Latent Geographical Factors for Analyzing the Evolution of Dialects in Contact

Yugo Murawaki
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
murawaki@i.kyoto-u.ac.jp

Abstract

Analyzing the evolution of dialects remains a challenging problem because contact phenomena hinder the application of the standard tree model. Previous statistical approaches to this problem resort to admixture analysis, where each dialect is seen as a mixture of latent ancestral populations. However, such ancestral populations are hardly interpretable in the context of the tree model. In this paper, we propose a probabilistic generative model that represents latent factors as geographical distributions. We argue that the proposed model has higher affinity with the tree model because a tree can alternatively be represented as a set of geographical distributions. Experiments involving synthetic and real data suggest that the proposed method is both quantitatively and qualitatively superior to the admixture model.

1 Introduction

How languages have changed over time is a question that has attracted a lasting interest. Observing the present state of a language, we typically want to trace it back to the past. Historical–comparative linguists have done this by systematically comparing related languages and representing them as a tree. The success of this approach led to the establishment of language families such as Indo-European and Austronesian (Campbell, 2004). The recent adoption of computer-intensive statistical methods offer additional insights (Gray and Atkinson, 2003; Bouckaert et al., 2012; Chang et al., 2015).

When it comes to dialects, or closely-related languages,[1] the situation is very different. When we draw an *isogloss*, or the geographical boundary of a linguistic feature, and collect such isoglosses, it often happens that they conflict with each other (Kalyan and François, 2018). Conflicting isoglosses violate the assumption of the tree model, where after a branching event, two daughter languages evolve without any contact.

Nevertheless, some historical–comparative linguists have recently tried to apply the tree model to dialects in intense contact, with the assumption that at least some portion of observed data reflects tree-like vertical inheritance while the rest may result from horizontal contact (Lawrence, 2006; Pellard, 2009; Igarashi, 2017). While these efforts have been met with some success, it seems to us that the inherent difficulty in disentangling the two modes of transmission remains unresolved. This motivates us to turn to statistical modeling because computers are better at handling uncertainty than humans.

As a statistical model to analyze the evolution of dialects, admixture analysis has received attention in recent years (Bowern, 2012; Syrjänen et al., 2016; Cathcart, 2020). It assumes that each dialect is generated from a mixture of latent ancestral populations. Unfortunately, such ancestral populations can hardly be used for humans to infer a tree. Covering all the dialects with varying degrees of membership, an ancestral population only offers vague information about subgrouping if it does.

In this paper, we propose a probabilistic generative model that represents latent factors as geographical distributions (Figure 1). The geographical distribution of an observed feature is assumed to be stochastically generated from a weighted combination of the latent geographical factors. These factors are much more easier to interpret in the context of the tree model than latent ancestral population of the admixture model because an internal node of a tree can be geographically represented as the set of its descendant leaves. Some latent factors may be associated with vertical inheritance while others reflect horizontal transfer. We revisit this point in Section 5.3.

To evaluate the proposed method, we begin by

[1] The language/dialect distinction is not clear-cut. In this paper, the two terms are used interchangeably.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 959–976,
November 16–20, 2020. ©2020 Association for Computational Linguistics

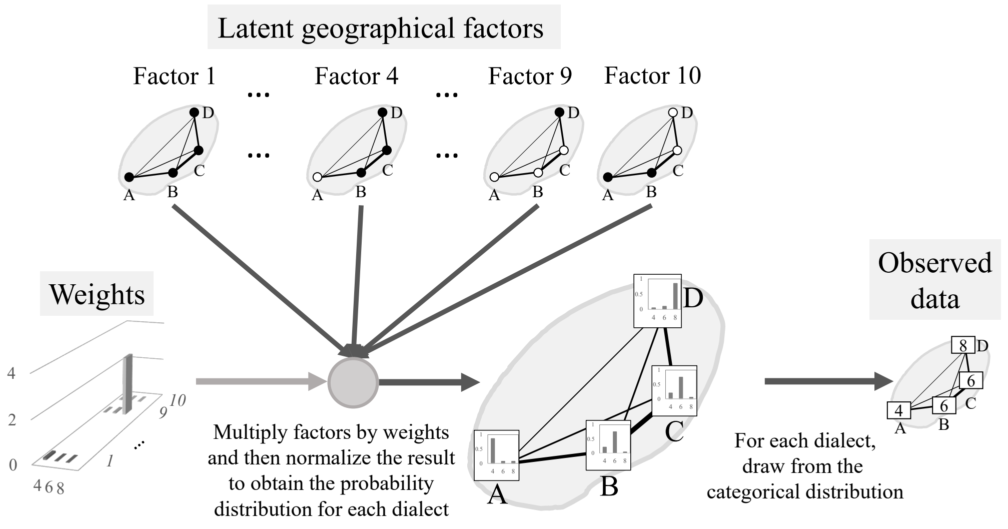

Latent geographical factors
Factor 1
Factor 4
Factor 9
Factor 10
A
B
C
D
Weights
4
2
0
4 6 8
1
9
10
Multiply factors by weights and then normalize the result to obtain the probability distribution for each dialect
Observed data
For each dialect, draw from the categorical distribution
4
6
6
8

Figure 1: An overview of the proposed method. There are $L = 4$ dialects, A, B, C, and D, on the island. The figure focuses on one of N features, for which each dialect takes the value 4, 6, or 8 (bottom right). The proposed method decomposes the observed data into $K = 10$ latent factors and the corresponding weights. Each latent factor has its own geographical distribution (top). Filled circles indicates the dialects are covered by the latent factor, while the dialects represented by hollow circles are not. Each feature value is tied to K weights (bottom left). Multiplying the binary factors by the weights and normalizing the resultant scores, we obtain a probability distribution for each dialect (bottom center). The value of each dialect is assumed to be drawn from the categorical distribution.

simulation experiments, where we know the ground truth. We demonstrate that the proposed method recovers tree-based and geographical clusters better than the admixture model. We then switch to a basic vocabulary database of Fijian dialects, whose evolutionary history is yet to be uncovered. We confirm that the proposed method detects major dialect groups. Although the proposed method in its current form focuses on spatial inference, the quantification it offers shows the potential of making temporal reasoning. The code is available at `https://github.com/murawaki/dialect-latgeo`.

2 Background

2.1 Dialectology

It is important to note that although we work on dialects, we methodologically lean toward historical–comparative linguistics. While historical–comparative linguistics is known for the Neogrammarian doctrine of *exceptionless* sound laws, dialectology is dominated by the dictum, "every word has its own history." In fact, the *Atlas linguistique de la France* (Gilliéron and Edmont, 1902–1910) and subsequent linguistic atlases that have been produced by dialectologists elaborate "the geography not of dialects but of linguistic traits" (Goebl, 2018).

Nevertheless, there have been several attempts in dialectology to aggregate over a large set of features (see Nerbonne and Wieling (2018) for an overview). Among the most popular ones are dimensionality reduction techniques such as principal component analysis (PCA) and multidimensional scaling (MDS). PCA is also routinely employed in population genetics to infer population structure from recombining genetic markers (Menozzi et al., 1978; Patterson et al., 2006). For visualization, each language is colored according to the value of a selected principal component (PC). In typical applications, at most the first three PCs are examined because subsequent PCs are hardly interpretable.

Recent applications of NLP techniques to dialectology and sociolinguistics (Eisenstein et al., 2010, 2014) make use of geotagged social media. While the big data allow us to analyze language variation and language change to the fine details, our interest lies in (1) applicability to unwritten languages and (2) language change on the order of hundred years or more. For these reasons, we work on data manually complied by field linguists.

2.2 Historical–Comparative Linguistics

Historical–comparative linguistics is characterized by careful manual selection of features (Sagart et al.

(2019) is a recent example). If two languages are phylogenetically closely related, they must be similar to each other, but not vice versa. It is because there are at least four ways to explain the fact that two languages share the same feature value: (1) inherited from a common ancestor (*vertical inheritance*), (2) borrowed from one language into another (*horizontal transfer*), (3) reflecting universal tendencies, and (4) coincidence. Only the first factor is a genuine phylogenetic signal. In order to establish phylogenetic relationships, linguists carefully count out features that have potential connections to the remaining three factors.

When three or more languages are involved, their subgroups need to be determined. To do so, linguists focus on *shared innovations* (Hoenigswald, 1966). A shared innovation is a change that occurred in an intermediate descendant from which a subset of modern languages have descended and that is not shared by the remaining languages. In other words, *shared retentions*, or feature values inherited from the common ancestor, are disregarded because they cannot be used as a criterion for subgrouping.

When the above-mentioned principle is applied to dialects in intense contact, an even more stringent feature selection is performed.[2] For example, Lawrence (2006) and Pellard (2009) discard a set of regular sound changes in favor of a conflicting *irregular* sound change, arguing that the former is more likely to occur in parallel (i.e., universal tendencies). However, they appear to have so much trouble distinguishing vertical inheritance from horizontal transfer. In addition, a large number of discarded features must constitute an important aspect of evolutionary history that awaits description. For these reasons, we choose a setting where no manual feature selection is performed. At this stage of research, our model is agnostic as to which factor has led to the current distribution of a given feature although we are much interested in tying some of the latent factors to the tree model.

Igarashi (2017) manually searched for *matryoshka*-like geographical distributions of shared innovations to construct a phylogenetic tree of dialects, with the assumption that if the distribution of one innovation is nested inside that of another, it reflects a branching event within the tree. We concur with his idea that spatial inference forms the basis for temporal reasoning. We note that an innovation that occurred in the past is not necessarily directly observable because it can be overshadowed by subsequent changes. As a probabilistic model, the proposed method has the potential to recover the original pattern given that it is supported by other observed features.

2.3 Admixture Analysis

Originally borrowed from population genetics (Pritchard et al., 2000; Alexander et al., 2009), what we collectively refer to as admixture analysis has been employed in recent studies on dialects (Bowern, 2012; Syrjänen et al., 2016; Cathcart, 2020). The same technique was also used to analyze typological features (Reesink et al., 2009; Longobardi et al., 2013).

Like the more familiar latent Dirichlet allocation (LDA) (Blei et al., 2003), an admixture model assumes that each individual (document) is stochastically generated from a mixture of K ancestral populations (topics). A major difference is that while LDA ties a single vocabulary distribution to each topic, each ancestral population has N discrete distributions, one per feature type.

We argue that this is not a natural assumption for languages although it is for genetic data. A population (a collection of individuals) normally maintains multiple values for a genetic marker. In contrast, a speech community would have trouble communicating if it uses multiple values for a single feature (e.g., multiple words for a given concept). To guarantee efficient communication, a language must take a single value for each feature, except for transitional periods. To address this problem, Cathcart (2020) explicitly imposes sparsity on his model.

2.4 Phylogenetic Networks

When horizontal transfer is non-negligible, a network model is often used as an alternative to the tree. NeighborNet (Bryant and Moulton, 2004) is arguably the most famous implementation of the idea and has been applied to dialect data (Lee and Hasegawa, 2011; Saitou and Jinam, 2017).

However, it must be noted that NeighborNet does not explicitly indicate any single evolutionary scenario but simply visualizes multiple conflicting trees as a single network. Nichols and Warnow

[2] Ignoring the methodology of historical–comparative linguistics, Lee and Hasegawa (2011) applied a computer-intensive phylogenetic method to a lexical dataset of dialects. Not surprisingly, the resulting phylogenetic tree is judged totally unreliable by an expert linguist (Pellard, 2018).

(2008) give warning against applying the model to dialects under intense contact.

3 Proposed Method

3.1 Basic Idea

The key insight behind the proposed method is that both vertical and horizontal signals can be represented as geographical distributions. If horizontal contact occurs in a certain area, leading to multiple feature values being shared by the dialects there, we can identify the corresponding geographical cluster. Similarly, a group of dialects that exclusively share the same ancestor usually occupies a continuum geographical space. Because their shared evolutionary history results in many shared feature values, the corresponding geographical subspace can be identified. Note, however, that we do not necessarily observe geographical distributions in their original forms because a state in the past can be overshadowed by subsequent changes. Therefore, our goal is to induct latent, typically clearer geographical factors from observed geographical distributions, as illustrated in Figure 1.

Each latent geographical factor is responsible for spreading certain feature values. Ideally, a binary variable should indicate the presence or absence of a feature value in the latent geographical factor. However, observed data are too complex and noisy to be explained by a deterministic generative process, and we want to reserve clear-cutness for latent geographical distributions. For these reasons, we introduce soft membership to feature values: A non-negative continuous weight indicates how strong the feature value is associated with the latent geographical factors.

3.2 Bayesian Generative Model

The proposed method is a Bayesian generative model that is based on the model of Murawaki (2019) even though at first glance, our task has little in common with that of Murawaki (2019). The differences between the two are summarized in Appendix A.

Formally, the observed data[3] are an $L \times N$ matrix X, where L is the number of languages and N is the number of features. Its element $x_{l,n}$ represents language l's n-th feature. Features are categorical and feature n takes one of F_n values.

We assume that X can be reorganized into an $L \times K$ binary matrix Z, where K is the number of latent factors and is specified a priori. The latent factor k is represented by the vector $\mathrm{z}_{*,k} = (z_{1,k}, \cdots, z_{L,k})$, in which $z_{l,k} \in \{0, 1\}$ indicates whether the latent factor k is active for language l.

Each latent factor has a geographical interpretation. Filled and hollow circles in the top of Figure 1 indicate one- and zero-valued $z_{l,k}$'s, respectively.

To incorporate our prior expectation that nearby languages are likely to take the same value for each k, we use an autologistic model (Besag, 1974; Towner et al., 2012). Relationships between languages are represented as a *neighbor graph*, which is indicated by edges between dialects in Figure 1.

We use a weighted variant of the graph. The probability of language l taking the value $b \in \{0, 1\}$, conditioned on the rest of the languages, $\mathrm{z}_{-l,k} = (z_{1,k}, \cdots, z_{l-1,k}, z_{l+1,k}, \cdots, z_{L,k})$, is

$$P(z_{l,k} = b \mid \mathrm{z}_{-l,k}, h_k, u_k) \propto \exp\left(h_k \sum_{l' \in \mathcal{G}(l)} \omega_{l,l'} I(z_{l',k} = b) + u_k b\right). \quad (1)$$

The parameter $h_k > 0$ controls the degree of influence from neighboring languages while $u_k \in (-\infty, +\infty)$ serves as a bias term. Their prior distributions are: $h_k \sim \text{Gamma}(\kappa, \theta)$ and $u_k \sim \mathcal{N}(0, \sigma^2)$. $\mathcal{G}(l)$ returns a set of l's neighbors and $\omega_{l,l'} > 0$ indicates how strongly the pair is connected. Both $\mathcal{G}(l)$ and $\omega_{l,l'}$ are given a priori. Eq. (1) encodes our assumption that the more neighboring languages take the value b, the more likely language l also takes the value b.

This model is called an *auto*logistic model because the target variable $z_{l,k}$ depends on explanatory variables of the same kind, $z_{l',k}$'s. To solve the chicken-and-egg problem, we define a joint distribution, $P(\mathrm{z}_{*,k} \mid h_k, u_k)$ (Besag, 1974).

The generation of Z is followed by that of the weight matrix $W \in \mathbb{R}_{>0}^{K \times M}$, where $M = \sum_{n=1}^{N} F_n$. Suppose that feature n's i-th value corresponds to the m-th weight. We map the two indexes using $f(n, i) = m$. An element of W, $w_{k,m}$, is drawn from $\text{Gamma}(1, 1)$.

Next, we compute $\tilde{\Theta} = ZW \in \mathbb{R}_{\geq 0}^{L \times M}$ and then normalize $\tilde{\Theta}$ for each feature n using the softmax function:

$$\begin{aligned} \theta_{l,f(n,i)} &= \text{softmax}_i(\tilde{\theta}_{l,f(n,1)}, \cdots, \tilde{\theta}_{l,f(n,F_n)}) \\ &= \frac{\exp(\tilde{\theta}_{l,f(n,i)})}{\sum_{i'=1}^{F_n} \exp(\tilde{\theta}_{l,f(n,i')})}. \end{aligned} \quad (2)$$

[3] To be precise, a language can have missing features.

Finally, $x_{l,n}$ is drawn from the corresponding categorical distribution:

$$x_{l,n} \sim \text{Categorical}(\theta_{l,f(n,1)}, \cdots, \theta_{l,f(n,F_n)}). \tag{3}$$

To see how Z and W affect the generation of X, we should note that $\theta_{l,f(n,i)}$ indicates how likely language l takes the value i for feature n. Recall that $\tilde{\theta}_{l,f(n,i)}$, the unnormalized counterpart of $\theta_{l,f(n,i)}$, is calculated as

$$\tilde{\theta}_{l,f(n,i)} = \sum_{k=1}^{K} z_{l,k} w_{k,f(n,i)}. \tag{4}$$

If $z_{l,k} = 0$, the latent factor k has no effect on $\theta_{l,f(n,i)}$; otherwise $w_{k,f(n,i)}$ raises the probability of language l's taking the value i for feature n. Let $\tilde{\theta}_{*,f(n,i)} = (\tilde{\theta}_{1,f(n,i)}, \cdots \tilde{\theta}_{L,f(n,i)})$. For each latent factor k, $w_{k,f(n,i)}$ is added to the vector $\tilde{\theta}_{*,f(n,i)}$, but zero-valued $z_{l,k}$'s mask the operation.

To complete the generative story, we define the joint distribution (hyperparameters are omitted for brevity):

$$P(A,Z,W,X) = P(A)P(Z|A)P(W)P(X|Z,W), \tag{5}$$

where $A = (H,U)$, $H = (h_1, \cdots, h_K)$ and $U = (u_1, \cdots, u_K)$.

3.3 Inference

Following Murawaki (2019), we use Gibbs sampling to perform posterior inference. Given observed values $x_{l,n}$, we iteratively update $z_{l,k}$, h_k, u_k, and $\mathrm{w}_{k,*} = (w_{k,1}, \cdots, w_{k,M})$, and missing values $x_{l,n}$.

Update $x_{l,n}$ $x_{l,n}$ is sampled from Eq. (3).

Update $z_{l,k}$ and $\mathrm{x}^{\text{mis}}_{l,*}$ We use the Metropolis-Hastings algorithm to update $z_{l,k}$ and $\mathrm{x}^{\text{mis}}_{l,*}$, the missing portion of $\mathrm{x}_{l,*} = (x_{l,1}, \cdots, x_{l,N})$. We find that jointly updating $\mathrm{x}^{\text{mis}}_{l,*}$ drastically improves the mobility of $z_{l,k}$.

Update h_k and u_k We want to sample h_k (and u_k) from $P(h_k \mid -) \propto P(h_k)P(\mathrm{z}_{*,k} \mid h_k, u_k)$. Since this belongs to a class of problems known as sampling from doubly-intractable distributions (Møller et al., 2006; Murray et al., 2006), we adopt an approximate sampler (Liang, 2010).

Update $\mathrm{w}_{k,*}$ We block-sample $\mathrm{w}_{k,*} = (w_{k,1}, \cdots, w_{k,M})$ using Hamiltonian Monte Carlo (HMC) (Neal, 2011).

4 Simulation Experiments

4.1 Synthetic Data

Evaluating the proposed method is a tough challenge. Here we turn to synthetic data. While rare in NLP, simulation is an established practice in evolutionary biology as a means of quantitatively evaluating statistical models.

Specifically, we consider a general scenario where dialects follow tree-shaped evolutionary paths but a high degree of borrowing obscures the phylogenetic signal. The resultant leaf nodes (modern dialects) are given to the proposed model to perform inference while the tree is used for evaluation. Our simulator is similar in spirit to the TraitLab software package extended with lateral transfer,[4] which is used extensively to test the robustness of the tree model with respect to contact phenomena (Greenhill et al., 2009; Kelly and Nicholls, 2017). There are, however, two important differences that make our simulation more realistic:

1. Instead of independently simulating the birth and death of each trait along branches, we group traits into features. Having a new trait born at a branch, we randomly choose a feature type and update the feature value of the dialect in question to the new one (i.e., the old value dies there).

2. We simulate spatial diffusion using a 2D Brownian random walk process. While the local borrowing variant of the TraitLab model makes dialects borrow traits from phylogenetically close dialects, we control the degree of borrowing according to spatial proximity.

We set the number of observed dialects to 50, the number of features to 100, and the root date to 1,000 BP (before present). The simulation was repeated 5 times using different random seeds. We removed features that had only one value (i.e., no variation) and merged dialects that were too similar to each other to be documented separately.

To make simulation experiments realistic, we tuned hyperparameters by manually checking neighbor-joining trees (Saitou and Nei, 1987) and NeighborNets (Bryant and Moulton, 2004) drawn from generated data, in addition to monitoring several statistics. We found that only a small subspace in the hyperparameter space led to realistic-looking data. As a result, we obtained 44.4 ± 2.6 languages

[4] https://github.com/lukejkelly/TraitLabSDLT

and 92.4 ± 3.0 features with 524.6 ± 86.8 unique values.

To assess how realistic the synthetic data were, we checked the δ score (Holland et al., 2002). Ranging from 0 to 1, the δ score indicates how tree-like the data are (lower is more tree-like). We obtained the score of 0.246 ± 0.057, which was roughly comparable to those calculated from real datasets known for non-tree-like evolution (Murawaki, 2015).

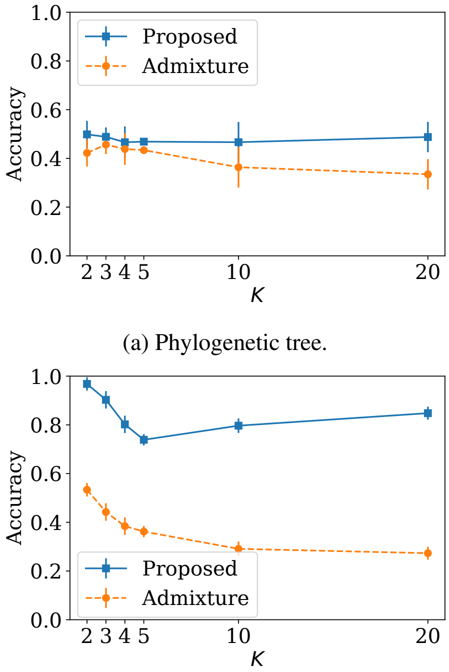
4.2 Model Settings and Evaluation Metric

We compared the proposed method with an admixture model. The settings for the proposed method is described in detail in Appendix B. We implemented a simple, fully Bayesian variant of admixture analysis, which is explained in Appendix C. For both models, we varied the number of latent factors, K, to be 2, 3, 4, 5, 10, and 20.

As the evaluation metric, we used a variant of many-to-one mapping accuracy. The induced latent factors were compared against gold standard clusters, and more than one latent factor may be mapped to the same gold standard cluster. Each latent factor was first mapped to the gold standard cluster that had the highest similarity score. We used the Jaccard index as the similarity score. The accuracy was then obtained by averaging each latent factor's score.

We considered two types of gold standard clusters: (1) phylogenetic tree and (2) spatial hierarchical clustering. For the ground-truth phylogenetic tree, each node was mapped to the set of its descendant leaves, and it was used as a gold standard cluster if it covered at least 10% of the leaves. We also conducted spatial hierarchical clustering using the UPGMA algorithm with the Euclidean distance, and generated clusters in the same manner.

Although the proposed method assumes clear-cut latent geographical distributions, posterior inference entails uncertainty about membership. To determine hard membership, we applied the threshold of 0.5 to the posterior probability $P(z_{l,k} \mid -)$. Obtaining clusters with hard membership from the admixture model is non-straightforward because it assumes soft membership by design. For each l, we averaged the ancestral population assignment $z_{l,n}$ over N and over posterior samples, and applied the threshold of 0.2. We confirmed that changing the threshold did not have much impact on accuracy.



(a) Phylogenetic tree.

(b) Spatial hierarchical clustering.

Figure 2: Many-to-one mapping accuracy of the induced latent factors, with varying K.

4.3 Results

The results are shown in Figure 2. We can confirm that the proposed method consistently outperformed the admixture model. The proposed method was particularly better at recognizing spatial patterns. It is understandable given that the geography is explicitly encoded to the proposed method while it is ignored by the admixture model.

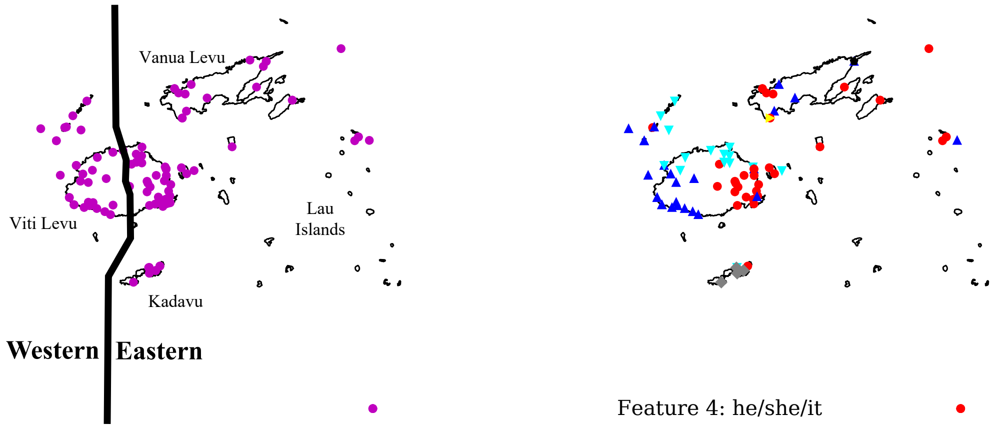
For the admixture model, the accuracy dropped more noticeably as K increased. In contrast, the proposed method retained a relatively high accuracy even with $K = 20$. It used additional latent factors to capture minor but genuine patterns.

5 Analysis of Real Data

5.1 Fijian Basic Vocabulary Database

Next, we analyzed a dataset of Fijian dialects, which was originally collected by Paul Geraghty and is in process of digitization by the Fijian Language GIS Project.[5] The details of the dataset will be published in the near future. We combined a lexical dataset with coordinate data. For each dialect, the dataset contains word form(s) that describe each of 100 basic concepts. Coordinate data were based on the Fiji Map Grid system, where the

[5] https://fijigis.github.io/



(a)

(b)

Figure 3: Linguistic maps of Fiji. (a) Locations of Fijian dialects, with an approximate boundary between Eastern and Western Fijian. (b) An example of features (seven more are shown in Figure A.1). The shape and color of a language indicates the value it takes. We can see that the feature value indicated by cyan down-pointing triangles (word form *ka*) transgresses the east–west boundary.

x- and y-axes correspond to local horizontal, and local vertical coordinates, respectively. As a result of preprocessing described in Appendix D.1, we obtained data with $L = 106$ and $N = 97$. The δ score was 0.286.

Figure 3(a) shows the locations of Fijian dialects in the dataset. It is well known that two major dialect groups, Eastern and Western Fijian, are demarcated by a boundary crossing the largest island of Viti Levu (Geraghty, 1983). As exemplified by Figure 3(b), however, features do not necessarily align with the boundary. Although Geraghty (1983) proposed multiple subgroups of Fijian by identifying shared innovations, he refrained from constructing a phylogenetic tree, arguing that they were likely to have resulted from intense contact. In short, no ground-truth is known for Fijian language history.

5.2 Qualitative Analysis

Due to lack of gold standard for the Fijian data, we chose to perform qualitative analysis. To do this, we first identify several desiderata for a model: (d1) intuitive geographical visualization of patterns, (d2) identification of Eastern and Western Fijian, (d3) identification of many more common patterns, and (d4) identification of conflicting patterns.

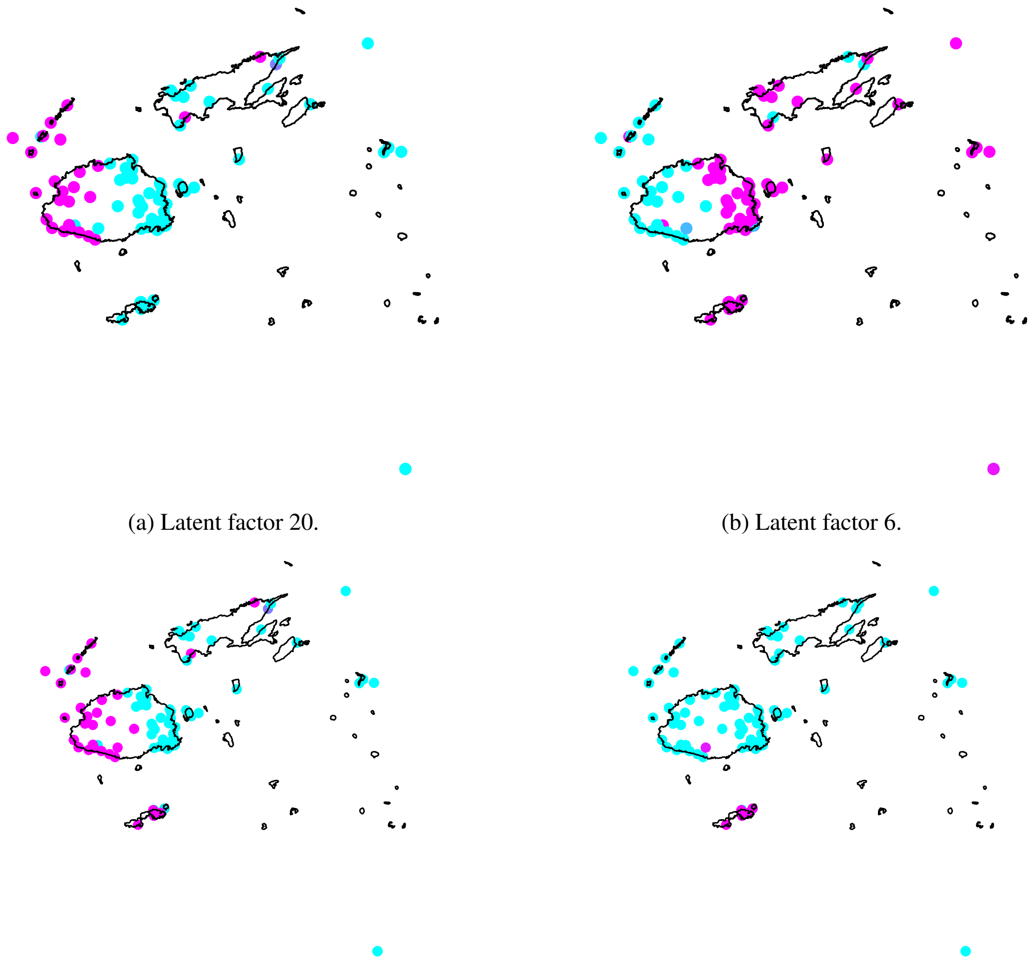
We performed posterior inference in the same manner as in Section 4. Figures 4 and A.3 visualize latent factors induced by the proposed method ($K = 20$). The visualization is intuitive (d1) and latent factors 20 and 6 (Figures 4(a–b)) correctly identified Western and Eastern Fijian, respectively (d2). At the same time, latent factor 2 (Figures 4(c)) covers Western Fijian and Kadavu in the southwest, transgressing the the east–west boundary (d4).

Impressionistically, other latent factors also appear to capture genuine patterns (d3), but the proposed model's superior performance with respect to desideratum 3 becomes more apparent when it is compared against other methods (Appendix D.2). Most importantly, admixture analysis was interpretable only with $K \leq 4$. Indeed, it is a standard practice in admixture analysis that K is carefully incremented from 2 until the output becomes uninterpretable. Confirming the result of the quantitative evaluation, the proposed method had no problem with $K = 20$. Although how to determine the optimal number of latent factors is an unresolved question, the proposed method safely allows us to try a large K. It is also worth noting that in admixture analysis, ancestral populations obtained with different Ks are routinely compared although they cannot necessarily be aligned in a consistent manner. In contrast, the proposed method does not necessitate incremental exploration.

In summary, only the proposed method satisfied the four desiderata at the same time. Although this does not necessarily guarantee the correctness of the model, we believe that the proposed method is worth further exploration.

5.3 Discussion

Our ultimate goal is to uncover spatio-temporal dynamics of languages although in this paper we



(a) Latent factor 20.

(b) Latent factor 6.

(c) Latent factor 2.

(d) Latent factor 19.

Figure 4: The visualization of four latent factors induced by the proposed method ($K = 20$). Other eight latent factors are shown in Figure A.3. The warmest color indicates that the latent factor k is active for language l ($z_{l,k} = 1$) while the coolest color corresponds to the opposite ($z_{l,k} = 0$). Intermediate colors indicate uncertainty.

concentrate on spatial inference. How does the proposed method provide a basis for temporal reasoning? To gain a toehold on this question, recall that the proposed method piles up multiple, potentially conflicting geographical clusters for each feature n. Since their relative strengths are controlled by $w_{k,f(n,i)}$'s, we expect that in case of conflict, a newer feature value gets a larger weight to supersede an older one.

Figure 6 shows a portion of the weight matrix W corresponding to feature 4 in Figure 3(b). We can see that although we did not explicitly impose sparsity on W, the overwhelming majority of elements in it were close to zero.

The feature value indicated by gray diamonds (word form *i*) was used by many, but not all, dialects on the southwestern island of Kadavu. Not surprisingly, this group gave the largest weight to latent factor 19, which also concentrated on Kadavu (Figures 4(d)). Interestingly, this conflicted with the feature value indicated by red circles (word form *e*) because it assigned a relatively large weight to latent factor 18, which covered Kadavu in addition to southeastern Viti Levu, Vanua Levu and some other small islands (Figure A.3(b)). However, latent factor 19 for *i* had a much larger weight than latent factor 18 for *e*, and as a result, the former overwhelmed the latter.

This seems to suggest that *e* was once widely used in Kadavu but was later replaced by *i*. Needless to say, however, a different run of the model may provide a different interpretation. We need to devise a statistical measure to quantify how likely the hypothesis is.

At this stage of research, temporal reasoning is left to human interpretation. Can we directly incor-

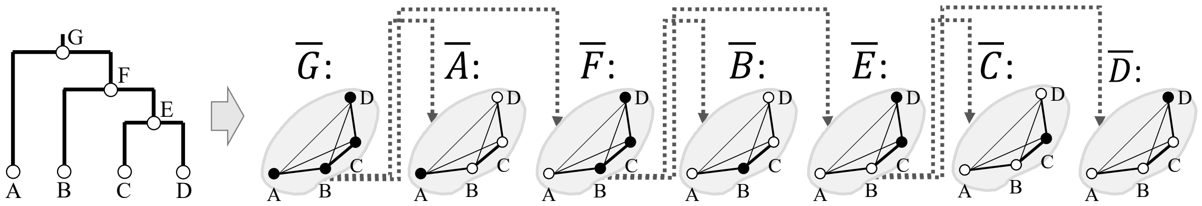


Figure 5: A geographical representation of a phylogenetic tree. We assume that the four modern dialects in Figure 1 have followed evolutionary paths shown on the left. We label internal nodes as E, F, and G. Each node X in the tree can be uniquely mapped to the set of its descendant leaves, which we denote as $\bar{X}$. A dotted arrow corresponds to a branch in the tree.

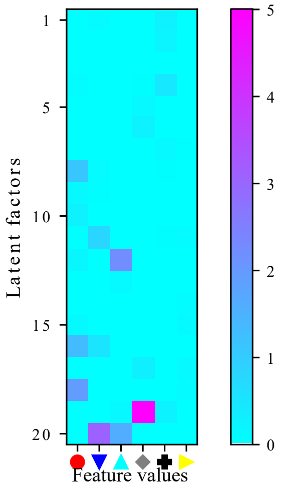


Figure 6: A portion of the weight matrix W corresponding to feature 4 in Figure 3(b).

porate it to the model? A hint is given in Figure 5. A node in a tree can be reinterpreted as a latent factor of the proposed method because it can be mapped to the set of its descendant leaves. All we have to do is to force the set of latent factors to satisfy the tree constraint: The two sets of active dialects in the children are a partition of the set of active dialects in their parent. As such, the proposed model has the potential of incorporating the tree model. With additional latent factors that are outside of the tree, the extended model can straightforwardly capture contact phenomena.[6]

Incorporating the tree constraint into the proposed method, especially as a hard constraint, is highly challenging. It is because each latent factor alone forms so complex a network that we resort to approximate sampling (Møller et al., 2006; Murray et al., 2006). However, this extension deserves further investigation. If a trait is observed in geographically fragmented regions and the possibility of parallel innovation is ruled out, linguists assume that it once had a wider geographical distribution connecting them. The proposed method in its present form has no mechanism to favor such a scenario, but the tree constraint does.

A caveat is that the proposed model does not keep track of the birth and death (or replacement by a new trait) of traits but lets multiple layers of historical changes simply pile up. This means that the state of an ancestral node cannot be reconstructed. This limitation appears inevitable especially if we want to model both vertical inheritance and horizontal contact, because it is hard to date contact events relative to an ancestral node.

6 Conclusions

In this paper, we proposed a Bayesian generative model to analyze dialectal variation. With this model, we successfully induced a large number of latent factors from a set of noisy surface features. Each latent factor is associated with an intuitively appealing geographical interpretation.

In the experiments, we used synthetic data and Fijian lexical data. Future directions include the incorporation of phonological and morphosyntactic features, application to other languages, and most importantly, a model extension to infer temporal ordering.

Acknowledgments

We thank Paul Geraghty, John Lowry, Ritsuko Kikusawa, Susumu Okamoto, Fumiya Sano, and the *itaukei ni vosa* (owners of the language) for providing the Fijian lexical/GIS data. We also thank Chundra A. Cathcart for valuable comments and suggestions on an earlier draft of this paper. This work was partly supported by JSPS KAKENHI Grant Numbers 18K18104 and 18KK0012.

[6] To analyze typological data, Daumé III (2009) presented a mixture model of a phylogenetic tree and a set of areal clusters. Although we share similar motivations with Daumé III (2009), our key idea is to represent vertical and horizontal signals in a unified manner, rather than given them completely different representations.

References

David H. Alexander, John Novembre, and Kenneth Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664.

Julian Besag. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236.

Balthasar Bickel, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga, and John B. Lowe. 2017. The AUTOTYP typological databases. version 0.1.0.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

Claire Bowern. 2012. The riddle of Tasmanian languages. *Proceedings of the Royal Society B: Biological Sciences*, 279(1747):4590–4595.

David Bryant and Vincent Moulton. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21(2):255–265.

Lyle Campbell. 2004. *Historical Linguistics: An Introduction (2nd edition)*. Edinburgh University Pres.

Chundra A. Cathcart. 2020. A probabilistic assessment of the Indo-Aryan inner-outer hypothesis. *Journal of Historical Linguistics*, 10(1):42–86.

Will Chang, Chundra Cathcart, David Hall, and Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 91(1):194–244.

Hal Daumé III. 2009. Non-parametric Bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 593–601.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLOS ONE*, 9(11):1–13.

Paul A. Geraghty. 1983. *The History of the Fijian Languages*. University of Hawai'i Press.

Jules Gilliéron and Edmond Edmont, editors. 1902–1910. *Atlas linguistique de la France*. Champion. (in French).

Hans Goebl. 2018. Dialectometry. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, pages 123–142. John Wiley & Sons.

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439.

Simon J. Greenhill, Thomas E. Currie, and Russell D. Gray. 2009. Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society B: Biological Sciences*, 276(1665):2299–2306.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101:5228–5235.

Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. 2005. *The World Atlas of Language Structures*. Oxford University Press.

Henry M. Hoenigswald. 1966. Criteria for the subgrouping of languages. In Henrik Birnbaum and Jaan Puhvel, editors, *Ancient Indo-European Dialects*. University of California Press.

B. R. Holland, K. T. Huber, A. Dress, and V. Moulton. 2002. δ plots: A tool for analyzing phylogenetic distance data. *Molecular Biology and Evolution*, 19(12):2051–2059.

Yosuke Igarashi. 2017. Phylogenetic classification of Japanese dialects using a shared innovation-based cladistic method: A proposal for the Southern Japanese branch (including Ryukyuan) and the Eastern Japanese branch (including Hachijo). First Meeting on the Reconstruction of the Proto-language of Japanese–Ryukyuan Dialects and the Construction of a Phylogenetic Tree by Means of Comparative Linguistic Methods. (in Japanese).

Eppie R. Jones, Gloria Gonzalez-Fortes, Sarah Connell, Veronika Siska, Anders Eriksson, Rui Martiniano, Russell L. McLaughlin, Marcos Gallego Llorente, Lara M. Cassidy, Cristina Gamba, Tengiz Meshveliani, Ofer Bar-Yosef, Werner Muller, Anna Belfer-Cohen, Zinovi Matskevich, Nino Jakeli, Thomas F. G. Higham, Mathias Currat, David Lordkipanidze, Michael Hofreiter, Andrea Manica, Ron Pinhasi, and Daniel G. Bradley. 2015. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications*, 6.

Siva Kalyan and Alexandre François. 2018. Freeing the comparative method from the tree model: A framework for historical glottometry. *Senri Ethnological Studies*, 98:59–89.

Luke J. Kelly and Geoff K. Nicholls. 2017. Lateral transfer in stochastic Dollo models. *Annals of Applied Statistics*, 11(2):1146–1168.

Wayne Lawrence. 2006. On the subclassification of the Okinawan dialects. *The Okinawa Bunka*, 40(2):101–118. (in Japanese).

Sean Lee and Toshikazu Hasegawa. 2011. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proceedings of the Royal Society B: Biological Sciences*, 278(1725):3662–3669.

Faming Liang. 2010. A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022.

Johann-Mattis List, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi, and Robert Forkel. 2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, 3(2):130–144.

Giuseppe Longobardi, Cristina Guardiano, Giuseppina Silvestri, Alessio Boattini, and Andrea Ceolin. 2013. Toward a syntactic phylogeny of modern Indo-European languages. *Journal of Historical Linguistics*, 3(1):122–152.

Paolo Menozzi, Alberto Piazza, and Luigi Cavalli-Sforza. 1978. Synthetic maps of human gene frequencies in Europeans. *Science*, 201(4358):786–792.

Jesper Møller, Anthony N. Pettitt, R. Reeves, and Kasper K. Berthelsen. 2006. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Yugo Murawaki. 2015. Spatial structure of evolutionary models of dialects in contact. *PLoS ONE*, 10(7):1–15.

Yugo Murawaki. 2019. Bayesian learning of latent representations of language structures. *Computational Linguistics*, 45(2):199–228.

Yugo Murawaki and Kenji Yamauchi. 2018. A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features. *Journal of Language Evolution*, 3(1):13–25.

Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. 2006. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 359–366.

Radford M. Neal. 2011. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 113–162. CRC Press.

John Nerbonne and Martijn Wieling. 2018. Statistics for aggregate variationist analyses. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, pages 400–414. John Wiley & Sons.

Johanna Nichols and Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass*, 2(5):760–820.

Nick Patterson, Alkes L. Price, and David Reich. 2006. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190.

Thomas Pellard. 2009. *Ōgami: Éléments de description d'un parler du Sud des Ryukyu*. Ph.D. thesis, Ecole des Hautes Etudes en Sciences Sociales (EHESS). (in French).

Thomas Pellard. 2018. On phylogenetic classification and bifurcations of Japonic languages. Phylesis and the History of the Japonic Languages from Philological and Field Linguistic Perspectives. (in Japanese).

Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Ger Reesink, Ruth Singer, and Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biology*, 7(11):e1000241.

Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Sciences*, 116(21):10317–10322.

Naruya Saitou and Timothy A. Jinam. 2017. Language diversity of the Japanese Archipelago and its relationship with human DNA diversity. *Man in India*, 97(1):205–228.

Naruya Saitou and Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Morris Swadesh. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of American Philosophical Society*, 96:452–463.

Kaj Syrjänen, Terhi Honkola, Jyri Lehtinen, Antti Leino, and Outi Vesakoski. 2016. Applying population genetic approaches within languages: Finnish dialects as linguistic populations. *Language Dynamics and Change*, 6:235–283.

Mary C. Towner, Mark N. Grote, Jay Venti, and Monique Borgerhoff Mulder. 2012. Cultural macroevolution on neighbor graphs: Vertical and horizontal transmission among western north American Indian societies. *Human Nature*, 23(3):283–305.

Appendix A A Comparison between the Model of Murawaki (2019) and the Proposed Method

The proposed method is a Bayesian generative model that is based on the model proposed by Murawaki (2019) even though at first glance, our task has little in common with that of Murawaki (2019). Table A.1 summarizes key differences between the two models. The most obvious difference is scale. While Murawaki (2019) worked on languages around the world, our target is a group of closely related dialects that usually occupies a relatively small area of the globe.

The difference in scale leads to the difference in the choice of features. In order to compare any pair of languages, which may have no known phylogenetic relationships, one has a limited choice. For this reason, Murawaki (2019) used features of linguistic typology (Haspelmath et al., 2005; Bickel et al., 2017). In contrast, we have a wide range of options for comparing dialects. While we used lexical features in the experiments, phonological and morphosyntactic features can readily be incorporated into the model although these features may be more prone to parallel innovation.

Both models encode our assumption that languages related to each other in some way tend to take the same feature value. However, whereas Murawaki (2019) used two neighbor graphs, one for phylogenetic relations and the other for spatial relations, we only use a spatial neighbor graph. It is because we are interested in cases where no ground truth is available for the internal phylogenetic classification of the languages in question.

The weight matrix W, which connects latent and surface representations, also differs slightly. We constrain $w_{k,m}$ to be positive whereas in Murawaki (2019), $w_{k,m}$ can be negative. Negative weights are hard to interpret in our task because we assume that multiple layers of historical changes simply pile up.

Finally, we look at $\tilde{\Theta}$ from a different angle. Murawaki (2019) interpreted $\tilde{\Theta}$ row-wise (fixing language l and discussing how the feature values (n_1, i_1) and (n_2, i_2) depend on each other). On the other hand, we present a column-wise interpretation (fixing the feature value (n, i) and discussing how l's get their probabilities).

Appendix B Settings of the Proposed Method

We constructed the neighbor graph as follows. First, we connected any pair of languages that are within the distance of 300 km. The edge weight ω_{l_1,l_2} for the pair of languages l_1 and l_2 was then given as

$$\max(d_{l_1,l_2}/3, 1)^{-1/2}.$$

σ^2, the hyperparameter for u_k, was set to 5. Recall that h_k is drawn from $\mathrm{Gamma}(\kappa, \theta)$. We set $\kappa = \hat{h}/5$ and $\theta = 5$. This means that the gamma distribution had mean $\hat{h}$ and variance $5\hat{h}$. Using the Fijian data, we estimated $\hat{h}$ using the autologistic models for N surface features (Murawaki and Yamauchi, 2018), with the assumption that the range of the parameter for latent factors should not deviate too much from the range for surface features. Specifically, we tied a single single parameter h to N autologistic models, sampled h's using an MCMC algorithm, and calculated their geometric mean. As a result, we obtained $\hat{h} \approx 0.009$.

Before collecting posterior samples, we ran 1,000 burn-in iterations. Following Murawaki (2019), we applied simulated annealing to the sampling of $z_{l,k}$ and $\mathrm{x}_{l,*}^{\mathrm{mis}}$. For the first 100 iterations, the inverse temperature was increased from 0.1 to 1.0. After the burn-in iterations, we collected 100 samples, one per iteration.

Appendix C Admixture Model

We implemented a simpler version of admixture analysis (Pritchard et al., 2000; Alexander et al., 2009). While population geneticists have devoted much effort to make inference scale to large genetic data, linguistic data are so small that a naïve Markov chain Monte Carlo algorithm suffices.

The generative story of the admixture model is as follows:

1. For each ancestral population $k \in \{1, \cdots, K\}$:
   (a) For each feature type $n \in \{1, \cdots, N\}$:
      i. Draw a categorical distribution from a symmetric Dirichlet distribution $\phi_{k,n} \sim \mathrm{Dir}(\beta_n)$.
2. For each language $l \in \{1, \cdots, L\}$:
   (a) Draw a mixing proportion from a symmetric Dirichlet distribution $\theta_l \sim \mathrm{Dir}(\alpha)$.

	Murawaki (2019)	Proposed method
Target	Worldwide	Dialects
Linguistic domain	Typology	Lexicon
Neighbor graphs	2	1
Weight range	$(-\infty, +\infty)$	$(0, +\infty)$
Interpretation of $\tilde{\Theta}$	Row-wise	Column-wise

Table A.1: A summary of key differences between the model of Murawaki (2019) and the proposed method.

(b) Then for each feature type $n \in \{1, \cdots, N\}$:

   i. Draw an ancestral population assignment $z_{l,n} \sim \text{Categorical}(\theta_l)$.

   ii. Draw a feature $x_{l,n} \sim \text{Categorical}(\phi_{z_{l,n}})$.

We marginalize out $\phi_{k,n}$ and θ_l and run a collapsed Gibbs sampler (Griffiths and Steyvers, 2004) to draw posterior samples. In the experiments, we ran 1,000 burn-in iterations and after that, collected 500 samples, one per iteration. As routinely done in population genetics (Jones et al., 2015), we increased the number of ancestral populations, K, one by one, starting from 2.

Appendix D Fijian Dataset

D.1 Details of Preprocessing

The lexical data of Fijian dialects[7] covered 100 basic concepts. The list was inspired by but is not identical with Swadesh's famous list (Swadesh, 1952) since it was tailored to Fijian.

We converted word forms into categorical features. To do this, we adopted a sequence comparison tool named LingPy (List et al., 2018). For each concept, it automatically clustered word forms into cognate groups, to which we assigned unique numbers. We discarded 3 concepts that were covered by single cognate groups. This means that each language was represented as a sequence of 97 lexical features. Note that since the proposed method only requires features to be discrete, it can also deal with phonological and morphosyntactic features.

Finally, we removed languages for which we were unable to determine coordinates. As a result, we chose 106 languages for further analysis. The ratio of missing features was 17.4%.

The Fijian dataset is still a work in progress, and a finished version is expected to be published in the near future. Needless to say, automatic cognate detection was not without errors, and the alignment between lexical and coordinate data was a source of additional complications. Nevertheless, we believe that the result of preprocessing was good enough to evaluate the proposed method, even if it may be too early to draw Fijian-specific linguistic insights.

Figure A.1 visualizes the dataset. A high degree of contact is evident from the NeighborNet analysis (Bryant and Moulton, 2004).

D.2 Additional Analysis with Baseline Methods

In addition to NeighborNet, several baseline methods were used to analyze the Fijian dataset.

Isogloss bundles The map is partitioned using a Voronoi diagram. An edge is drawn between two nearby languages, and its width is proportional to the number of features over which they disagree. Thus, thick lines indicate major dialect boundaries.

PCA Principal component analysis maps languages into lower dimensions (Nerbonne and Wieling, 2018). We visualize the first two principal components (PCs).

Admixture The admixture model used in simulation experiments in Section 4.

In NeighborNet (Figure A.1(a)), we can recognize Eastern Fijian (right) and Western Fijian (left), and also some of their subgroups. However, it is not easy to draw insights from reticulations, except for the obvious fact that the tree model does not fit well. Also, since NeighborNet visualizes clusters without reference to location, it does not provide any intuitive geographical interpretation.

Isogloss bundles in Figure A.2(a) illuminated so many dialect boundaries that even the most important east–west boundary got buried. The result partly explains why dialectologists are reluctant to generalize.

As for PCA, the first PC shown in Figure A.2(b) clearly identified the east–west boundary. The in-

[7] Called *communalects* in Fijian language studies (Geraghty, 1983).

	NeighborNet	Isoglosses	PCA	Admixture	Proposed
Aggregate & geographical	✗	✓	✓	✓	✓
Detect the east–west boundary	✓	? ∼ ✓	✓	✓	✓
Detect many more factors	? ∼ ✓	? ∼ ✓	✗	✗∼ ✓	✓
Detect conflicts	✓	? ∼ ✓	✗	✗	✓

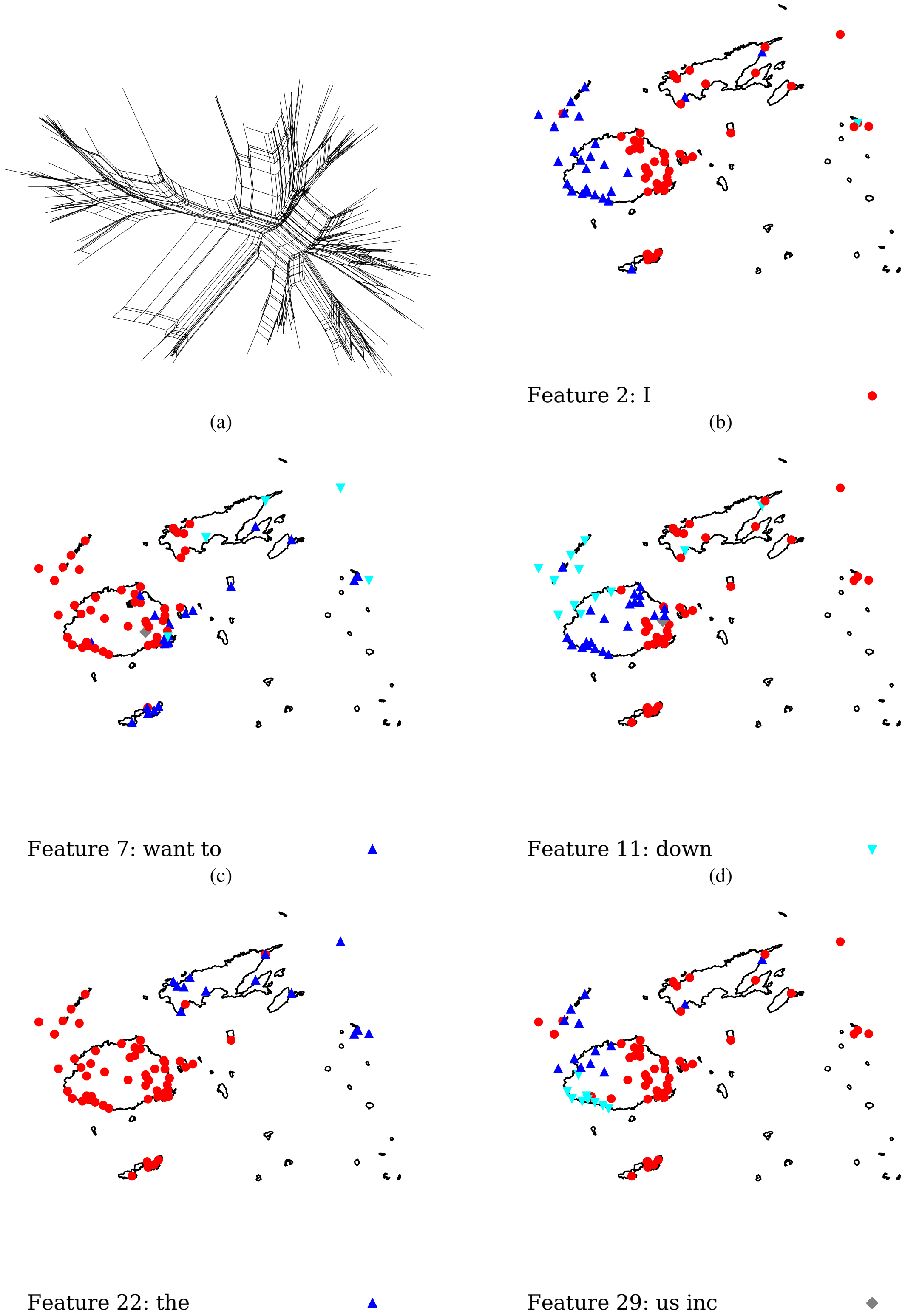
Table A.2: A summary of the comparison of various methods.

termediate colors found in the middle of Viti Levu suggest that the two groups are in contact. They explain why isoglosses were not clearly bundled together. It turns out, however, that PCA uncovered only one factor since the second PC, visualized in Figure A.2(c), discouraged any geographical interpretation.

In admixture analysis (Figure A.2(d–f)), each language is given a pie chart indicating the mixing proportion of ancestral populations. At first glance, admixture analysis generated a beautiful high-level picture of the dataset although the outputs with $K \geq 5$ were hard to interpret. With $K = 2$, it identified Eastern and Western Fijian, again with traces of contact in the middle of Viti Levu. With $K = 3$, Eastern Viti Levu was separated from the rest of Eastern Fijian, and with $K = 4$, Eastern Viti Levu was further divided into the northeast and the southeast.

However, a close examination reveals that admixture analysis went against our intuition. As Figure A.1 demonstrates, non-overlapping isoglosses were the norm in the dataset, but admixture analysis far too often assigned a single ancestral population to a language. We conjecture that most conflicts were absorbed by over-expressive ancestral populations and escaped detection.

Recall that in Section 5.2, we enumerated several desiderata: (d1) intuitive geographical visualization of patterns, (d2) identification of Eastern and Western Fijian, (d3) identification of many more common patterns, and (d4) identification of conflicting patterns. Based on the discussion above, we summarize the comparison of various methods in Table A.2. Now we can see that only the proposed method satisfies all of the four desiderata.



(a)

Feature 2: I ●

(b)

Feature 7: want to ▲

(c)

Feature 11: down ▼

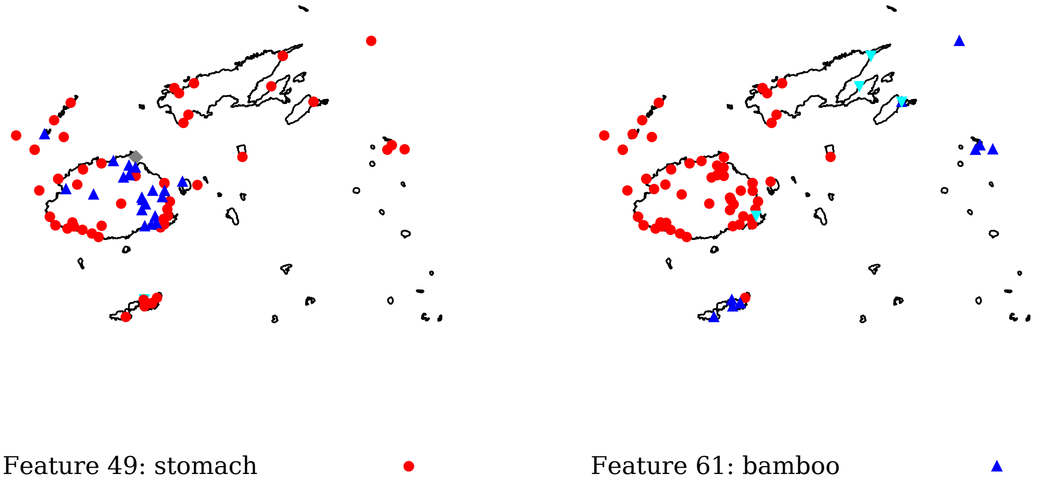
(d)

Feature 22: the ▲

(e)

Feature 29: us inc ◆

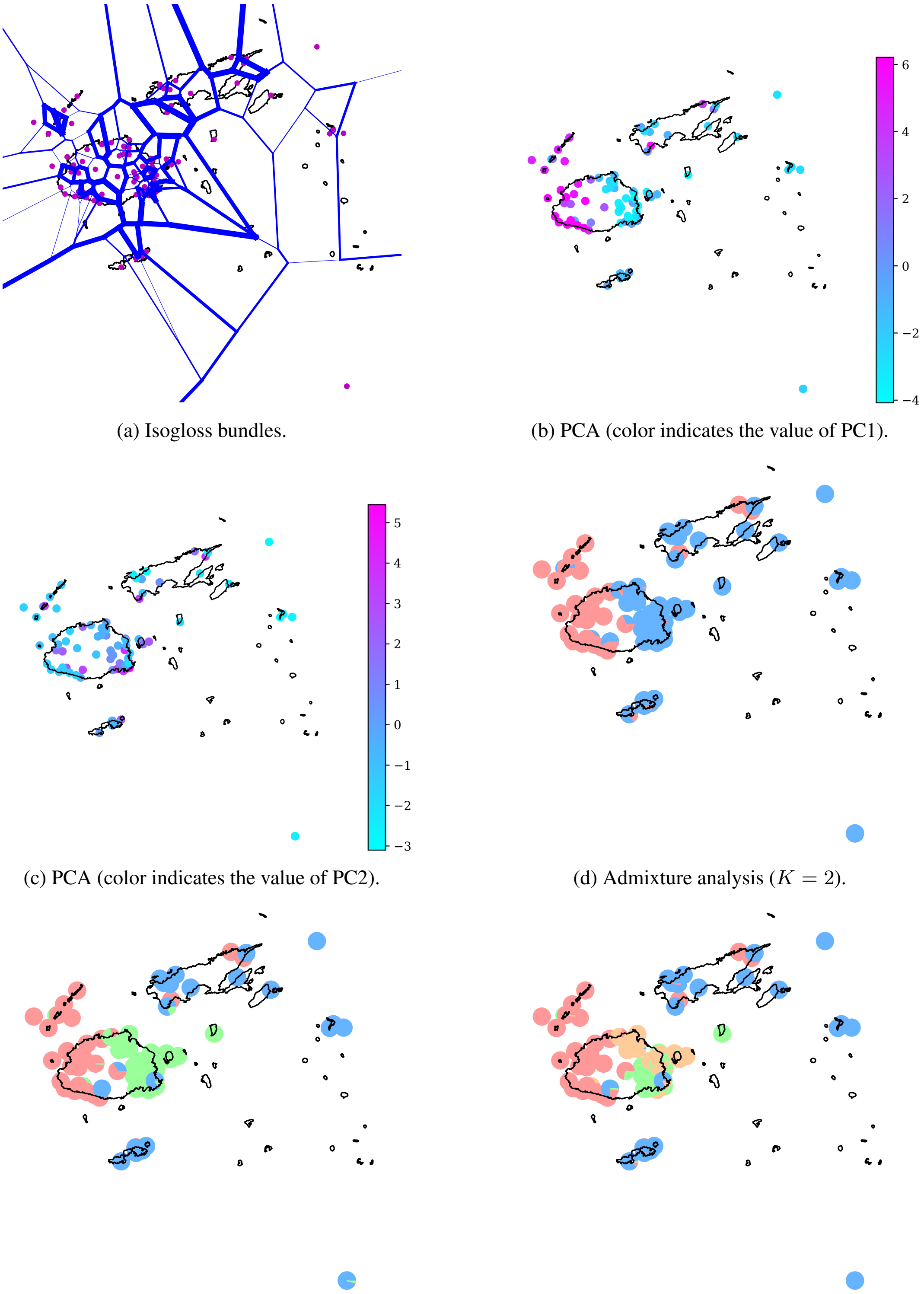
(f)



(g)

(h)

Figure A.1: (a) NeighborNet analysis visualizes the non-tree-like nature of the data. Leaves represent modern languages. Branch lengths are proportional to distances, and reticulations indicate conflicting signals. (b–h) Seven more examples of features, in addition to one shown in Figure 3(b). The shape and color of a language indicates the value it takes.



(a) Isogloss bundles.

(b) PCA (color indicates the value of PC1).

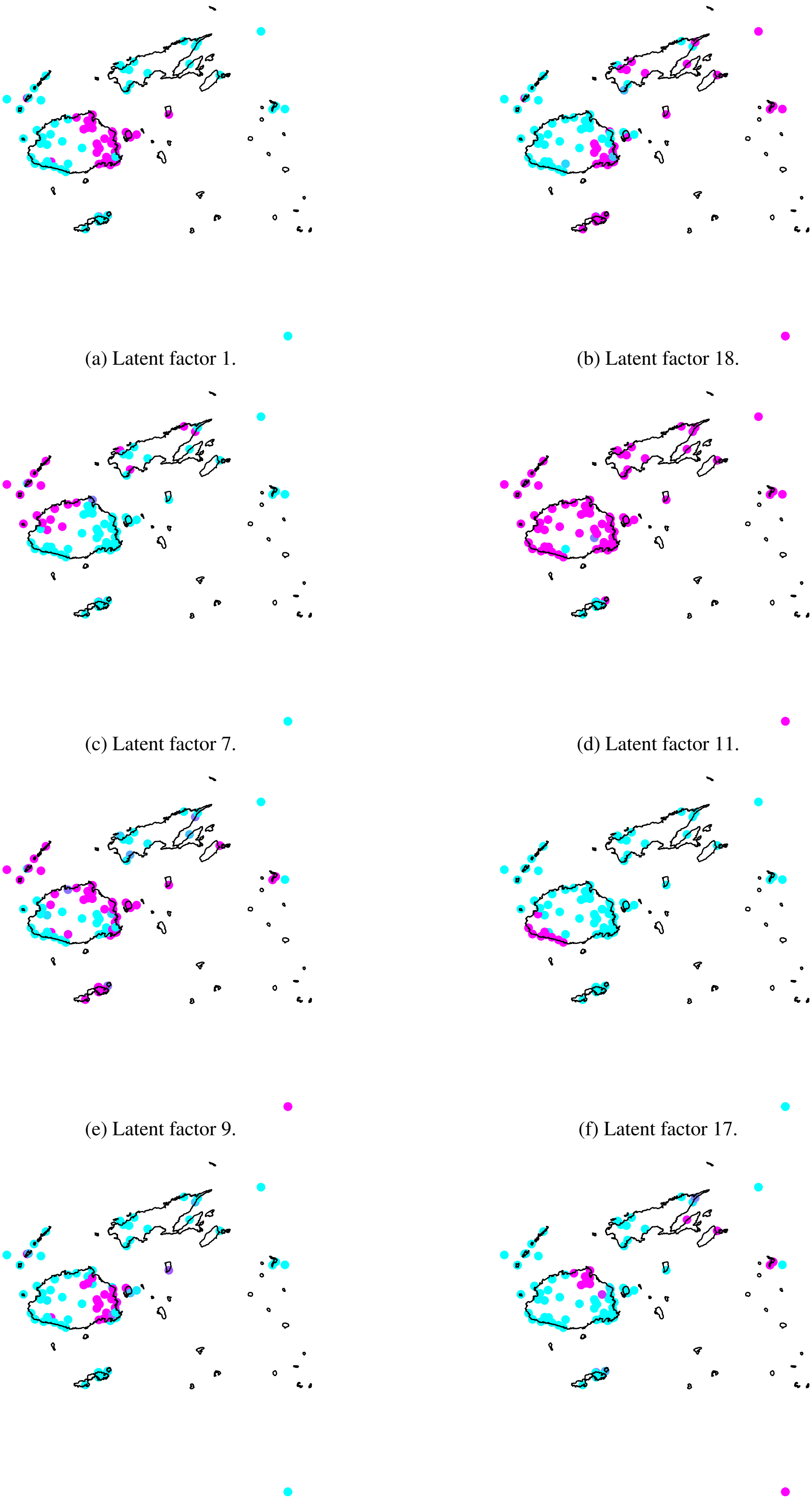
(c) PCA (color indicates the value of PC2).

(d) Admixture analysis ($K = 2$).

(e) Admixture analysis ($K = 3$).

(f) Admixture analysis ($K = 4$).

Figure A.2: The visualization of baseline methods.



(a) Latent factor 1.

(b) Latent factor 18.

(c) Latent factor 7.

(d) Latent factor 11.

(e) Latent factor 9.

(f) Latent factor 17.

(g) Latent factor 8.

(h) Latent factor 5.

Figure A.3: The visualization of eight latent factors induced by the proposed method ($K = 20$). Figure 4 visualized four other latent factors.