Inferring the Origin of Linguistic Features from an Atlas: A Case Study of Swiss-German Dialects.

Takuya Takahashi
Department of Geography, University of Zurich, Switzerland
NCCR Evolving Language, University of Zurich, Switzerland

Elvira Glaser
German department, University of Zurich, Switzerland
URPP Language and Space, University of Zurich, Switzerland

Peter Ranacher
Department of Geography, University of Zurich, Switzerland
NCCR Evolving Language, University of Zurich, Switzerland
URPP Language and Space, University of Zurich, Switzerland

Abstract

A linguistic atlas is a set of maps which visualize the geographical variation of linguistic features in a single language. We present a novel model for Bayesian statistics which infers when and where the variants of a linguistic feature were invented based on the geographical distribution shown in a linguistic atlas. Based on a spatial network representing the rate of diffusion between locations, our model evaluates the probability (likelihood) that the observed geographical pattern is realized by considering the genealogical relationship between variants at different locations. We apply our model to a linguistic atlas of Swiss-German dialects and infer the origin of three forms of the High-German word "nein" meaning "no".

2012 ACM Subject Classification Mathematics of computing → Bayesian computation

Keywords and phrases Dialectology, Linguistic geography, Geographic information science, Bayesian inference

Digital Object Identifier 10.4230/LIPIcs.COSIT.2024.18

Category Short Paper

Supplementary Material The code used for the case study is available on the Zenodo repository.
Software: `https://zenodo.org/records/13355152`

Funding *Takuya Takahashi*: JSPS Overseas Research Fellowship, NCCR Evolving Language with Swiss NSF Agreement No. 51NF40_180888 and the Swiss NSF Sinergia Project No. CRSII5_183578 (Out of Asia).
Peter Ranacher: NCCR Evolving Language with Swiss NSF Agreement No. 51NF40_180888.

Acknowledgements We thank Nico Neureiter for valuable discussion.

1 Introduction

It is estimated that approximately 7,000 languages exist worldwide, most of which contain a dialect-level variation in their linguistic properties such as phonology, lexicon, morphology, and syntax. The dialect-level variation of a single language has often been surveyed by fieldwork and recorded in the form of a linguistic atlas, a set of maps visualizing the geographical distribution of variants of a linguistic feature. So far, linguistic atlases have been created for many modern languages.

In dialectometry, researchers have profited from linguistic atlases to investigate the influence of geography on the variation of a language. Starting from the famous Séguy's curve [13], researchers have computed the linguistic distance between locations and investigated its relationship with the geographical distance, often finding a sublinear growth of the distance

16th International Conference on Spatial Information Theory (COSIT 2024).
Editors: Benjamin Adams, Amy Griffin, Simon Scheider, and Grant McKenzie; Article No. 18; pp. 18:1–18:9
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

between languages as a function of the distance in space [10]. Other previous studies in dialectometry include research into modeling the dialect borders [6] and investigating cultural or social factors underlying the variation of dialects [3]. More recently, methods borrowed from evolutionary biology were also applied to the spatial data of dialects, such as phylogenetic trees [8] and admixture analysis [11]. Many of the previous methods of spatial statistics applied to dialect data from linguistic atlases were aimed mainly at clustering dialects or studying the border or spatial transition between dialect areas. But can we use linguistic atlases to infer where and when a linguistic feature was invented?

In this article, we present a novel statistical method to infer the time and place of invention of a feature variant from its current geographical distribution recorded in a linguistic atlas. The method is an extension of the Bayesian framework by Takahashi and Ihara (2023)[14], which infers the genealogical tree of a focal cultural trait and its transmission and mutation. Their model employs a spatial network representing topography to retrace the genealogical (or phylogenetic) tree of the trait variants. Adding parameters representing the time and place of invention to their model, we infer when and where the focal feature variant arose on tree branches.

We apply the model to *Sprachatlas der deutschen Schweiz* (SDS), a linguistic atlas of Swiss-German dialects. We focus on a single map representing the geographical distribution of the word forms for “no” (High German “nein”) and infer the origin of each variant.

2 Model

2.1 Network model

We consider a spatial network with n nodes $P_1 \cdots P_n$, where P_i is the node with index i. The nodes represent locations in space, such as cities, villages, or, as in the SDS case study, survey locations. We use a discrete-time model, where at each timestep t, every node can be in one of two states, 0 or 1, indicating whether the focal variant is absent or present, respectively. So, the geographical distribution at any given time is represented by a n-dimensional binary vector.

2.2 Observed data

Focusing on a single linguistic feature, our model considers each of its variants as the focal variant in turn, assuming that they have evolved independently of each other. The observed data is the geographical distribution of the focal variant, represented by an n-dimensional binary vector $\mathbf{D} = \{D_1 \cdots D_n\}$, from which we aim to infer when and where the variant was invented.

2.3 Evolution of state values

At each timestep, the geographical distribution of the feature variant evolves following three rules: (1) the variant is invented, (2) the variant diffuses, (3) the variant is lost (Figure 1).

The invention of the focal variant (i.e., transition of the state value from 0 to 1) happens only once in history, in agreement with Dollo’s law. All occurrences of the focal variant in the present geographical distribution $\mathbf{D}$ are assumed to have the same origin. The time of invention is denoted by X_{time}, given as the number of timesteps before present, and the place as X_{space}, given as the index of the node. The present time is indexed by 0. All nodes have state value 0 (i.e., focal variant is absent) more than X_{time} timesteps ago, and all nodes except X_{space} have state value 0 at time $t = X_{time}$. The user must set the oldest possible time of origin τ, such that $0 \leq X_{time} \leq \tau$.

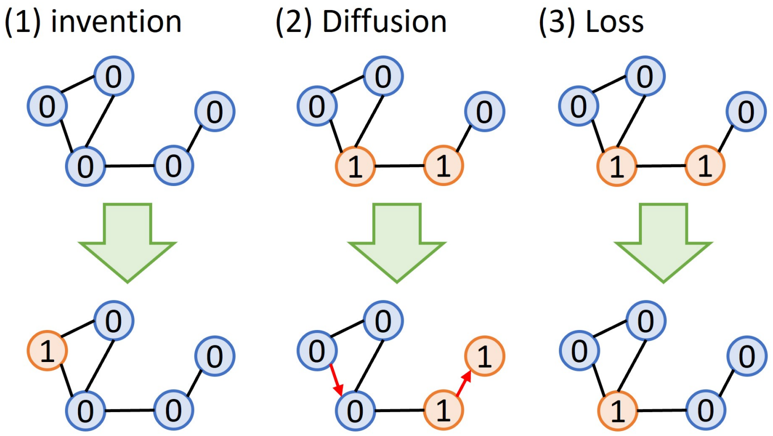


Figure 1 Evolution of state values in two consecutive timesteps.

As for diffusion, every node copies the state value of itself or one of its neighbouring nodes at the previous timestep. Formally, P_i copies the state value from the node P_j with probability (transmission rate) a_{ij}, where $\sum_{j=1}^{n} a_{ij} = 1$ for every i. In particular, a_{ii} represents the probability of copying from the same node, meaning the absence of diffusion. The value for a_{ij} must be defined by the user before inference. Note that the absence of the focal variant (state value 0) also diffuses in space, since another competing variant may also spread at rate a_{ij}, encroaching into the area originally occupied by the focal variant.

In addition to the diffusion event, each node may lose their variant with probability (loss rate) b, changing the state value from 1 to 0. The loss rate is a latent variable inferred simultaneously with the origin of the focal variant.

2.4 History of transmission

We extend the model by Takahashi and Ihara (2023)[14] and represent the history of a focal variant with the matrix $\mathbf{G}$ with dimension $\tau \times n$. The element g_{tj} of $\mathbf{G}$ represents the node from which P_j copied its state value $t-1$ timesteps ago (see Figure 2 for an example). In other words, $\mathbf{G}$ records all diffusion events between the oldest possible time of origin $t = \tau$ and the present time $t = 0$. Note that $\mathbf{G}$ is a random variable since the diffusion event is stochastic.

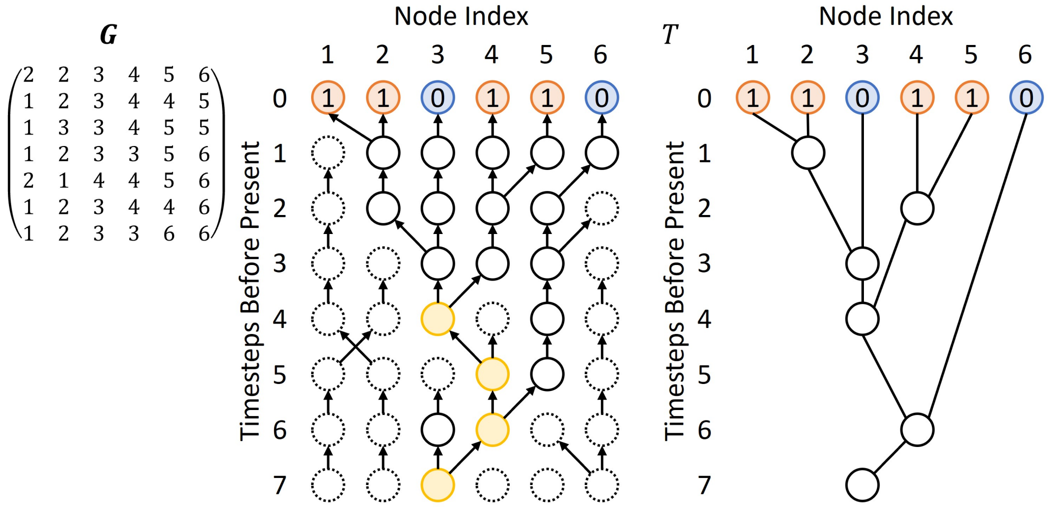
2.5 Bayesian inference

The observed variable $\mathbf{D}$ is probabilistically dependent on $\mathbf{G}$, b, X_{time}, and X_{space}. The joint posterior distribution of the model variables is given by

$$P(\mathbf{G}, b, X_{time}, X_{space}|\mathbf{D}) \propto P(\mathbf{G})\, P(b)\, P(X_{time})\, P(X_{space})\, P(\mathbf{D}|\mathbf{G}, b, X_{time}, X_{space}) \quad (1)$$

We sample from the posterior distribution using the Markov chain Monte Carlo (MCMC) algorithm by evaluating the right-hand side of 1.

First, the prior distributions $P(b)$, $P(X_{time})$ and $P(X_{space})$ are given by the user, reflecting specific domain knowledge. The prior $P(\mathbf{G})$ reflects the assumptions of the diffusion model: $P(\mathbf{G}) = \prod_{t=1}^{\tau} \prod_{j=1}^{n} P(g_{tj})$, such that $P(g_{tj} = k) = a_{jk}$ holds.



$$\boldsymbol{G} = \begin{pmatrix} 2 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 4 & 4 & 5 \\ 1 & 3 & 3 & 4 & 5 & 5 \\ 1 & 2 & 3 & 3 & 5 & 6 \\ 2 & 1 & 4 & 4 & 5 & 6 \\ 1 & 2 & 3 & 4 & 4 & 6 \\ 1 & 2 & 3 & 3 & 6 & 6 \end{pmatrix}$$

Figure 2 (Left) Example of matrix $\mathbf{G}$ with $\tau = 7$ and $n = 6$. (Center) The genealogy based on $\mathbf{G}$, with arrows representing transmission of state values. For example, we have $g_{25} = 4$, meaning that P_5 copied the state value from P_4 at timestep $2 - 1 = 1$, so we have an arrow starting at P_4 and ending at P_5 between timesteps 1 and 2. Circles with a solid line are past nodes upon which the observed state values are probabilistically dependent. The orange nodes represent possible time and place of invention. (Right) The genealogical tree T which is based on the matrix $\mathbf{G}$. The tree is rooted at timestep τ. The internal nodes show when and where the lineages split, which happens when multiple nodes copy the state value from a single node.

In computing the likelihood $P(\mathbf{D}|\mathbf{G}, b, X_{time}, X_{space})$, we trace the lineages representing the transmission of the focal variant over the τ timesteps. The process yields a genealogy represented by a tree T (see Figure 2 and also Figure 5 of [14]). T is a tree whose nodes are associated with temporal and spatial information, with leaves representing the nodes of the spatial network at the current timestep and internal nodes representing the time and place where the lineage split due to diffusion. The present geographical distribution $\mathbf{D}$ (i.e., observed data) in Figure 2 can only be realized if the values of (X_{time}, X_{space}) are either $(4,3)$, $(5,4)$, $(6,4)$ or $(7,3)$. In general, the invention must have occurred in an ancestor shared by all the nodes with the focal variant, as they are all descendants of the variant invented at $t = X_{time}$.

Here we illustrate the way to compute $P(\mathbf{D}|G, b, X_{time}, X_{space})$ based on Figure 2. Let $\mathbf{D}(t,i)$ denote the set of observed state values of the nodes which are descendants of the state value of P_i at time t. For instance, we have $\mathbf{D}(4,3) = \mathbf{D}(5,4) = \{D_1, D_2, D_3, D_4, D_5\}$ and $\mathbf{D}(6,4) = \mathbf{D}(7,3) = \{D_1, D_2, D_3, D_4, D_5, D_6\}$. Obviously, $\mathbf{D}(t,i)$ is a subset of the whole dataset $\mathbf{D}$. If $(X_{time}, X_{space}) = (4,3), (6,4)$, the invention happened at the internal node of T, in which case we can compute the likelihood as follows.

$$P(\mathbf{D}|\mathbf{G},\ b, (X_{time}, X_{space}) = (4,3)) = P(\mathbf{D}(4,3)\,|T,\ b,\ P_3\ \mathit{has\ state}\ 1\ \mathit{at}\ t = 4)$$
$$P(\mathbf{D}|\mathbf{G},\ b, (X_{time}, X_{space}) = (6,4)) = P(\mathbf{D}(6,4)\,|T,\ b,\ P_4\ \mathit{has\ state}\ 1\ \mathit{at}\ t = 6)$$

The conditional probability $P(\mathbf{D}(t,i)\,|T,\ b,\ P_i\ \mathit{has\ state}\ 1\ \mathit{at}\ t)$ in the right-hand side can be computed by Felsenstein's pruning [4], an algorithm to efficiently compute the probability that the state values at the tree leaves match the observed data. If $(X_{time}, X_{space}) = (5,4), (7,3)$, in order to apply the pruning algorithm, we look for the oldest internal node of

T, which is more recent than the timestep X_{time}. Let V_{time} and V_{space} respectively denote the timestep and index of the node where the lineage of the feature first split after invention. If $(X_{time}, X_{space}) = (5,4), (7,3)$, we have respectively $(V_{time}, V_{space}) = (4,3), (6,4)$. As the state value of the node indexed V_{space} at V_{time} is 1 with probability $(1-b)^{X_{time}-V_{time}}$, because the variant invented at X_{time} is lost with probability b at every timestep. Hence, we can compute the likelihood as follows.

$$P(\mathbf{D}|\mathbf{G},\ b,(X_{time},X_{space})=(5,4)) = (1-b)\,P(\mathbf{D}(4,3)\,|T,\ b,\ P_3\ \mathit{has\ state}\ 1\ \mathit{at}\ t=4)$$
$$P(\mathbf{D}|\mathbf{G},\ b,(X_{time},X_{space})=(7,3)) = (1-b)\,P(\mathbf{D}(6,4)\,|T,\ b,\ P_4\ \mathit{has\ state}\ 1\ \mathit{at}\ t=6)$$

To generalize the discussion above, the likelihood can be computed as follows.

$$P(\mathbf{D}|\mathbf{G},b,X_{time},X_{space}) = (1-b)^{X_{time}-V_{time}} P\left(\mathbf{D}(V_{time},V_{space})\,|T,b,P_{V_{space}}\ \mathit{has\ state}\ 1\ \mathit{at}\ t=V_{time}\right) \quad (2)$$

Note that variables X_{time} and X_{space} can be marginalized out by taking the summation of the above equation for possible pairs of X_{time} and X_{space}, in order to efficiently sample from the posterior distribution.

3 Case study

3.1 Data

Sprachatlas der deutschen Schweiz (SDS) is a linguistic atlas for over 1500 linguistic items surveyed at 565 locations in German-speaking Switzerland from 1939 to 1958. We used the digitalized dataset of SDS, available at `https://dialektkarten.ch/` [12], which shows the presence/absence of each feature variant in binary format. We focused on words for “no” (High-German “nein”) and applied our model to the geographical distribution of the three variants “nai”, “nei”, and “naa” (Figure 3). These three variants exhibit distinct geographical patterns, with “nai” occupying a wide region in the center, “nei” forming multiple distant geographical clusters, and “naa” being confined to a small area.

3.2 Spatial network

To apply our model to the SDS data, we treated each survey location as a node of the spatial network. The edge weight a_{ij} is given as a Gaussian function of the great circle distance, d_{ij}, between P_i and P_j, such that the probability that one location copies a variant from another decays exponentially with geographical distance. More specifically,

$$a_{ij} \propto exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right) \quad (3)$$

Note that a_{ij} is normalized so that its summation over j $(1 \leq j \leq n)$ equals 1. As the transmission between distant locations is extremely rare based on expression 3, this model describes the diffusion process through contacts between geographically close speakers.

3.3 Prior distributions

Based on linguistic expertise, we give priors on when the three variants were invented. It is considered that “nai” is a variant which derived from “nei” and that “naa” derived from one of the diphthong forms (most probably “nai”). We thus consider that the “nei” variant is the

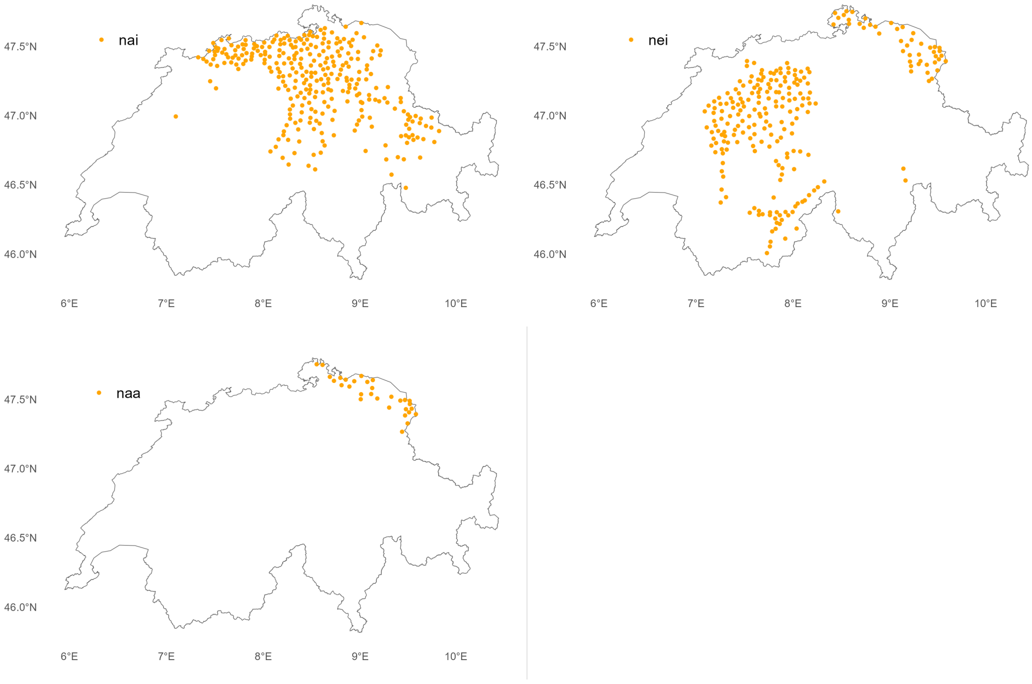


Figure 3 Present geographical distributions of the three variants. (top left) "nai" (top right) "nei" (bottom left) "naa".

oldest, whereas "nai" and "naa" are more recent innovations. We express our prior knowledge by setting $P\left(X_{time}\right) = U\left(300, 1000\right)$ for "nai" and "naa", and $P\left(X_{time}\right) = U\left(1000, 2000\right)$ for "nei", where $U\left(x, y\right)$ is a uniform distribution between x and y given in years before present.

As for the place of invention and the loss rate, we use the uninformative priors $P\left(X_{space}\right) = \frac{1}{n}$ and $P\left(b\right) = U\left(0, 1\right)$.

3.4 Other configurations

To establish the correspondence between the timestep and real time unit, we set one timestep of the model to be 20 years. The parameter in the Gaussian function σ in Expression 3 was set to 10 km. As for MCMC, we ran the Metropolis–Hastings algorithm and sampled parameters 10^3 times at the interval of 10^4 iterations, preceded by a burn-in period of 10^6 iterations.

3.5 Results

We sampled 1000 places of invention from the posterior distribution (Figure 4). A comparison with the current distribution of the variants (Figure 3) suggests that the place of invention is centered at the modern geographical distribution. Although the current distribution of the "nei" variant shows multiple distant clusters, the result suggests that the origin of this variant is most plausibly in Espace Mittelland or central Switzerland.

The results suggest that "naa", whose geographical distribution is confined in a small region, is likely to be the youngest variant. The median posterior of the time of invention is 980 years before present for "nai", 1840 for "nei" and 580 for "naa" (Table 1).

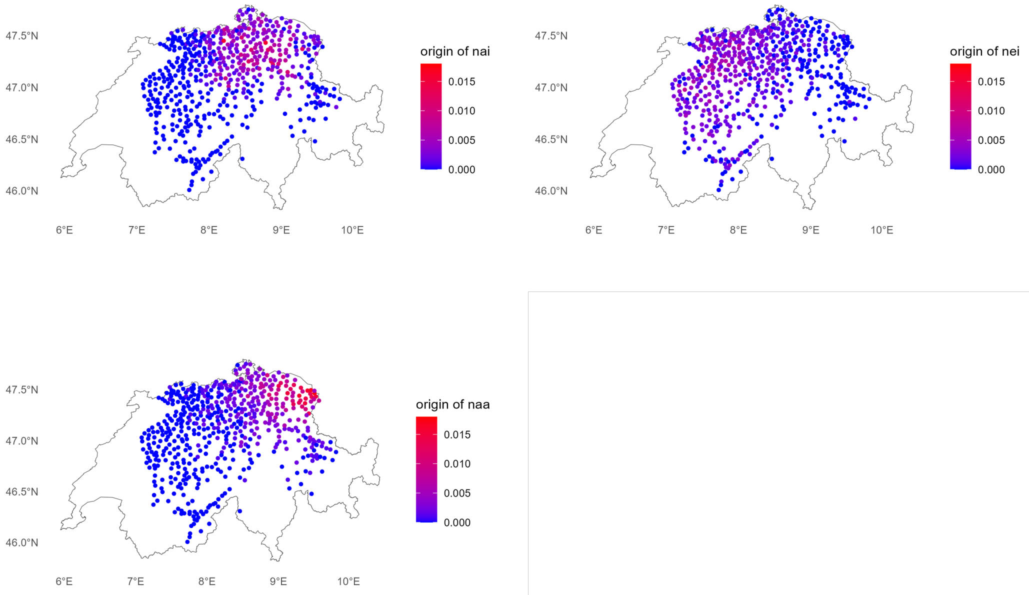


Figure 4 Place of invention sampled from the posterior distribution $P\left(X_{space}|\mathbf{D}\right)$. The value indicates the proportion of each survey location being sampled as the origin of the variant. This value approximates the posterior probability. (top left) “nai” (top right) “nei” (bottom left) “naa”.

Table 1 Posterior distribution of the time of origin $P\left(X_{time}|\mathbf{D}\right)$. The values are displayed in years before present.

variant	min.	first quartile	median	mean	third quartile	max.	95% CI
“nai”	720	940	980	956	1000	1000	840 - 1000
“nei”	1040	1700	1840	1799	1940	2000	1380 - 2000
“naa”	300	440	580	601	740	1000	300 - 960

4 Discussion

In this paper, we presented a novel model, to infer the time and place where a linguistic feature was invented, primarily tailored to data found in linguistic atlases. The model is Bayesian, enabling users to integrate prior knowledge in linguistics and geography to further narrow down the possible time period and the area of invention.

Our approach is based on a tree model, which represents the genealogical relationship between the copies of features which we observe today, so we consider similarities and differences between the current study and other tree models. The most related field is thus phylogeography, in which researchers have reconstructed the phylogenetic tree of pathogens [9] and languages [1][7], simultaneously inferring the time and location of ancestors. The phylogeographic approach has been applied to infer the origin of language families such as Indo-European [1] and Bantu [7]. The biggest difference is that, whereas standard phylogeographical models consider the position of language as a point location, our model considers the presence of linguistic features at every survey location of a linguistic atlas. Our model is thus expected to fully profit from the geographical pattern of the linguistic feature. In the other vein, while our model considers the diffusion of a single feature variant, phylogeographical models regard a language as a set of linguistic features and assume that all these features share the same diffusion pathway. The phylogenetic models can thus incorporate more data to infer the diffusion process. However, as it has been already pointed out by Takahashi and Ihara (2023) [14], considering frequent linguistic borrowing at the dialect level, modeling the diffusion of each linguistic feature separately may be appropriate in dialectology.

On a related note, one of the current limitations of the model is its assumption that variants of the same feature diffuse independently of each other. In reality, different lexical or phonological variants for the same concept compete, and it is rare for multiple variants to coexist. In addition, new variants may arise by altering an extant variant, which is true for our case study, where the "nai" and "naa" variants are thought to have derived from "nei". Analyzing the geographical distributions of multiple variants simultaneously should lead to more accurate inference. A potential direction for future study, therefore, is to extend the model so that the nodes are assigned a categorical variable with any number of possible states.

As for limitations concerning the case study, our dataset on the Swiss-German dialects may potentially be biased in two ways. First, the transcription of the phonetic data depends on explorers and the categorization of variants (symbols on the map) is an interpretation of the original data. Second, in the digitized dataset of SDS [12], some of the different symbols (variants) on the original maps are summarized for visualization purpose. Since phonetic variation is not discrete but continuous, the way to categorize the variants can be a source of bias, which may potentially violate our model assumption that the invention of a single variant took place only once in history.

Another limitation of the current model is that the special network must be defined in advance. In particular, the transmission rate (i.e., weight of edges) a_{ij} cannot be inferred from the data because of the convergence issue of the MCMC algorithm. In our case study on SDS data, we assumed that the transmission rate decays with geographical distance as a Gaussian function (i.e., expression 3). However, previous quantitative studies in dialectology suggested that the spatial variation of linguistic features in some regions were formed not only by contact between geographically continuous areas but also by migration and displacement of human populations [5]. To model the long-distance diffusion triggered by human migration, the transmission rate should perhaps be represented by a function with a long-tailed distribution, rather than a Gaussian function [2]. Since the mode of diffusion is different in each language and region, it would be definitely more practical if the model could infer the mode of diffusion which has formed the current geographical distribution of feature variants. Hence, future studies should conceive an algorithm to compute model selection measures like Bayes factor.

References

1 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. Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097):957–960, 2012. `doi:10.1126/science.1219669`.

2 James Burridge. Unifying models of dialect spread and extinction using surface tension dynamics. *Royal Society Open Science*, 5(1):171446, 2018. `doi:10.1098/rsos.171446`.

3 Curdin Derungs, Christian Sieber, Elvira Glaser, and Robert Weibel. Dialect borders—political regions are better predictors than economy or religion. *Digital Scholarship in the Humanities*, 35(2):276–295, June 2019. `doi:10.1093/llc/fqz037`.

4 Joseph Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, November 1981. doi:10.1007/BF01734359.

5 John LA Huisman, Asifa Majid, and Roeland Van Hout. The geographical configuration of a language area influences linguistic diversity. *PLoS ONE*, 14(6):e0217363, 2019. doi:10.1371/journal.pone.0217363.

6 Péter Jeszenszky, Philipp Stoeckle, Elvira Glaser, and Robert Weibel. A gradient perspective on modeling interdialectal transitions. *Journal of Linguistic Geography*, 6(2):78–99, 2018. doi:10.1017/jlg.2019.1.

7 Ezequiel Koile, Simon J. Greenhill, Damián E. Blasi, Remco Bouckaert, and Russell D. Gray. Phylogeographic analysis of the bantu language expansion supports a rainforest route. *Proceedings of the National Academy of Sciences*, 119(32):e2112853119, 2022. doi:10.1073/pnas.2112853119.

8 Sean Lee and Toshikazu Hasegawa. Bayesian phylogenetic analysis supports an agricultural origin of japonic languages. *Proceedings of the Royal Society B: Biological Sciences*, 278(1725):3662–3669, 2011. doi:10.1098/rspb.2011.0518.

9 Philippe Lemey, Andrew Rambaut, Alexei J. Drummond, and Marc A. Suchard. Bayesian phylogeography finds its roots. *PLOS Computational Biology*, 5(9):1–16, September 2009. doi:10.1371/journal.pcbi.1000520.

10 John Nerbonne. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1559):3821–3828, 2010. doi:10.1098/rstb.2010.0048.

11 Noemi Romano, Peter Ranacher, Sandro Bachmann, and Stéphane Joost. Linguistic traits as heritable units? spatial bayesian clustering reveals swiss german dialect regions. *Journal of Linguistic Geography*, 10(1):11–22, 2022. doi:10.1017/jlg.2021.12.

12 Yves Scherrer. dialektkarten.ch - interactive dialect maps for german-speaking switzerland and other european dialect areas. *Berichte aus der digitalen Geolinguistik (II): Akten der zweiten Arbeitstagung des DFG-Langfristvorhabens VerbaAlpina und seiner Kooperationspartner am 18.06.2019*, 2021.

13 Jean Séguy. La relation entre la distance spatiale et la distance lexicale. *Revue de linguistique romane*, 35:335–357, 1971.

14 Takuya Takahashi and Yasuo Ihara. Spatial evolution of human cultures inferred through bayesian phylogenetic analysis. *Journal of The Royal Society Interface*, 20(198):20220543, 2023. doi:10.1098/rsif.2022.0543.