

Inferring the History of Spatial Diffusion Processes

Takuya Takahashi ✉ 


Department of Geography, University of Zurich, Switzerland

Geneviève Hannes ✉ 

Department of Geography, University of Zurich, Switzerland

Nico Neureiter ✉ 

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

Peter Ranacher ✉ 

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

Abstract

When studying the spatial diffusion of a phenomenon, we often know its geographic distribution at one or more snapshots in time, while the complete history of the diffusion process is unknown. For example, we know when and where the first Indo-European languages arrived in South America and their current distribution. However, we do not know the history of how these languages spread, displacing the indigenous languages from their original habitat. We present a Bayesian model to interpolate the history of a diffusion process between two points in time with known geographical distributions. We apply the model to recover the spread of the Indo-European languages in South America and infer a posterior distribution of possible evolutionary histories of how they expanded their areas since the time of the first invasion by Europeans. Our model is more generally applicable to infer the evolutionary history of geographic diffusion phenomena from incomplete data.

2012 ACM Subject Classification Computing methodologies

Keywords and phrases Bayesian inference, geographic diffusion, language evolution, Indo-European, colonisation of the Americas

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.71

Category Short Paper

Funding Funding supports for this work were provided by the URPP Language and Space, University of Zurich, the NCCR Evolving Language with Swiss NSF Agreement No. 51NF40_180888, and the Swiss NSF Sinergia Project No. CRSII5_183578 (Out of Asia).

Acknowledgements We thank Gereon Kaiping for valuable discussion and ideas in the early phase of the project.

1 Introduction

Following the European colonisation of the Americas during the Age of Discovery, Indo-European (IE) languages, such as Spanish, Portuguese and French, spread extensively in South America, eliminating many indigenous languages. Historical records show when and where the IE languages arrived on the continent. The current spatial distribution of languages in South America is available in modern language maps. However, little is known about the spatio-temporal diffusion of the IE languages between the time of first contact at about 1500 CE and today. How have the IE languages spread between the time of contact and today? How can we infer probable evolutionary histories of this diffusion process in the absence of relevant historical records?



© Takuya Takahashi, Geneviève Hannes, Nico Neureiter, and Peter Ranacher;
licensed under Creative Commons License CC-BY 4.0

12th International Conference on Geographic Information Science (GIScience 2023).

Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 71; pp. 71:1–71:6



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In GIScience, similar questions frequently arise when studying diffusion phenomena, such as urban sprawl, deforestation, land cover change, segregation, or the spread of innovation. We know the spatial distribution at two points in time, and we would like to interpolate potential histories of how the diffusion has unfolded in between.

Various methods have been used to model the diffusion process in space. Cellular automata (CA) were applied to simulate the urban sprawl [4]. Reaction-diffusion equations were used to represent the demic-diffusion of modern humans theoretically [7]. Network models were used to describe the diffusion of human culture [6] and dialects [5]. Ising models were used to describe the diffusion of linguistic features in the UK [2, 3]. While these models can simulate the diffusion process from a given initial spatial distribution, they cannot infer the evolutionary history between two known points in time.

In this paper, we present a novel Bayesian model to interpolate potential histories of a spatial diffusion process, capturing the uncertainty of the process. The model reveals the distribution of the most likely evolutionary histories of a diffusion process, given the spatial distribution of the process at two points in time.

We applied the model in a case study to interpolate the spatial diffusion of the invasive IE languages in the Americas between the time of contact and today, capturing the uncertainty of the process. The model gives the posterior probability that an IE language occupied a given location in South America at a given time.

2 Methods

2.1 Model assumptions

We represent space as a network of n discrete nodes P_1, \dots, P_n , each assigned to one of K possible states. In the case study, P_i is a cell in a regular spatial grid over South America. The cell has two states: 1 means an IE language occupies the cell, and 0 means an indigenous language occupies the cell. We denote the geographical distribution of states at time t with the vector

$$\mathbf{M}(t) = \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix},$$

where $x_i(t)$ is the state of node P_i at time t . In the case study, $\mathbf{M}(t)$ is the geographical distribution of the IE languages in South America at a specific time in history. We model the spatial diffusion as a Markov process, where $\mathbf{M}(t)$ only depends on $\mathbf{M}(t-1)$ and is independent of earlier time steps. At each time t , every node copies the state from its neighbours at time $t-1$. The transmission rate a_{ij} , with $0 \leq a_{ij} \leq 1$ and $\sum_{j=1}^n a_{ij} = 1$, gives the probability that node P_i copies the state from P_j . The transmission rate is a constant and must be defined before the analysis. In the case study, each cell can copy the state from its eight neighbours and itself with equal probability:

$$a_{ij} = \begin{cases} \frac{1}{9} & \text{if } P_i = P_j \text{ or } P_i \text{ and } P_j \text{ are neighbours} \\ 0 & \text{otherwise} \end{cases}.$$

2.2 Bayesian inference

We use Bayesian inference to estimate the spatial diffusion $\mathbf{M}(0), \dots, \mathbf{M}(T)$, i.e. the history of the geographic distribution of states. The spatial diffusion follows a Markov process and has the probability

$$P(\mathbf{M}(0), \dots, \mathbf{M}(T)) = P(\mathbf{M}(0)) \prod_{t=1}^T P(\mathbf{M}(t) | \mathbf{M}(t-1)).$$

We know the geographic distribution at two points in time, the initial distribution at $t = 0$ and the final distribution at $t = T$. The history between initial and final distribution, $\mathbf{M}(1), \dots, \mathbf{M}(T - 1)$, has posterior probability

$$\begin{aligned} P(\mathbf{M}(1), \dots, \mathbf{M}(T - 1) \mid \mathbf{M}(0), \mathbf{M}(T)) &= \frac{P(\mathbf{M}(0), \dots, \mathbf{M}(T))}{P(\mathbf{M}(0), \mathbf{M}(T))} \\ &= \frac{P(\mathbf{M}(0))}{P(\mathbf{M}(0), \mathbf{M}(T))} \prod_{t=1}^T P(\mathbf{M}(t) \mid \mathbf{M}(t - 1)) \\ &\propto \prod_{t=1}^T \prod_{i=1}^n P(x_i(t) \mid \mathbf{M}(t - 1)). \end{aligned} \quad (1)$$

2.3 Markov chain Monte Carlo (MCMC)

We can use the Metropolis-Hasting (M-H) algorithm to draw samples from the posterior distribution in Equation 1, repeating the following steps:

1. Randomly choose one timestep t and one node P_i .
2. If $x_i(t) = k$, propose k' as a candidate state with the proposal distribution

$$q(k' \mid k) = \begin{cases} \frac{1}{K-1} & \text{if } k' \neq k \\ 0 & \text{otherwise} \end{cases}.$$

3. Compute the acceptance ratio

$$r = \frac{P(x_i(t) = k' \mid \mathbf{M}(t - 1))}{P(x_i(t) = k \mid \mathbf{M}(t - 1))} \prod_{j \in N(i)} \frac{P(x_j(t + 1) \mid x_i(t) = k', \cap_{(1 \leq l \leq n, l \neq i)} x_l(t))}{P(x_j(t + 1) \mid x_i(t) = k, \cap_{(1 \leq l \leq n, l \neq i)} x_l(t))}, \quad (2)$$

where $N(i)$ is the set of neighbours of P_i , or formally the set $\{j \mid 1 \leq j \leq n, a_{ji} > 0\}$. The conditional probabilities in expression 2 are computed with the transmission rates a_{ij} .

4. Accept the proposal with probability $\min(r, 1)$.

Letting m denote the average node degree of the network, one iteration of the MH-algorithm runs in $O(m)$ time.

3 Case study

In this section, we apply our model to explore the diffusion of the IE languages in South America, and interpolate probable evolutionary histories of how they have expanded their geographical area.

3.1 Network and diffusion model

We segmented the landmass of South America into a regular grid, each grid cell representing a node in the network. The Moore neighbourhood gives the transmission rate between cells: each cell may copy the state from its eight neighbours or itself with equal probability. Cells can take two states:

$k = 0$... an indigenous language occupies the cell

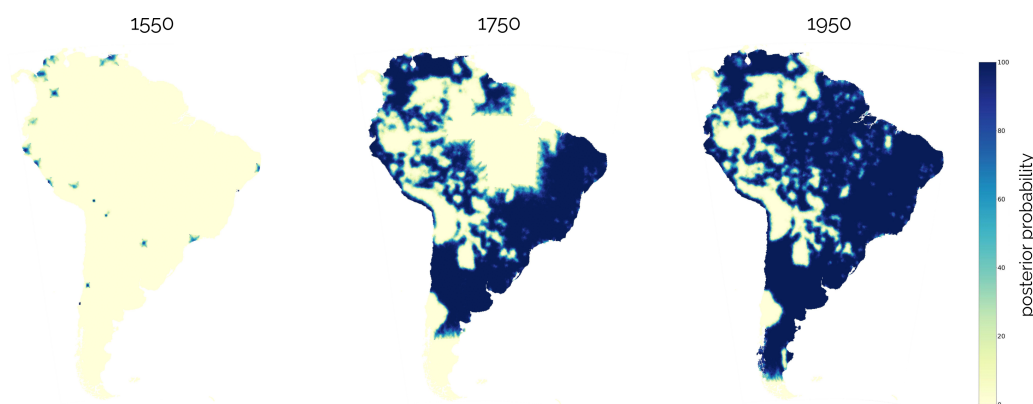
$k = 1$... an IE language occupies the cell

3.2 Data

The data comprise the geographical distributions of languages in 1510, the time of the first invasion, and 1990, the modern geographical distribution[1]. We included additional European arrivals between 1510 and 1990 from the literature. For example, the Spanish arrived in Santa Marta, modern-day Colombia, in 1525, and we fixed the state of the corresponding grid cells to 1 in the spatial distribution for this year.

3.3 Results

Figure 1 shows the posterior probability of the IE languages reaching each grid cell by 1550, 1750, and 1950. The IE languages gradually spread inland from the initial points of arrival at the coast. Figure 2 shows the posterior distribution of the arrival of the IE languages to selected cities along the Amazon basin.



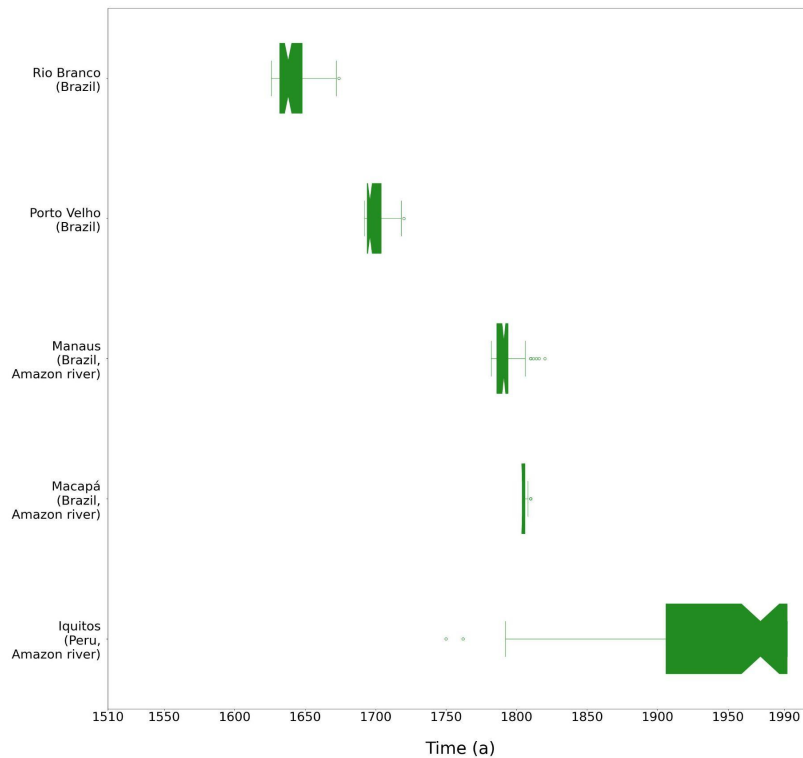
■ **Figure 1** Posterior distribution of IE languages reaching each grid cell by 1550, 1750, and 1950.

4 Discussion

In this paper, we presented a Bayesian model to interpolate the evolutionary history of a spatial diffusion process between two points in time with known geographic distributions. In a case study, the model showed likely scenarios of how the invasive Indo-European languages drove the indigenous languages of South America out of their original habitat.

In contrast to the conventional CA models, the model is fully Bayesian and returns a posterior distribution of possible evolutionary histories instead of just a single best history. In the case study, the model revealed the posterior probability of the IE languages reaching locations in South America between 1510 and 1990. Moreover, one can easily add prior information to Bayesian models and estimate the posterior distribution of potential evolutionary histories considering all available knowledge in a principled way. In the case study, for example, we added the locations of additional European entries to South America between the two known times in history.

In our model, the transmission rate reflects the influence of Geography on spatial diffusion. Since Bayesian models return a full posterior distribution, we can compare models with different transmission rates, e.g. using the Bayes factor, and evaluate the effect of geographic hypotheses on the diffusion process. For example, geographical barriers such as mountains and rivers might hinder the diffusion of languages, blocking the displacement of human groups. We could model this influence with lower transmission rates in mountainous terrain.



■ **Figure 2** Posterior distribution of the arrival of IE to selected cities.

Another possible extension includes mutation events, where a node may acquire a state not shared by any of its neighbours with a non-zero probability. Modelling the mutation event will enable the inference of an unrecorded arrival of an IE language not included in the data. Comparing two models with and without the mutation event could show whether today's geographical distribution has been formed by continuous diffusion or discontinuous state change. Since the geographical distribution at a given time still only depends on that at the previous time, including the mutation event does not violate the assumptions of a Markov process.

5 Conclusion

We present a method to infer potential histories of a spatial diffusion process between two points in time with known spatial distributions. We applied the method to infer the history of the IE languages spreading and displacing the indigenous languages in South America. Our method is more broadly applicable to infer the evolutionary history of geographic diffusion phenomena from incomplete data, frequently occurring in GIScience.

References

- 1 Ronald E Asher and Christopher Moseley. *Atlas of the world's languages*. Routledge, 2018.
- 2 James BurrIDGE. Spatial evolution of human dialects. *Phys. Rev. X*, 7:031008, July 2017. doi:10.1103/PhysRevX.7.031008.

71:6 Inferring the History of Spatial Diffusion Processes

- 3 James Burridge and Tamsin Blaxter. Using spatial patterns of english folk speech to infer the universality class of linguistic copying. *Phys. Rev. Res.*, 2:043053, October 2020. doi:10.1103/PhysRevResearch.2.043053.
- 4 Lingling Sang, Chao Zhang, Jianyu Yang, Dehai Zhu, and Wenju Yun. Simulation of land use spatial pattern of towns and villages based on ca-markov model. *Mathematical and Computer Modelling*, 54(3):938–943, 2011. Mathematical and Computer Modeling in agriculture (CCTA 2010). doi:10.1016/j.mcm.2010.11.019.
- 5 Takuya Takahashi and Yasuo Ihara. Quantifying the spatial pattern of dialect words spreading from a central population. *Journal of The Royal Society Interface*, 17(168):20200335, 2020. doi:10.1098/rsif.2020.0335.
- 6 Takuya Takahashi and Yasuo Ihara. Application of a markovian ancestral model to the temporal and spatial dynamics of cultural evolution on a population network. *Theoretical Population Biology*, 143:14–29, 2022. doi:10.1016/j.tpb.2021.10.003.
- 7 Joe Yuichiro Wakano, William Gilpin, Seiji Kadowaki, Marcus W. Feldman, and Kenichi Aoki. Ecocultural range-expansion scenarios for the replacement or assimilation of neanderthals by modern humans. *Theoretical Population Biology*, 119:3–14, 2018. doi:10.1016/j.tpb.2017.09.004.