている
  → 共役ではないので、 Pólya-Gamma補助変数法 (Scott 2013) を使用
- 効率的なGibbs samplerが得られる
- 観測点の数が多い場合があるので、最近傍ガウス過程 (Datta+ 2016, JASA) で効率的に近似
- 詳細は、我々のarXiv論文をご参照ください
  https://arxiv.org/abs/2409.01943

学習されたフィジー語の潜在因子



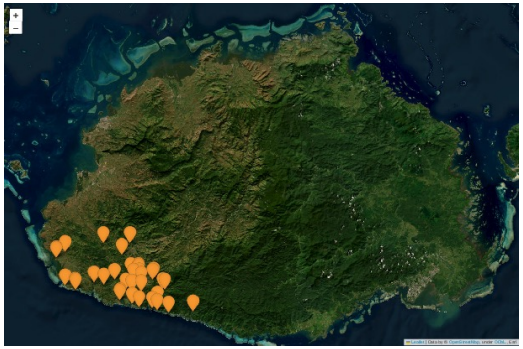
Factor 1



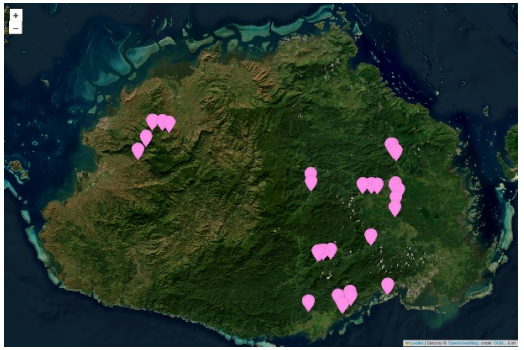
Factor 2

Factor 3

Factor 4



Factor 5

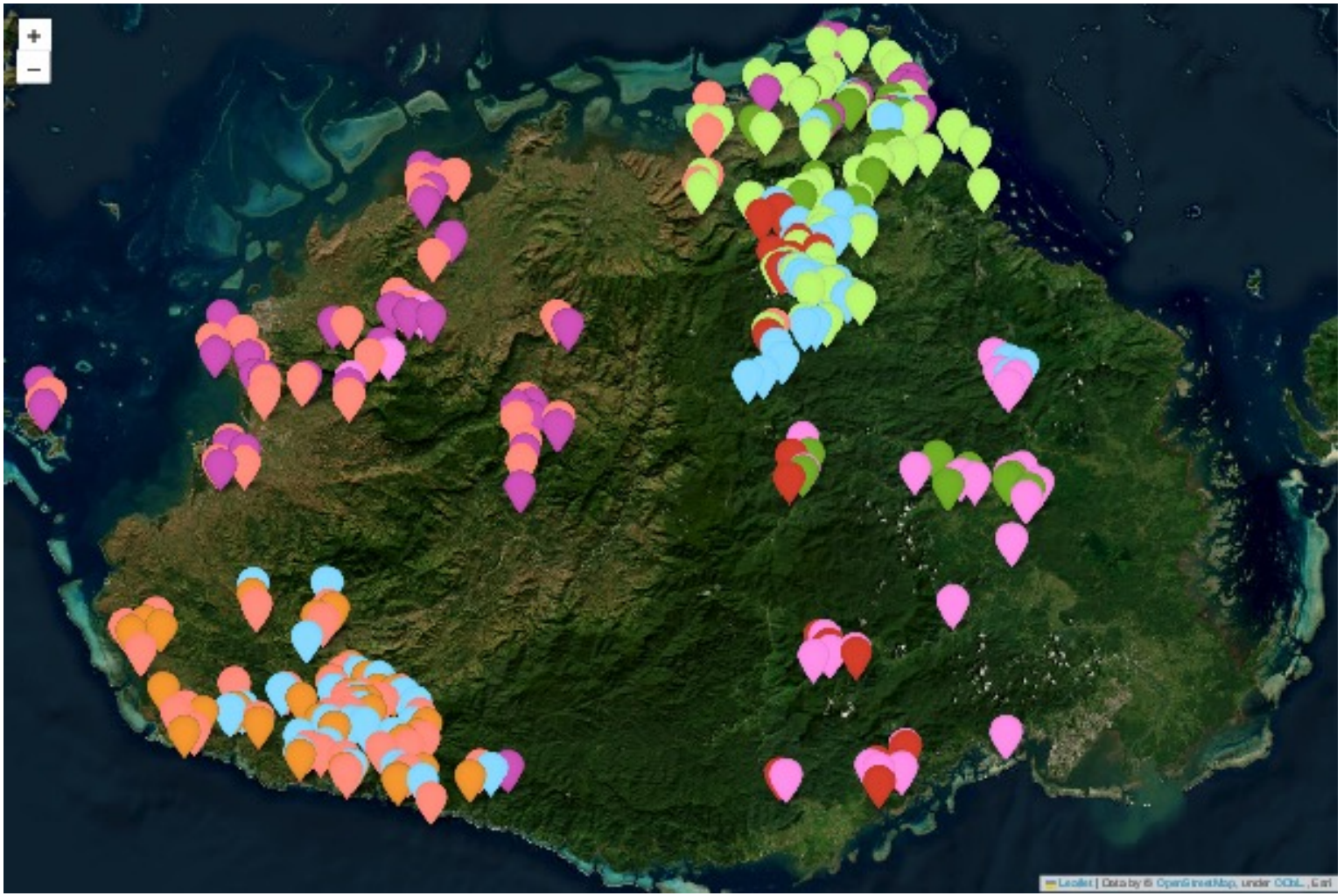


Factor 6

Factors 7,8

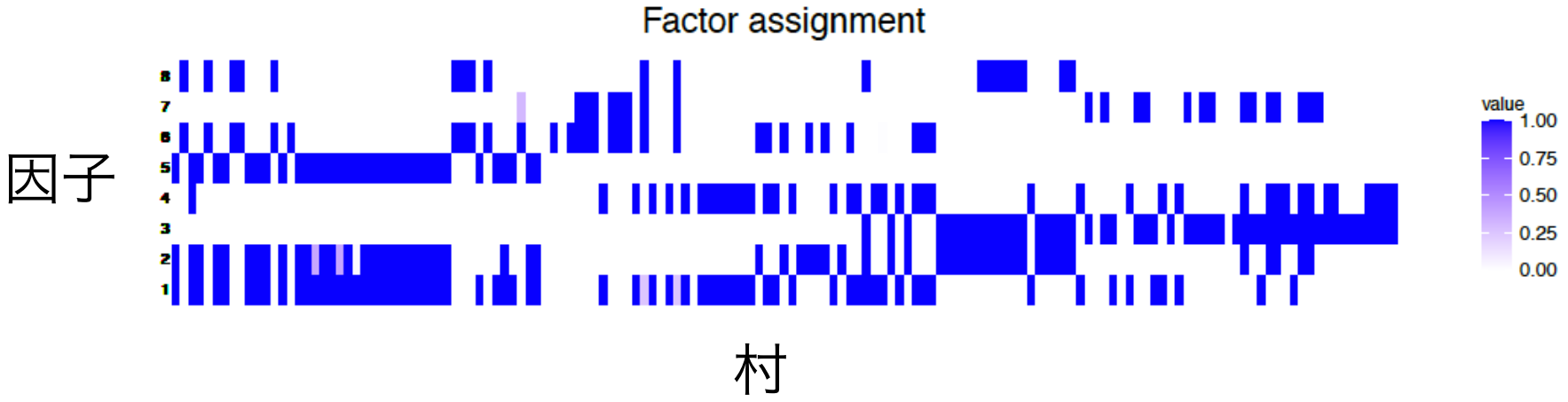
PythonのFoliumを使用してプロット

フィジー語の潜在因子 (全体表示)



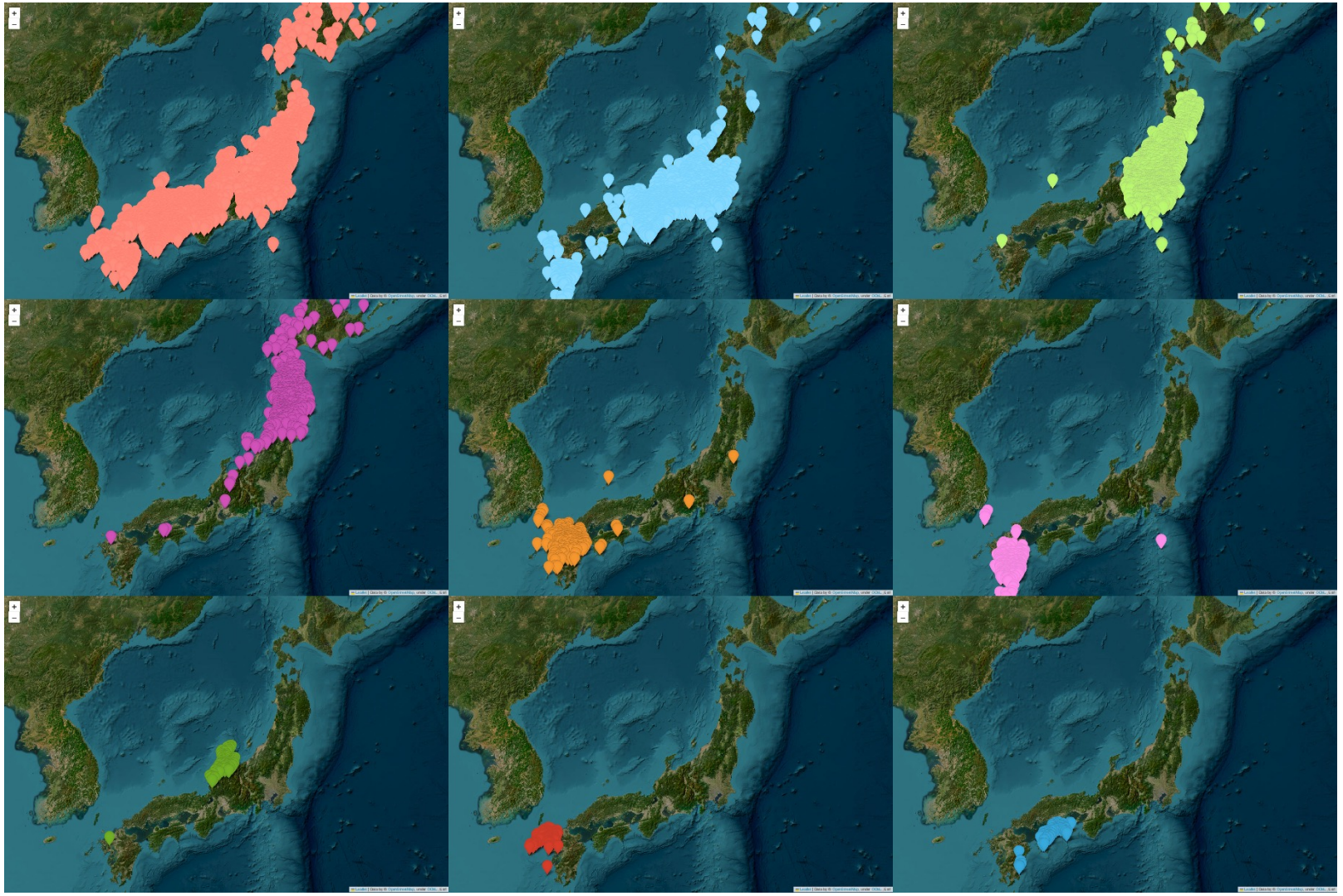
Using ArcGIS tiles in Python!

村と潜在因子の関係

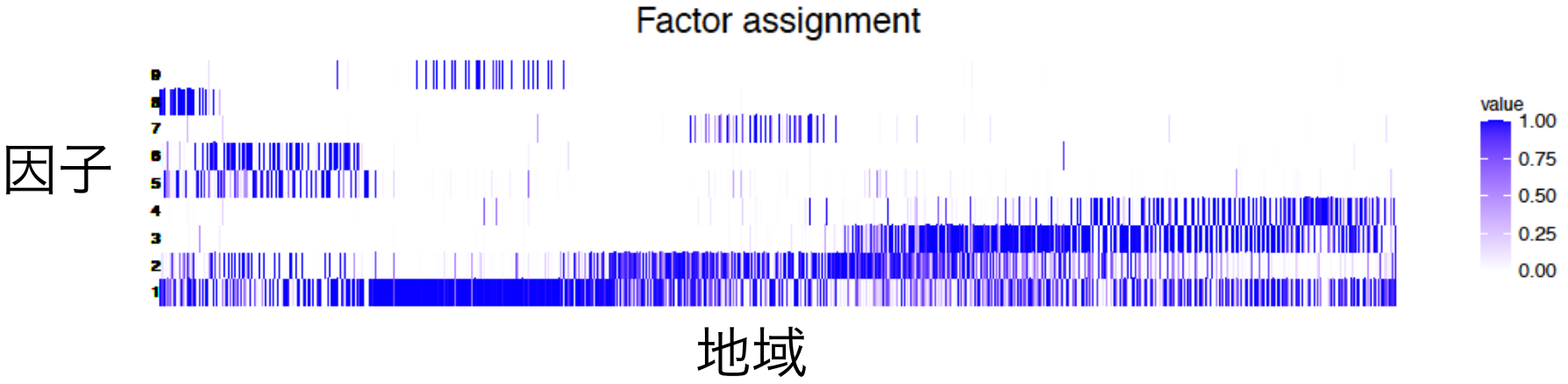


- それぞれの村が、複数の潜在因子を持っている
  - 横軸の村はだいたい地理的に近い順になっているので、空間相関があることがわかる
  - “例外”も少しあるが、それらが複数の村で共有されている

「日本言語地図」で学習された潜在因子

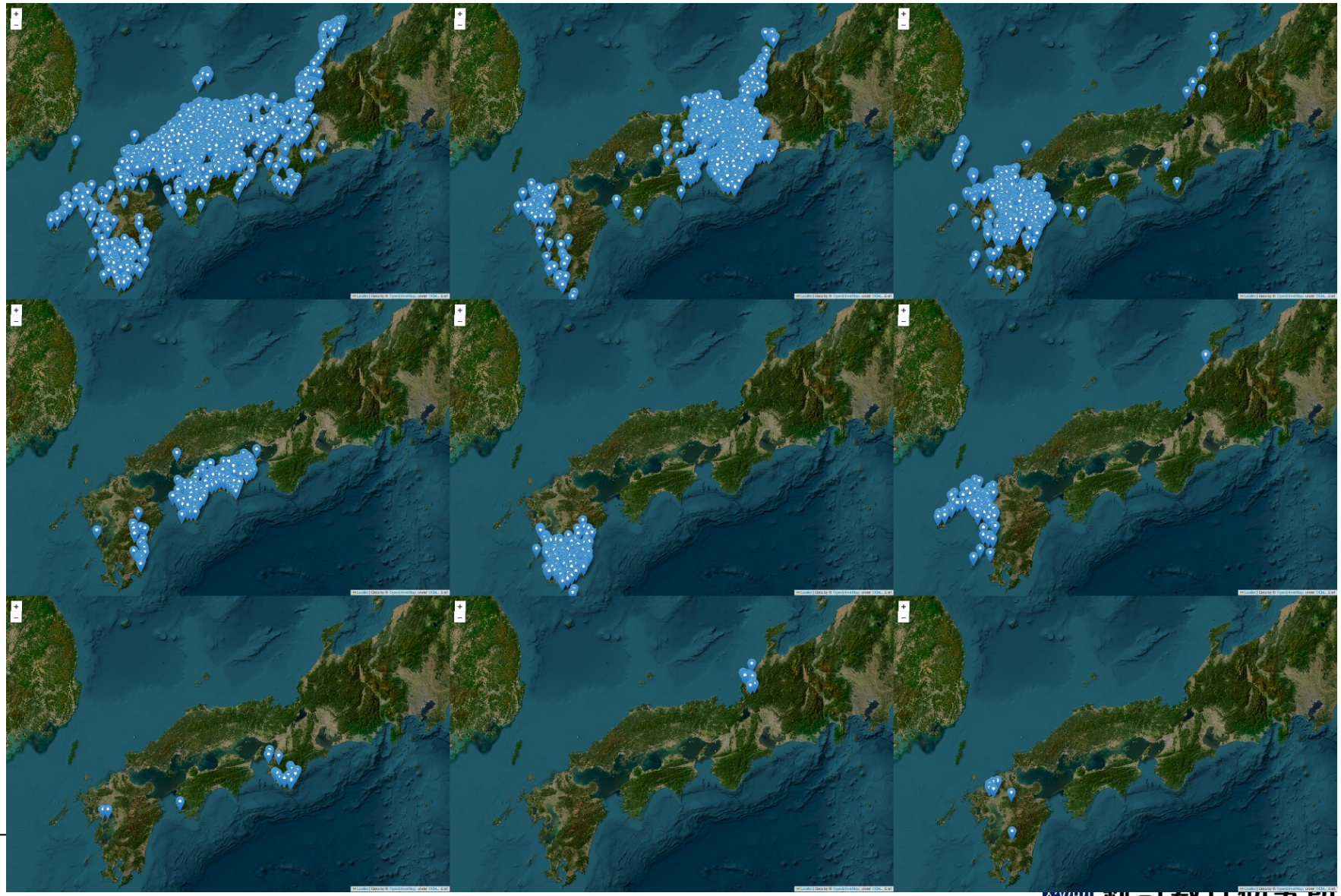


日本の地域と潜在因子の関係

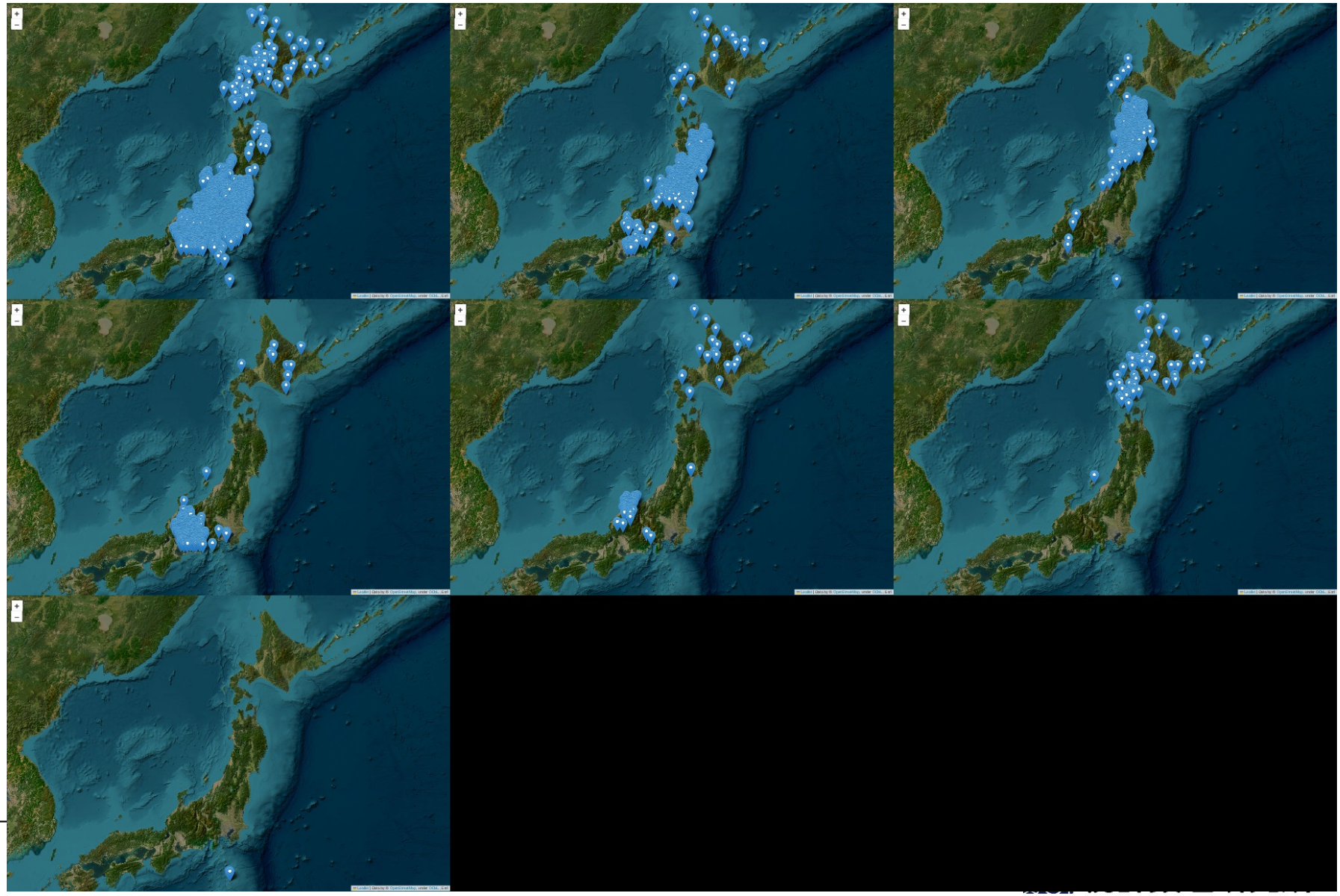


- 日本の各地域が、異なる潜在因子に確率的に所属している
- 植民された北海道、比較的単純な東日本、より歴史のある西日本の様子が読み取れる

「日本言語地図」の分析 (西日本)



「日本言語地図」の分析 (東日本)



まとめ

- 国立国語研究所などでも蓄積されてきた
  方言のデータを統計的にモデル化するために、
  空間的インド料理店過程を提案し、潜在的な
  因子を自動的に推定した
  - IBP+ガウス過程+ロジスティック回帰
- 統計的には、空間相関のある離散データを
  扱うための空間統計学の拡張
- 得られた因子の解釈には、これまでの言語学での
  知見と対照する必要がある (今後の課題)

おまけ

- 岩波書店より、春頃に発売予定
  http://chasen.org/~daiti-m/textmodel/