



**University of
Zurich**^{UZH}

Interpolating spatial language evolution in South America

GEO 511 Master's Thesis

Author

Geneviève Bénédicte F. E. Hannes
13-941-976

Supervised by

Dr. Peter Ranacher
Dr. Gereon Kaiping (gereon.kaiping@gmail.com)
Dr. Takuya Takahashi

Faculty representative

Prof. Dr. Robert Weibel

28.04.2023

Department of Geography, University of Zurich

Abstract

Mapping language richness is essential not only for a better understanding of the languages themselves, but also to gain new insights into related cultural phenomena such as migration or expansion. However, spatial language distribution data can be sparse to non-existent, depending on the time and location. In the framework of this thesis, a probabilistic method is developed to interpolate spatial language distributions over time in the case of South America, where overall information on the distribution of Indigenous languages families and Indo-European languages is provided at only two points in time: around the time of contact and around 1990. The newly developed algorithm, that allows to interpolate between given points in time, is composed of a cellular automaton as core underlying mechanism and Bayesian inference as statistical method. Follow-up research is suggested to further test the transferability of this model, thereby building a solid foundation for a globally applicable model allowing to conduct linguistic research in various regions across the globe.

Acknowledgments

The generous support of several people allowed this thesis to evolve. First and foremost, I would like to thank Dr. Peter Ranacher for his dedicated supervision, his input ideas and for always having an open ear. I further thank Dr. Gereon Kaiping for guiding me through the world of programming and for continuing to co-pilot my thesis even after leaving university. I also thank Dr. Takuya Takahashi for joining this project while it was already running and for disclosing a whole new, fascinating chapter in my studies: Bayesian inference. The valuable expert feedback of all three supported me in letting this thesis take form.

I would like to thank Dr. Nico Neureiter and Alex Sofios, who were not only office colleagues, but with whom I shared many laughs. Thank you for making this time so special! Dr. Nico Neureiter furthermore provided invaluable counseling and assistance in many study-related questions, for which I am eternally grateful. I would also like to thank Dr. Philipp Striedl for his feedback concerning linguistics-related inquiries and Prof. Dr. Robert Weibel for the opportunity to write this thesis within the Geographic Information Systems unit and for making me feel welcome right from the beginning.

Last but not least, I would like to express my gratitude to my parents, who have supported me throughout all of my studies. I would also like to thank my friends, who have always been there for me during these last years and without whom I cannot imagine my study years. Finally, a special mention goes to my partner: thank you for your patience, laughs, and calming presence throughout the time of this thesis.

Contents

List of Figures	viii
List of Tables	x
1 Introduction	1
1.1 Motivation and study aim	1
1.2 Research goal	2
1.3 Outline	2
2 Background	3
2.1 Languages	3
2.1.1 A few linguistics definitions: language, language family and language richness	3
2.1.2 Driving factors of language spreading at a global scale	4
2.1.3 Language diversity in South America	5
2.2 Process-based models (PBMs)	6
2.2.1 Theoretical background	6
2.2.2 Current state of research	6
2.3 Cellular automata (CA)	7
2.4 Bayesian inference	8
3 Data	11
3.1 Study area	11
3.1.1 South America	11
3.1.2 Suitability of the study area	11
3.2 Language data	12
3.3 First settlements	13
3.4 Geographical data	15
3.4.1 Habitat data	15
3.4.2 Topographical data	15
4 Methods	16
4.1 Development of the CA	16
4.1.1 CA concept	16
4.1.2 Pre-processing the language data for the CA	18
4.1.3 Running the CA with the language data	20
4.2 MCMC for Bayesian inference	23
5 Results	26
6 Discussion	30
6.1 Choice of methods	30
6.1.1 A process-based simulation model (PBM) as overall concept	30
6.1.2 A cellular automaton (CA) as core underlying mechanism	30
6.1.3 Bayesian inference as statistical method	31
6.2 RG: Development of a probabilistic interpolation method for spatial language distributions	31

6.2.1	Data uncertainty	31
6.2.2	Restrictions due to modelling choices and limitations	32
6.2.3	Optimization potential: Example of the Amazon Basin	34
7	Future Research	38
7.1	Implementation of geographical factors	38
7.1.1	Selected factors and processes of language spreading	38
7.1.2	Weighted adjacency lists for geographical costs	39
7.1.3	Outline on implementing selected geographical factors	40
7.2	Running the interpolation with non-binary language data	42
8	Conclusion	45
	References	46
A	Appendix	I
A1	European settlements: References	I
A2	Amazonian settlements: References	I
	Personal Declaration	II

List of Figures

2.1	An illustration of the complexity of sorting and linking languages within a language family at the example of the Indo-European language family. (Source: Sundberg [2014])	3
2.2	Von Neumann (a) and Moore (b) neighbourhood functions. (Source: Das [2011])	8
3.1	The defined study area of this master's thesis in WGS 84: South America and Southern Panama within a bounding box (in blue).	11
3.2	Digitized Kaufman maps at "Time of Contact – 1500 A.D." (left) and "Contemporary – 1990 A.D." (right): Snapshots from the "Glottography" project. (Source: Kaufman [2007])	12
4.1	Moore neighbourhood for a grid cell $P_{i,j}$. (Source: Yassemi et al. [2008])	16
4.2	Example of the CA used to gain a potential interpolation between two given spatial distributions representing a "past" and a "present" map. Only two language families – represented by the colors green and orange – are present and the CA comprises four time steps.	18
4.3	Rasterized spatial language distributions for the retained 110 language families in South America at 1510 A.D. and 1990 A.D.	20
4.4	Rasterized binary "TOC" and "C" spatial language distribution maps of South America between which the CA interpolates.	21
4.5	CA-induced interpolation steps for binary spatial language distributions at 100-year-intervals.	22
5.1	Spatial extension probability of the Indo-European language family as dominant language in South America for selected years.	26
5.2	Early colonization patterns in the 16 th century (left) and Iberian colonies around 1780 in South America (right). (Sources: Dastrup [2020] (left), Britannica [2019b] (right))	27
5.3	Proportion of grid cells within South America primarily speaking an Indo-European language, cumulated over 100 samples/evolutionary histories.	29
5.4	Proportion of grid cells within South America primarily speaking an Indo-European language for selected samples/evolutionary histories. Left: medium grid cell proportion increase after 1850 with a sudden dip around 1990. Right: steady, but very slow grid cell proportion increase after 1850.	29
6.1	Map of the Amazon Basin depicting the Amazon river and its most important tributary rivers as well as important local centres. (Source: Kmusser [2013])	34
6.2	Selected cities in the Amazon Basin.	35
6.3	Time by which a Indo-European language first reaches selected settlements in the Amazon Basin based upon 100 samples.	36
7.1	Example of a graph with 6 nodes and ordered arcs comprising costs. E.g., the arc between the nodes 1 and 2 has the costs $c_{i,j}=3$. (Source: Pressl [2012])	39
7.2	Nine ordered arcs with costs $c_{P,j}$ leaving a central node P according to a Moore neighbourhood.	40
7.3	CA-induced interpolation steps for spatial language distributions containing 110 language families at 100-year-intervals.	43

List of Tables

3.1	Important European settlements in South America between 1510 A.D. and 1600 A.D. Settlements with a * have been built upon existing Indigenous dwellings or cities.	14
6.1	Summary of the median years by which an Indo-European language reaches selected settlements in the Amazon Basin. Furthermore, the approximate founding years of the settlements are added. (Source for founding dates: <i>Encyclopaedia Britannica</i> , see Table A2 in the appendix)	36
7.1	Typical adjacency list containing as many entries as the graph contains nodes and listing all outgoing arcs per node. Adjacency list based upon the example in Figure 7.1	39
7.2	Adjacency list containing as many entries as the graph contains ordered arcs and listing the start node, end node and weight per arc. Adjacency list based upon the example in Figure 7.1	40
7.3	Table showing the slope categories, i.e., the geographical costs given slope information, for certain slope ranges	42
A1	References for the implemented European settlements. Settlements with a * have been built upon Indigenous dwellings or cities.	I
A2	References for the implemented Amazonian settlements.	I

1 Introduction

1.1 Motivation and study aim

Language richness, i.e., the number of languages in a given region, and the spatial distribution of languages, especially over time, are tightly linked to cultural development and changes since language evolution reflects society-shaping natural and sociocultural influences. Mapping language richness is therefore essential not only for a better understanding of the languages themselves, but also to gain new insights into related cultural phenomena such as migration or expansion (Grollemund et al. 2015; Bouckaert et al. 2018). Or, following the train of thoughts by Lameli: “[...] our understanding of human language and communication benefits from the geographical approach, especially the knowledge organized via maps and atlases” (2009).

Currently, two big databases of the world’s languages exist: Ethnologue (Eberhard et al. 2022) and Glottolog (Hammarström et al. 2022). Ethnologue presents the disadvantages of not including any academic references and of not being freely accessible. Glottolog only contains point locations, no counts for languages nor speaker ranges, i.e., geographical areas occupied by the speakers of a language (Gavin et al. 2017). The GIScience department at UZH therefore launched the Glottography project to establish a unified presentation manner for working with world-wide language samples which takes into account geography by using already published geographical distributions of languages.

The goal of this master’s thesis is to develop a probabilistic method for interpolating spatial language distributions in a given area and over a certain time span, preferably of historical dimensions. The motivation behind developing this new probabilistic method is to map potential spatial language distributions over a time span for which only limited data exists. As introduced in the first paragraph, filling these gaps will hopefully not only allow a better understanding of the mapped languages themselves, but also facilitate historical studies of related cultural phenomena. The developed probabilistic method is embedded in a process-based simulation model and includes a cellular automaton as core underlying mechanism and Bayesian inference as statistical method.

To conduct my thesis, I chose to work with data from the Glottography project as it presents the advantages of both included academic references and speaker ranges - the latter representing an advantage for a study dealing with spatial language distributions. Within the available Glottography data, I selected the continent of South America as study site due to certified data only being available around the time of first contact, 1500 A.D. (Evers 2023), and 1990 A.D. This perfectly ticked the box of a large time span over which to develop my probabilistic interpolation method. Furthermore, this choice will hopefully also lead to a broadened knowledge of language richness and related cultural phenomena in South America.

However, I had to diverge from my initial plan to include all of the languages represented in the data from around 1500 A.D. and 1900 A.D. Indeed, due to time and computing constraints, which will be discussed in more detail in chapter 6.2.2, I had to focus on the distribution patterns of, on the one hand, the Indigenous language families as a whole and, on the other hand, the Indo-European language family. I hope though that, for more in-depth analysis, my developed method will, in the future, be reused to assess South America’s phylogenetic diversity patterns over time with more than just two groups of language families.

In summary, the goal of this thesis is to develop a probabilistic interpolation method for spatial language distributions based upon the example of South America between roughly 1500 A.D.

and 1990 A.D. My research aims at filling existing knowledge gaps in current language and phylogenetic, i.e., language family related, diversity studies for South America and at contributing to a better understanding of related cultural developments on a continent heavily influenced by European colonisation during this time period. Moreover, I'm positive that the outcome of this thesis will help to further spread the use of process-based simulation models when working with spatial language distributions. Finally, my newly developed method, based upon the combined use of a cellular automaton and Bayesian inference, will hopefully set the basis for a globally applicable model, thereby allowing to conduct research also in other regions across the globe.

1.2 Research goal

Within this study, I will focus on the following research goal:

How is it possible to interpolate spatial language distributions in a given area and over a certain time span with a known distribution both in the beginning and in the end?

1.3 Outline

In the next chapter, background information about the following topics will be introduced in the form of short summaries: linguistics definitions, the driving factors of language spreading at a global scale, South America's specific linguistic situation, process-based simulation models (PBMs) and related current research, cellular automata (CA) as well as Bayesian inference. Chapter 3 then introduces the study site - South America - and the data sets used for this thesis. Chapter 4 subsequently explains the chosen, applied and developed methodologies, before the results are presented in chapter 5. In chapter 6, the choice of methods and the research goal are discussed. Possible improvements are propounded and an outlook to future research is given in chapter 7. Finally, conclusions are drawn in chapter 8.

2 Background

2.1 Languages

2.1.1 A few linguistics definitions: language, language family and language richness

About 7000 languages and 400 language families exist worldwide (Campbell 2019; Pacheco Coelho et al. 2021). The term “language family” describes a set of languages for which there exists proof of a common ancestor (Campbell 2019). The 400 language families also comprise isolates, i.e., languages with no known relative forming single membership language families, and extinct language families, i.e., language families which do not contain any language with remaining native speakers (Campbell 2012; Campbell 2019). A glimpse at the complexity of sorting and linking languages within a language family can be seen in Figure 2.1.

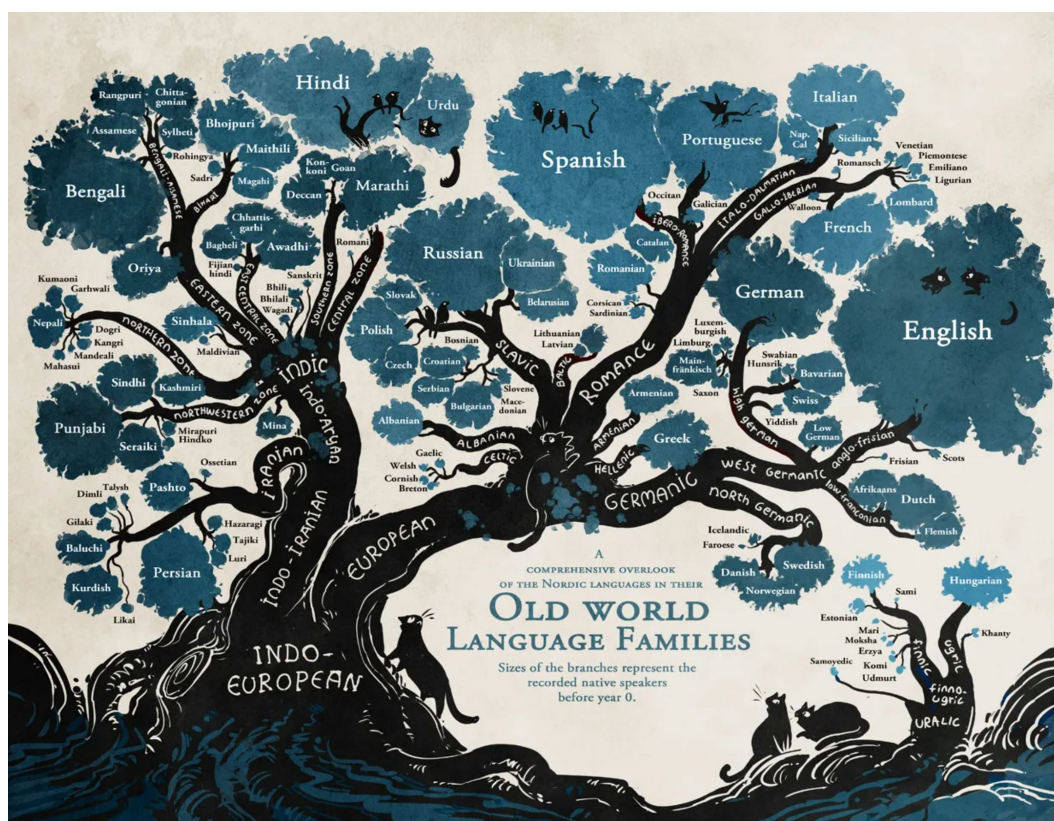


Figure 2.1: An illustration of the complexity of sorting and linking languages within a language family at the example of the Indo-European language family. (Source: Sundberg 2014)

However, it is often difficult to decide whether a detected language is merely a dialect or a language of its own – a correct classification based upon size or mutual intelligibility can be ambiguous and sometimes impossible to find (Campbell 2012; Oppenheim et al. 2019). Furthermore, as Oppenheim et al. state, “distinctions often incorporate politicized notions like ‘prestige’ (Hudson, 1996; Wei, 2000), leading to the popular aphorism that “language is a dialect with an army and a navy” (Weinreich, 1945)” (2019). In this thesis, I will use the term ‘language’ according to the conception used by Kaufman (2007) as I am using his language data (see chapter 3.2). This means that languages are categorized and distinguished from dialects based upon their phonology, lexicon, morphology, texts, and syntax (Kaufman 1990).

While the number of languages in a given region can either be referred to as language richness or language diversity, the amount of language families in an area is called phylogenetic diversity. The term “linguistic diversity” should be avoided if possible as it is ambiguous (Pacheco Coelho et al. 2019; Pacheco Coelho et al. 2021).

A “spatial language distribution” describes how languages or language phenomena like language families are organized in space, i.e., how they are regionally distributed (Lameli 2009). The spreading of languages happens according to human agencies such as military, economical, or religious activities and is the reason behind changing spatial language distributions (Phillipson 2008). “Language spreading” is therefore defined as the gain, over time, of a language’s speaker numbers (Cooper 1982) with the critical step being the corresponding language’s adoption by at least one foreign population group (Mufwene 2006). Hence, the mere demographic increase of a language’s original population group is not enough to speak of language spreading (Mufwene 2006). A good example for language spreading is the dispersal of the Portuguese and Spanish languages in South America: the number of Portuguese and Spaniards did not simply increase, but South American speaker communities stopped using their traditional languages and started speaking Spanish and Portuguese instead. This process happened mostly forcefully and was marked by the Europeans’ military invasion of South America.

2.1.2 Driving factors of language spreading at a global scale

There exist multiple non-spatial and spatial factors and processes which, over time, have been discussed to influence language spreading (Greenhill 2014). Spatial processes alone explain over 1/3 of the world’s spatial language distributions. Important spatial factors and processes of language spreading are water bodies, mountains, ecosystem boundaries and the ecosystem boundaries’ universal role as spatial barriers. As for non-spatial processes, at a global scale, population size seems to be the only relevant factor (Bentz et al. 2018; Moore et al. 2002). Population size is a very valuable factor since it indirectly also covers socio-cultural and political elements which heavily impact the population size (Gavin et al. 2017). For example, wars and diseases generally reduce the population size which therefore implicitly indicates such a political or socio-cultural event.

Water bodies, especially rivers, have a very ambiguous role when it comes to language spreading. On the one hand, rivers, as well as coast lines, tend to favour a high language diversity and reduced language spreading due to high individual but very low group mobility (Bouckaert et al. 2018; Greenhill 2014). This means that generally, many individual population groups speaking different languages live along a river or coastline in their own niches. Their contact and exchange is mostly reduced to individual mobility, e.g., trade or marriage. Since nearly no larger group movements like migration take place, the various languages are neither massively spreading nor suppressing other languages along the rivers or coastlines: there is a high language diversity but reduced language spreading in these areas. However, in certain cases, usually over very long distances, rivers and coastlines can also be important traffic and transport routes (Nichols 1997). In these cases, since people speaking various languages from very different regions either move permanently (water body as a traffic route) or meet frequently (water body as a trading route) along the river or coastline, these water bodies can become a factor of increased language spreading (Nichols 1997). Good examples of it are the increased language diversity along the coastlines of New Guinea as well as the extremely high number of language families along the world’s largest river system, the Amazonas (Ranacher et al. 2017).

Since at least the late Middle Ages, mountains allow for language niches with high language

diversity while language spreading occurs bottom-up with the valley languages slowly eradicating the high altitudes' languages. The spreading mechanism behind this bottom-up pattern is the so-called “vertical bilingualism”, which describes the fact that highland settlers generally know lowland languages but not vice versa. The reason for the bottom-up language spreading is climate. Indeed, the beginning of the Little Ice Age led towards highland economies becoming more precarious and lowland economies becoming more prosperous due to comparably longer growing seasons. Therefore, lots of highland settlers at least partially migrate for work towards the lowlands and “vertical bilingualism” increases (Nichols 1997).

People and therefore languages tend to move along similar habitats if possible: a longer migration route to the same or a similar ecosystem has proven to be preferred over shorter routes to foreign ecosystems where new adaptation mechanisms have to be learnt (Grollemund et al. 2015).

2.1.3 Language diversity in South America

South America comprises roughly 420 spoken Indigenous languages grouped into about 100 language families, including isolates (Campbell 2012; Campbell 2019). In total, the South American language families represent about a quarter of the world's language families (Campbell 2012; Campbell et al. 2012). However, Indigenous South American languages and language families are not strictly confined to the continental territory of South America: some of them spread into lower Central America as well as the Caribbean islands known as Antilles (Campbell 2012; Kaufman 2007).

Their uniquely high language diversity makes Indigenous South American languages somewhat special (Campbell 2012) and a very interesting study area. However, despite their unique position, the systematic study of Indigenous South American languages only started in the 1940s with a more in-depth understanding of them beginning to take place even later, around the early 2000s (Campbell et al. 2012). This might also be due to the obstacle of naming issues: indeed, South America comprises many single languages with multiple names as well as the opposite case where one name refers to several languages (Campbell 2012).

South America nowadays has the highest language diversity in both mountain (Nichols 1997) and latitudinal gradient niches (Nettle 1998). While the former reinforces suspicions that, despite bottom-up language spreading, the overcoming of high slopes is not easily done by languages and language families (Nichols 1997), the latter proves Nettle's “Ecological Risk Hypothesis” true (Nettle 1998). The “Ecological Risk Hypothesis” claims that ecosystem richness, based upon climatic drivers like precipitation, temperature, and seasonality as well as the ecosystem's stable productivity conditions, favours a latitudinal gradient with language and niche diversity being higher closer to the Equator (Greenhill 2014; Grollemund et al. 2015; Mace et al. 1995; Nettle 1998; Pacheco Coelho et al. 2019). However, due to the European colonization and its heavy impact upon Indigenous South American culture, language diversity close to the South American equator is not as high as expected in comparison with other equatorial regions (Nettle 1998).

2.2 Process-based models (PBMs)

2.2.1 Theoretical background

Process-based simulation models (PBMs) originate in macro ecology (Gavin et al. 2017) and are custom-built models made to fit data (Connolly et al. 2017). They are opposed to “purely statistical” models (Connolly et al. 2017) and have clearly defined rules and interpretable parameters (Pacheco Coelho et al. 2021). Their working principle is “allowing investigators to hold certain factors constant to isolate and assess the impact of certain chosen processes” (Gavin et al. 2017). This means that there is a shift from single-factor correlative studies towards multi-causal approaches which include statistical methods but are more than just a statistical model (Gavin et al. 2013). Furthermore, PBMs are a computer-simulated experience (Pacheco Coelho et al. 2021).

Since central topics of ecology -- like explaining patterns or using heterogenous, gridded environmental data as basis for species richness (Gotelli et al. 2009) — are similar to working with language diversity patterns, macro ecology’s methodological advance in using PBMs is considered useful in future work on language distribution patterns (Gavin et al. 2013).

Indeed, factor-driven analysis of language spreading have so far mostly been empirical, correlative studies. However, correlation does not necessarily infer causation, this being the reason why PBMs can be used to properly determine the specific drivers of changing language diversity patterns by detecting if, how, and to what extent the chosen factors determine the number and spatial pattern of languages in a certain place, e.g., South America (Gavin et al. 2017; Pacheco Coelho et al. 2021). Furthermore, PBMs allow for regionally different drivers of language diversity changes (Pacheco Coelho et al. 2021). Studies using PBMs for assessing language diversity patterns have already been conducted in Australia and North America and will be discussed in more detail in chapter 2.2.2 (Gavin et al. 2017; Pacheco Coelho et al. 2019; Pacheco Coelho et al. 2021).

2.2.2 Current state of research

Currently, three main studies using process-based models (PBMs) for assessing language distribution patterns have been conducted (Gavin et al. 2017; Pacheco Coelho et al. 2019; Pacheco Coelho et al. 2021).

In the first one, by Gavin et al., the authors investigate a potential causal relation between, on the one hand, the number of languages and the spatial language distribution in modern Australia and, on the other hand, some major factors and processes determining them. The study focuses on isolating and assessing a strict minimum of underlying processes and factors. The idea here is to get a best possible approximation by using and testing the least possible hypotheses. The result is stunning: the average predicted number of languages corresponds to the observed number of languages and the estimated spatial language distribution shares 56 percent with the real-world language distribution of today.

To achieve this, Gavin et al. used an underlying hexagonal grid. Besides the finding that very few processes and factors already allow for some rather well matched number of languages and spatial language distributions, the study also shows that the before untested hypotheses of environmentally limited group size per area, i.e., carrying capacity, and climatic conditions are indeed key causal factors. Furthermore, Gavin et al. presume that such key causal factors are undergoing regional changes since the carrying capacity is quickly outshined where other

processes like natural barriers are likely to determine language distributions. This is supported by the findings of Pacheco Coelho et al. (2019) and Pacheco Coelho et al. (2021), stating that universal indicators, similar to macro ecology, do not exist.

Pacheco Coelho et al. ((2019)) based their work upon the study by Gavin et al. (2017) and tested for more processes and factors determining the number of languages and the geographical language distribution in a given region, North America, using a geographically weighted path analysis. Besides additionally assessing topological complexity, river density, ecosystem diversity, and population density, they also tested for both direct and indirect effects of all the factors. Their main findings are that the impact of causal factors, especially ecological ones, varies spatially and that the factors are “connected in a complex web of causality, consisting of both direct and indirect effects”. Moreover, results indicate population size to be the most influent factor while the absence of sociocultural and historical factors due to the use of gridded map cells is presumably responsible for at least some of the unexplained variation.

Pacheco Coelho et al. (2021) then built a more complex simulation model based upon their study from 2019. The implementation is done within a hexagonal grid and based upon artificial algorithmic cycles. The language ranges are depicted as non-overlapping polygons in order to simplify the modelling. Pacheco Coelho et al. also introduced and modelled the concept of “shocks” for rapid changes, both positive and negative, in population sizes. This is important since population size change can lead to range expansion or contraction and therefore induce processes like language fragmentation and diversification or language extinction - crucial processes which were not represented in previous simulation models focusing only on population size itself. Important for modelling shocks is the notion of “carrying capacity”, i.e., the environmentally limited group size per area, already introduced by Gavin et al. in 2017. The shocks can then be “ [...] limited to a particular group [...] or felt by all the groups within a given region [...] ”, therefore emphasizing the previous findings of regional, not global, processes and factors determining the number of languages and the geographical language distribution. Furthermore, Pacheco Coelho et al. (2021) introduce the concept of languages being able to emerge from randomly selected cells.

2.3 Cellular automata (CA)

Cellular automata (CA) are mathematical representations of complex systems which evolve within time and space, first introduced by Von Neumann in the 1960s. They are constituted of a set of discrete elements, normally a grid, where each cell is in one of several finite states, i.e., where each cell can only take one single value per time step (Beltran et al. 2010; Das 2011; Sarkar 2000; Yassemi et al. 2008). For example, in a binary grid, each cell can at each time step t either take value 0 or value 1.

Each cell’s value can change once per time step. The change occurs according to a set of rules called “transition rules” which are a function of the cell’s own value at the preceding time step and the neighbouring cells’ values at the current time step. The transition rules can contain deterministic, probabilistic, or stochastic elements and use either a Von Neumann — only the cells connected to the sides of the cell in question — or a Moore — all the cells connected to the sides and vertices of the cell in question — neighbourhood function (see Figure 2.2) (Beltran et al. 2010; Yassemi et al. 2008). This means that “each cell is restricted to the local neighborhood interactions only [...]” (Das 2011).

In the beginning, a cellular automaton is in an initial configuration from which it proceeds deter-

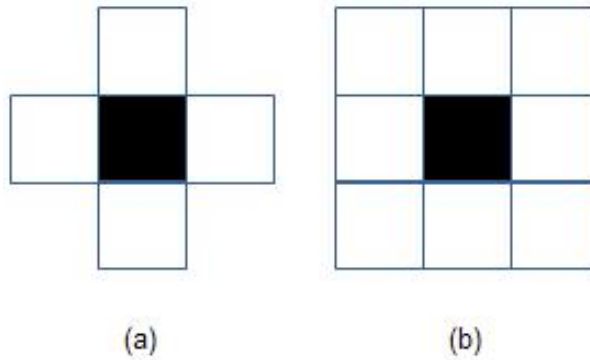


Figure 2.2: Von Neumann (a) and Moore (b) neighbourhood functions. (Source: Das [2011](#))

ministically according to the transition rules. Since it has no other input, it is an autonomously working model (Sarkar [2000](#)).

Cellular automata present several advantages for spatio-temporal modelling compared to other methods: they can deal with both the spatial and temporal component while being inherently spatial, they are compatible with geospatial data sets, i.e. GIS data, and, most importantly, they allow us to model complex situations while using rather simple and local rules (Yassemi et al. [2008](#)). The complexity and predictability of a cellular automaton’s output heavily depend on the number of cells and their potential values, as well as the transition rules and the chosen neighbourhood function (Beltran et al. [2010](#)). This can be considered a disadvantage of cellular automata.

Nevertheless, cellular automata, originally developed as “formal models of self-reproducing organisms”, are nowadays not only commonly used for modelling in biology, but in many other domains (Sarkar [2000](#)).

2.4 Bayesian inference

Bayesian inference is a concept for data analysis which has its roots in statistics, but is widely applied in all types of research. It is based on probabilities and the previous knowledge of events, i.e., the learning from data. This means that Bayesian inference reflects our knowledge about the world around us and asks the question: “How does one get from mere sample data towards a probability distribution of a population and its parameter(s)?” (McElreath [2016](#))

For example, one can make a survey at the university cafeteria to ask whether people like tea. The goal of such a survey would be to better restock the cafeteria’s offerings. The interviewed university staff’s answers – Yes or No – are data generated through a process of sampling. The question to answer then is: “How can one infer the probability distribution of the complete university staff’s positioning towards tea from the data gathered through the few staff members interviewed?” Knowing the model of our case study, i.e., sampling as data-generating process, one can subsequently infer the single parameter – probability of tea liking within all of the university staff – from the data collected in the sampling process.

To answer this type of question in a structured way, Bayesian inference uses the Bayesian theorem:

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters}) * P(\text{parameters})}{P(\text{data})} \quad (1)$$

Which equals to:

$$\text{posterior (probability)} = \frac{\text{likelihood (function)} * \text{prior (probability)}}{\text{marginal likelihood (function)}} \quad (2)$$

The posterior probability gives the probability distribution of a specific model parameter according to the observed data. The highest peak of this probability distribution corresponds to the highest posterior probability (y-axis value) and the most likely value for the analyzed specific parameter (x-axis value) (McElreath 2016). In our example, the posterior probability is the probability distribution of the whole university staff’s tea fondness given the results from the survey conducted at the cafeteria.

The likelihood function indicates how likely certain data is to be produced for each parameter value of the model. The likelihood function is the information we have in any case of applying the Bayesian theorem (McElreath 2016). In our example, the likelihood function expresses how probable the answers “Yes” or “No” are for the question of how well university staff overall likes tea.

The prior probability is the knowledge of each specific parameter value before seeing any data. A core element of the Bayesian inference concept is that the prior needs to be updated whenever new information is available. The prior can either be informative (strong) or uninformative (flat): the former is used when past experiences or domain knowledge exists, and the latter is used if no information is available (McElreath 2016). In our example, the prior is the knowledge about tea preferences at university from former surveys. In case such a former survey exists, the prior is strong, otherwise, due to lack of previous knowledge, the prior is weak.

The marginal likelihood is a normalizing constant – since it does not depend on a specific parameter – which can be omitted without losing proportionality. In our example, the marginal likelihood corresponds to the data collected in the cafeteria survey. Due to the omitting of the normalizing constant, the Bayesian theorem is reduced to:

$$P(\text{parameters} \mid \text{data}) \propto P(\text{data} \mid \text{parameters}) * P(\text{parameters}) \quad (3)$$

The posterior can mainly be influenced in two ways. The first one is an increased amount of data which leads towards a less influential prior while narrowing the posterior. Most importantly, though, the posterior probability better approximates the true population parameters in that case. The second way is a very strong prior probability, i.e., specific, definitive information about a parameter is known. The prior then pushes that parameter towards specific – potentially biased – values, and the posterior is narrowed. However, contrary to the case with the increasing amount of data, it is not sure if the posterior approaches the true population parameters due to the potentially biased prior information. Therefore, a good prior probability is helpful, but gathering more data is even better for running a Bayesian inference. Indeed, with few observations, all we get is a blurry picture, while with more observations, we can be more certain about the true nature of things. Nevertheless, as with every statistical method, some uncertainty always remains (McElreath 2016).

Bayes' theorem is, in practice, applied to many possible values for each parameter. For each potential value, it is retained how well they explain the data. In case the model and population have only one parameter, grid approximation is used to find suitable estimates. For simple models, there may even exist an analytical solution instead of sampling. If the model and population comprise several parameters though, a Markov Chain Monte Carlo (MCMC) approach is usually the method of choice. Contrary to the grid approximation, the MCMC mainly focuses on values that explain the data well and neglects those that don't. Furthermore, it already needs a good initial value to start its journey through parameter space. Both grid approximation and MCMC are tools that estimate the posterior distribution of a model and its parameters given the data (McElreath [2016](#)).

The major advantages of Bayesian inference are the possibility to capture the uncertainty of a process, that the concept can be extended to model processes of almost any complexity and that its results are very straightforward to interpret (McElreath [2016](#)).

3 Data

3.1 Study area

3.1.1 South America

The study area of this thesis is South America itself, defined as the landmass lying south of the isthmus of Panama (Wallenfeldt 2018), as well as Southern Panama. For better readability, I will however address the study area as “South America” and not mention Southern Panama separately every time.

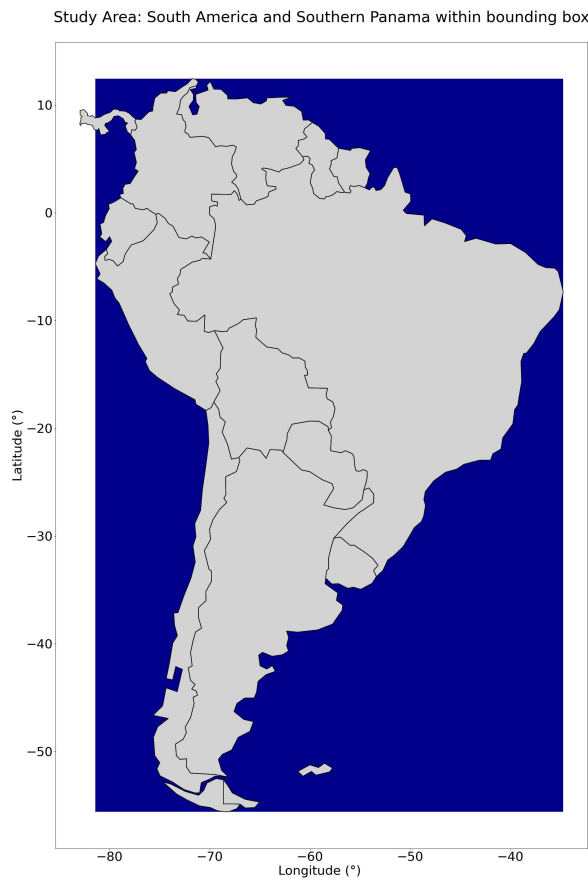


Figure 3.1: The defined study area of this master’s thesis in WGS 84: South America and Southern Panama within a bounding box (in blue).

The exact study area ranges from $55^{\circ}36'42.6''\text{S}$ to $12^{\circ}26'14.3''\text{N}$ and from $81^{\circ}24'39.4''\text{W}$ to $34^{\circ}43'48.0''\text{W}$ within the World Geodetic System 1984 (WGS 84) (see Figure 3.1). This bounding box has been extracted from the low-resolution “World” dataset by Natural Earth which is a standard dataset included in the Python module “GeoPandas”. The bounding box corresponds to the corresponding minimum and maximum values of the combined geometry attributes of the data set’s entries where the continent is “South America”.

3.1.2 Suitability of the study area

South America is a suitable study area for developing my probabilistic method for interpolating spatial language distributions since comparable, certified language data – from the Glottography

project, see chapter 3.2 – exists for two timestamps in that area: around the time of first contact, 1500 A.D. (Evers 2023), and around 1990 A.D. In South America, interpolating between these two years means interpolating over a time period of historic dimensions because of colonization. Indeed, colonization in South America started around 1500 A.D. and did not only impact language and phylogenetic diversity as well as spatial language distributions, but also had heavy cultural consequences (Nettle 1998). Such a vast and meaningful time period is, on the one hand, more challenging concerning the results, while also, on the other hand, guaranteeing big enough changes in spatial language distributions for my interpolation method to register.

3.2 Language data

While modern studies of Indigenous South American languages mostly focus on spatially or contentwise limited topics like certain areas or individual language families (Campbell et al. 2012), there exist some overarching large-scale classifications of Indigenous South American languages (Campbell 2012). One of these is a comprehensive mapping of both "Time of Contact" and "Contemporary" language distribution patterns by Terrence Kaufman (2007).

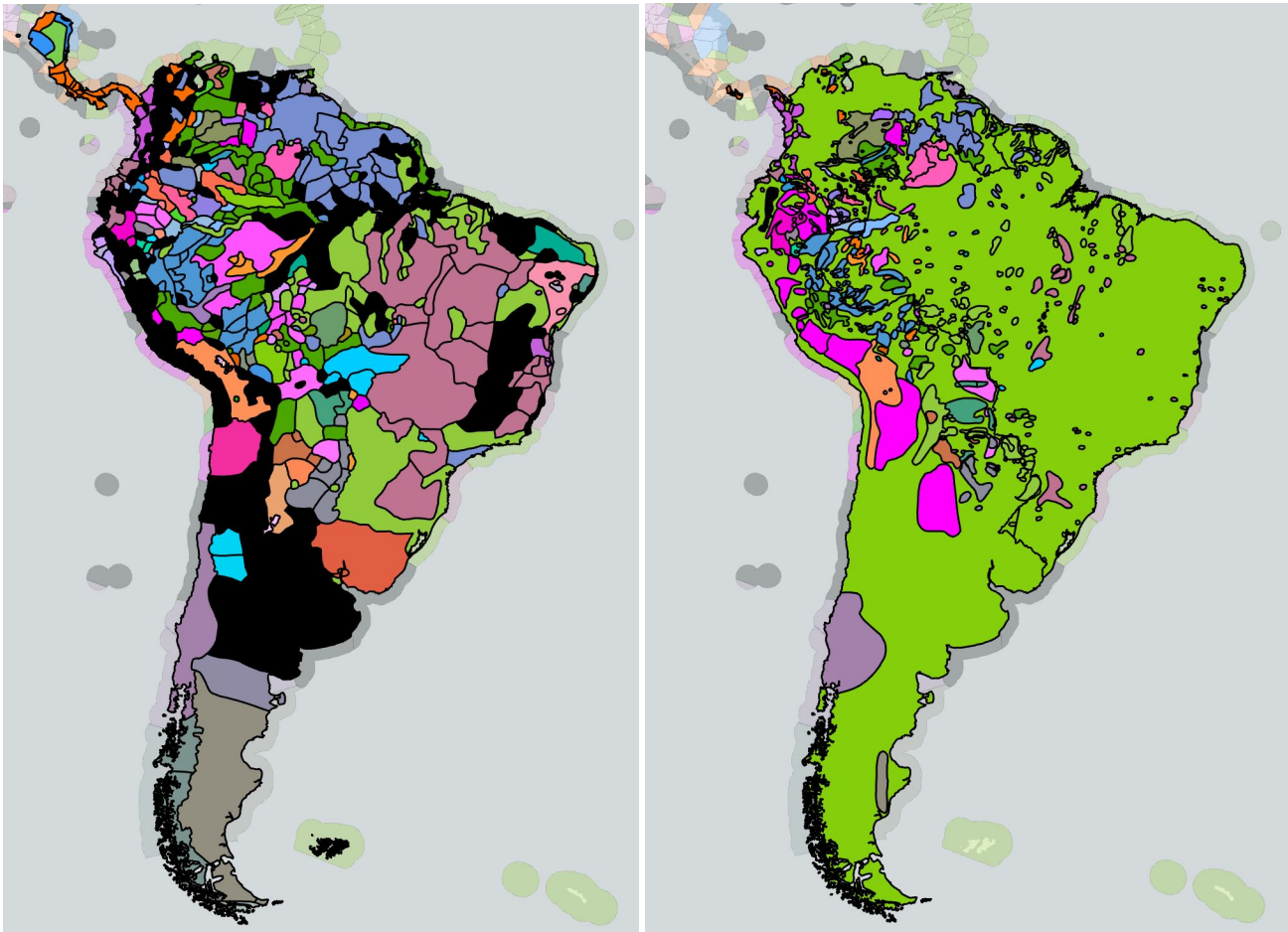


Figure 3.2: Digitized Kaufman maps at "Time of Contact – 1500 A.D." (left) and "Contemporary – 1990 A.D." (right): Snapshots from the "Glottography" project. (Source: Kaufman 2007)

The classification and subsequent mapping by Kaufman is the base data set used in this thesis.

It is based on the overall common points between previous overarching large-scale classifications by Swadesh (1959), Loukotka (1968), Suárez (1974) and Greenberg (1968) as seen in Campbell 2012 and was continuously updated between 1985 and 1989 (Kaufman 2007). While the "Contemporary" maps therefore refer to the time around 1990 at the latest, the "Time of Contact" maps depict the areas occupied by speakers of the represented Indigenous South American languages at the corresponding time of first contact with European colonizers around 1500 A.D. (Asher et al. 2007). As described in chapter 2.1.3, naming issues are a main complication of Indigenous South American languages' classification. In the mapping by Kaufman, for standardization, languages are therefore referred to with their most common English name or the name used by scholars to describe them (Asher et al. 2007; Kaufman 2007). Common spelling variations are also added (Asher et al. 2007).

For this thesis, I use digitized polygon versions of Kaufman's maps (see Figure 3.2) which are also used for the Glottography project. I obtained the data from the ERC Consolidator project "South American Population History Revisited (SAPPHIRE)" headed by Erik van Gijn at Leiden University. The digitized polygons were handed to me as two separate .json data sets: one for "Time of Contact", one for "Contemporary".

Using the digitized polygon versions of Kaufman's large-scale classification of Indigenous South American languages presents two main assets for me: First, the fact that the original assessment and classification of both "Contemporary" and "Time of Contact" language distribution patterns were done by the same person – Kaufman – allows for sensible comparisons between both, i.e., makes my probabilistic interpolation method statistically sound. Second, the digitization was done by specialists in the field of linguistics and should therefore contain no errors.

3.3 First settlements

The digitized language distribution map from around "Time of Contact" (TOC) only contains indigenous language families. To retrace the spreading of the Indo-European language family in South America, I therefore need to implement European settlement seeds which I assume to introduce Indo-European languages as dominant languages in their corresponding areas. At least one seed needs to be implemented within the language distribution map from around "Time of Contact" for an interpolation between the TOC and "Contemporary" (C) maps to be possible. More seeds can then be added to the CA-generated follow-up maps in order to better represent the historic reality.

Since the TOC map by Kaufman (Kaufman 2007) describes the language diversity patterns in South America roughly around 1500 A.D., I decided to use the first stable, yet short-lived, European settlement in continental South America as starting point: Santa María la Antigua del Darién, founded in 1510 (Cubero-Hernández et al. 2022; Keen 2023). Further European settlement seeds, chosen upon the criteria of successful, non-temporary foundations undertaken between 1510 A.D. and 1600 A.D., are added to the CA-generated follow-up maps at fitting time stamps. The reason for choosing 1600 A.D. as limit for introducing new settlement seeds is that more and more cities and dwellings started to be founded then and that, with me not having the historical background to properly assess which of these new settlements are most important to South American colonization, I preferred to favour missing information over potentially wrong information. In an updated version of my probabilistic method for interpolating spatial language distributions, it would however strongly be recommended to reassess and extend the choice of European settlement seeds for higher accuracy.

The chosen European settlement seeds (see Table 3.1) represent the two most powerful colonizing forces of South America: Spain and Portugal. The other three present forces, France, the Netherlands and the United Kingdom, only started founding durable settlements in South America during the 17th century (British Library 2018; Ebert 2019; Webster et al. 2023) and are therefore not represented due to my chosen seed-placing limit of 1600 A.D. Since I am studying the spreading of the Indo-European language family as a whole, discharging the still scarcely represented languages French, Dutch, and English (Kaufman 2007) is not a problem for the modelling process.

	Settlement	Founding Year	Modern Country	Colonizing Force
1	Santa María la Antigua del Darién	1510	Colombia	Spain
2	Panamá Viejo* (Panama City)	1519	Panama	Spain
3	Nueva Toledo (Cumaná)	1521	Venezuela	Spain
4	Santa Marta	1525	Colombia	Spain
5	Piura	1532	Peru	Spain
6	São Vicente	1532	Brazil	Portugal
7	Cartagena*	1533	Colombia	Spain
8	Cuzco*	1533	Peru	Spain
9	Quito*	1534	Ecuador	Spain
10	Trujillo	1534	Peru	Spain
11	Lima	1535	Peru	Spain
12	Asunción	1537	Paraguay	Spain
13	Olinda	1537	Brazil	Portugal
14	Chuquisaca* (Sucre)	1538	Bolivia	Spain
15	Santa Fé de Bacatá* (Bogota)	1538	Colombia	Spain
16	Santiago de Chile	1541	Chile	Spain
17	Potosí	1545	Bolivia	Spain
18	Nuestra Señora de la Paz* (La Paz)	1548	Bolivia	Spain
19	Salvador	1549	Brazil	Portugal
20	Concepción	1550	Chile	Spain
21	Huancavelica	1563	Peru	Spain
22	Rio de Janeiro	1565	Brazil	Portugal
23	Caracas	1567	Venezuela	Spain
24	Cochabamba	1574	Bolivia	Spain
25	Buenos Aires	1580	Argentina	Spain

Table 3.1: Important European settlements in South America between 1510 A.D. and 1600 A.D. Settlements with a * have been built upon existing Indigenous dwellings or cities.

The European settlement seeds to include were mostly determined based upon historical evidence from the online version of the *Encyclopaedia Britannica* (see references in Table A1 in the appendix) and are listed according to their name at the founding time. In case their modern name diverges, it is added in brackets. The settlements' corresponding coordinates, needed to implement the seeds into my interpolation method, were gathered through the website *Geohack* (N.N. 2023b). The settlements marked by an * in Table 3.1 were built upon pre-existing Indigenous dwellings or cities.

3.4 Geographical data

3.4.1 Habitat data

Habitats are defined as natural landscapes providing food and shelter to a given organism or group of organisms (NGS [2022](#)). However, the only publicly available data covering all of South America was an ecoregions data set called “Ecoregions 2017” (Dinerstein et al. [2017](#)) and found at <https://ecoregions.appspot.com/>.

Ecoregions are not centered upon a specific species and its needs like habitats. Instead, they cover broader areas classified according to their interacting fauna, flora and climate (IPBES [2023](#)). Contrary to habitats, ecoregions are geographically specific and exist only once, e.g., the “Venezuelan Andes montane forests” are an ecoregion only located in Venezuela (IPBES [2023](#), [Ecoregions2017](#)). However, despite being different from each other, habitats and ecoregions are still closely related due to being defined spatial areas based upon their environmental features. Therefore, in the context of this thesis and due to the aforementioned data constraints, I make use of ecoregions instead of habitat data.

3.4.2 Topographical data

The topographical data is provided by the GLOBE digital elevation model (DEM) (GLOBE [1999](#)) available at <https://www.ngdc.noaa.gov/mgg/topo/globe.html>. This is a 30-second (approximately 1-km) DEM provided and quality-controlled by the National Oceanic and Atmospheric Administration (NOAA).

4 Methods

The goal of this thesis is to develop a probabilistic method for interpolating spatial language distributions in a given area, South America, over a certain time span, i.e., the years between 1510 A.D. and 1990 A.D. The motivation behind developing this new probabilistic method is to map potential spatial language distributions over a time span for which only limited language data exists.

In order to achieve this, my chosen statistical method is Bayesian inference which infers the posterior distribution of potential evolutionary language histories, i.e., potential interpolation sets of spatial language distributions between 1510 A.D. and 1990 A.D. To be able to use Bayesian inference with good context data and within a reasonable computation time, I need to produce a “starting history” as input data for it. For this, I use a cellular automaton (CA): the final output of the CA is a first potential interpolation of spatial language distributions in South America between 1510 A.D. and 1990 A.D. Despite being non-probabilistic, this output is a good input data set for the Markov Chain Monte Carlo (MCMC) method, the tool I use for implementing Bayesian inference in this thesis.

I use a cellular automaton because CA have already successfully been used over a longer period within language studies (e.g., Beltran et al. 2010). Furthermore, their grid-like nature will hopefully allow for a good diffusion simulation of colonial languages in South America (Gavin et al. 2017; Pacheco Coelho et al. 2019). Bayesian inference is suited for my interpolation project since it allows me to roughly assess the reliability of my interpolation method and therefore to continuously improve the generated evolutionary language histories. The choice of the MCMC as tool for implementing the Bayesian theorem is due to my Bayesian inference model containing many parameters and an MCMC being typically applied in such cases.

4.1 Development of the CA

All programming related to the CA and the data used within it is done in Python.

4.1.1 CA concept

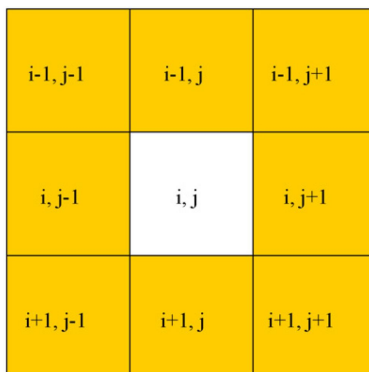


Figure 4.1: Moore neighbourhood for a grid cell $P_{i,j}$. (Source: Yassemi et al. 2008)

I use a CA (see chapter 2.3) to interpolate an initial evolutionary history between the “Time of Contact” around 1510 and “Contemporary” around 1990. The spatial language distribution around “Time of Contact”, called TOC, is the CA’s initial configuration from which it proceeds.

The second spatial language distribution, called C for ‘‘Contemporary’’, is the final step which needs to be reached through the interpolation. Both TOC and C are grids in order for the CA to work.

Each grid cell of the CA contains exactly one language family per time step. The language families are coded as integer values, and the grid steps are algorithmic cycles which do not necessarily correspond to a real time unit. The number of time steps can therefore be chosen deliberately. The transition rule of the CA uses a Moore neighbourhood (see Figure 4.1). If $w_{i,j}$ represents the weight of the eight neighbouring cells of $P_{i,j}$ and $P_{i,j}$ itself, $P_{i,j}$ copies the current language family of one of its neighbours or its previous own value based upon this weighing scheme. The language family of each grid cell can change once per time step according to the basic transition rule:

$$w_{i,j} = \begin{cases} \frac{1}{9} & \text{for a current neighbour or } P_{i,j} \text{ itself} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

However, this version of the CA (using equation 4) turns out to struggle to converge towards C. I therefore modify the basic transition rule to also consider the language family of each cell in C. The eight neighbouring cells and $P_{i,j}$ itself have now a higher weight if their cell value matches the corresponding cell value in C. For example, if a neighbouring cell of $P_{i,j}$ contains the same language family, e.g., Indo-European, in both the current time step t and C, the weight of that cell grows tenfold. That means that there is a ten times higher chance for Indo-European being copied from that cell into $P_{i,j}$ at time step $t+1$. This variation of a CA uses non-random transition rules, which would be problematic if used as the sole method for inference. However, in this thesis, the CA only serves as a way to generate an initial evolutionary history for the MCMC, therefore this issue can be neglected. The updated transition rule used in the CA hence is:

$$w_{i,j} = \begin{cases} \left(\frac{1}{9}\right) * 10 & \text{if the value at } t \text{ matches the value in the C map} \\ \left(\frac{1}{9}\right) * 1 & \text{if the value at } t \text{ does not match the value in the C map} \\ 0 & \text{for cells that are neither a direct neighbour nor } P_{i,j} \text{ itself} \end{cases} \quad (5)$$

Moreover, several different grids are iterated at each time step and compared to C to find the most fitting one. For this comparison, the number of grid cells containing the same value as in C are calculated for each of the produced grids at a time step t . The grid with the highest amount of identically filled out grids cells is kept within the interpolation process. This means that this grid is the output data from which the next time step $t+1$ is randomly iterated. The reason behind this comparison is the prevention of too many and abrupt changes between the last iterated time step and the C map, i.e., an irregular progression of the interpolation.

An example of the CA used to gain a potential interpolation between two given spatial distributions representing a ‘‘past’’ and a ‘‘present’’ map with randomly generated test data can be seen in Figure 4.2. Only two language families – represented by the colors green and orange – are present. Over the course of the four time steps, the original spatial language distribution gets transformed into a distribution very similar to the given ‘‘present’’ map. Furthermore, the grid at the last iterated time step (time step 4) is already very similar to the given final language distribution and the interpolation process overall seems to be fluid.

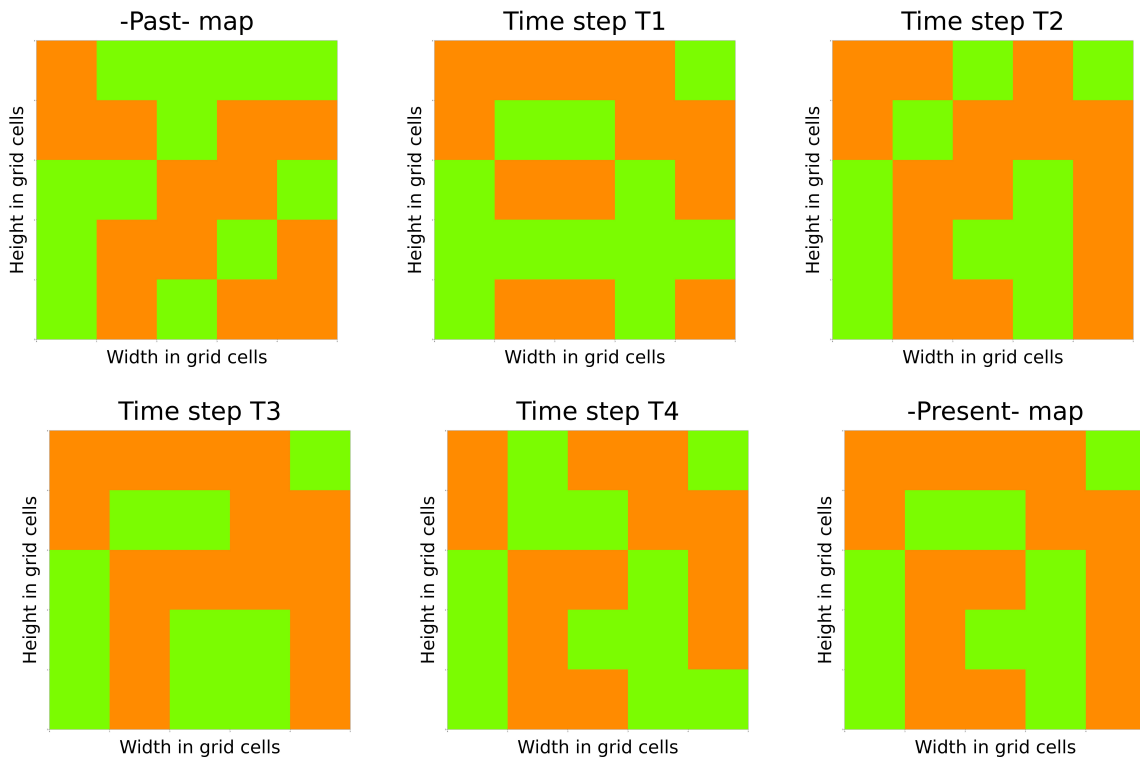


Figure 4.2: Example of the CA used to gain a potential interpolation between two given spatial distributions representing a "past" and a "present" map. Only two language families – represented by the colors green and orange – are present and the CA comprises four time steps.

4.1.2 Pre-processing the language data for the CA

Using my CA on the real language data (see chapter [3.2](#)) requires pre-processing. All language families within the pre-processed data are kept despite only focusing on the expansion of the Indo-European language family at the cost of non-Indo-European language families as a whole for the development of my interpolation method. This allows me to show in a later excursus (see chapter [7.2](#)) that the CA also works for non-binary data. A feature which will hopefully increase the attractiveness of the here developed method for future work with spatial language distributions.

In an initial step, I load the two .JSON language data sets into Python and turn them into geodataframes. The two geodataframes are then rearranged so that they are ready for further calculations. The most important step here is to convert both geodataframes into the WGS 84 coordinate system.

In a second step, since the CA only works with numerical values, I filter out each unique language family in both the "Time of Contact" (TOC) and the "Contemporary" (C) geodataframes and create a dictionary allocating each of these language families a unique integer value, starting at 1. Through the dictionary, it is moreover possible to conserve the information of which newly assigned integer refers to which language family. Among the 115 detected language families, there are two categories representing no proper language family. Instead, the two "language families" "Unclassified" and "Bookkeeping", as defined by Glottolog (Hammarström et al. [2022](#)), include languages which have either not been properly classified or have in the meantime been reclassified. However, since my goal is to develop a probabilistic method for interpolating spatial

language distributions and not to improve the data sets I got, I will consider "Unclassified" and "Bookkeeping" to be proper language families. As I will later on focus on the expansion of the Indo-European language family at the cost of non-Indo-European language families as a whole, this should also not impact my results. The freshly allocated integer values – called integer codes in the following sections – are added in a supplementary column to both geodataframes and the geodataframes are regrouped around that column.

Since the digitized language data do not continuously cover all of South America, I need to deal with these data gaps. Otherwise, after rasterizing the geodataframes, there will be void cells dispersed over the continent. To do so, I import the 1:50m "Land" shapefile by Natural Earth and add it as first entry to both geodataframes. Since during the following rasterization process of the geodataframes, entries situated more towards the bottom of the geodataframes are superimposing entries situated more towards the top of the geodataframes, putting the "Land" shapefile in first position means that it will only be visible for areas where no other language data is available. As it got assigned its own integer code (0), all areas with no language data in the original .JSON files will therefore be treated as one additional language family within the CA: "Land Default". However, as this research does not focus on single language families but on Indo-European and non-European as whole, this does not impact my results.

The superimposing of the further down entries in the geodataframes during the rasterization process also deals with the issue of multilingualism within the TOC and C grids rasterized for the CA. Indeed, if several language families are overlapping, only the language family situated the furthest down in the corresponding geodataframe will be retained during rasterization. In case of multilingualism, I want to keep the language family with the smallest speaker ranges in order to represent the highest possible phylogenetic diversity within my CA. To do so, I add a supplementary column to both geodataframes for which the area of each language family's speaker ranges is calculated. After that, the two geodataframes are rearranged by descending area size. Eliminating multilingualism reduces the complexity of reality and will lead towards information loss in the results. However, the advantage of it is simplicity, which makes inference – and therefore the development of my probabilistic interpolation method for spatial language distributions – possible.

However, due to my decision concerning multilingualism, five Indigenous language families with medium sized speaker ranges – Andoque, Kanoê, Naduhup, Taushiro, and Waorani – are completely superimposed by language families with smaller language ranges in the TOC geodataframe, but not in the C geodataframe. This can be explained by the fact that most of the small language families go extinct between TOC and C while the middle-sized ones are more prone to survival. Since the developed CA does not allow new language families to be added during the interpolation process, I need to reintroduce these five languages, present in C but not in TOC, into TOC. This is done by later implementing seeds into the rasterized TOC language grid in a similar fashion as for the European settlement seeds (see chapter 3.3). To do so, the coordinates for the five seeds are first determined based upon the language families' point locations in Glottolog (Hammarström et al. 2022). Then, knowing the coordinates of the bounding box's vertices as well (see chapter 3.1.1), I determine the grid cell corresponding to each of the five coordinates and implement a 5x5 cell-sized language seed around that grid cell.

Furthermore, three language families – Misupalman, Puri-Coroado, and Timote-Cuica – are present neither in TOC nor in C due to the multilingualism solution. All three of them are, for completeness reasons, being kept in the dictionary, but ignored within any further calculations. Two more language families – Nuclear Trans New Guinean and Sino-Tibetan – are present in

C but not in TOC. However, these language families were and are not present at all in South America. Instead, they were wrongly identified somewhere during the map production process and are hence discarded.

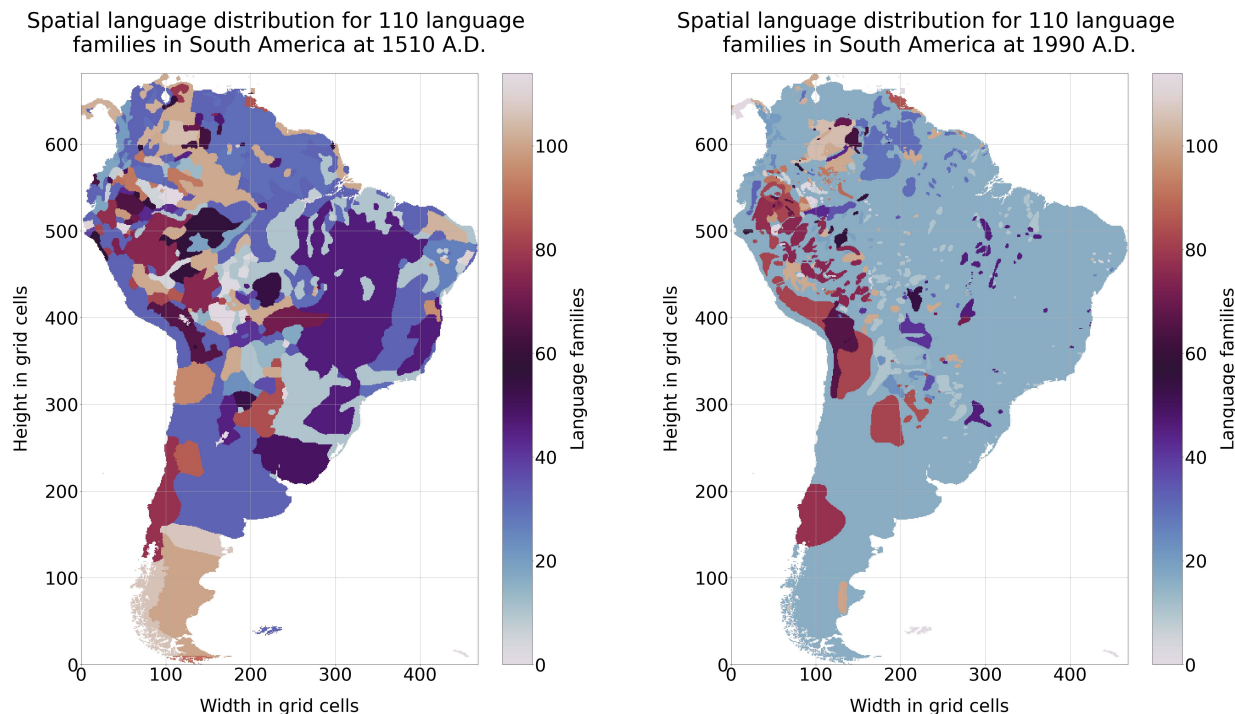


Figure 4.3: Rasterized spatial language distributions for the retained 110 language families in South America at 1510 A.D. and 1990 A.D.

The final step in pre-processing the South-American language data is to rasterize the two geodataframes. To do so, I use the bounding box defined in chapter 3.1.1 and a resolution of 0.1° . After adding the Indo-European language family into the TOC grid by implementing the first European settlement seed, the language grids contain the same 110 language families each (see Figure 4.3). The grid cells within the boundary box which represent the ocean contain no value, i.e., are void, to prevent the language families from spreading into the ocean. The dictionary comprises two wrongly classified and eliminated language families and three language families which are hidden in both TOC and C in addition to the 110 language families. TOC translates to 1510 A.D. (the founding year of the first implemented European settlement, Santa Mariá Del Darién) and C to 1990 A.D. (the last update of Kaufman’s contemporary map happening around 1989, see chapter 3.2). The shape of both language grids is 682 x 468 cells.

4.1.3 Running the CA with the language data

To run the CA with the language data for South America, I use the two pre-processed language grids for TOC and C, described in chapter 4.1.2. An interpolation between these two spatial language distributions is only possible if both contain exactly the same language families.

In a first step, I reduce the amount of language families to work with to two: Indo-European gets integer code 1 and all the Indigenous language families get integer code 0. The reason for

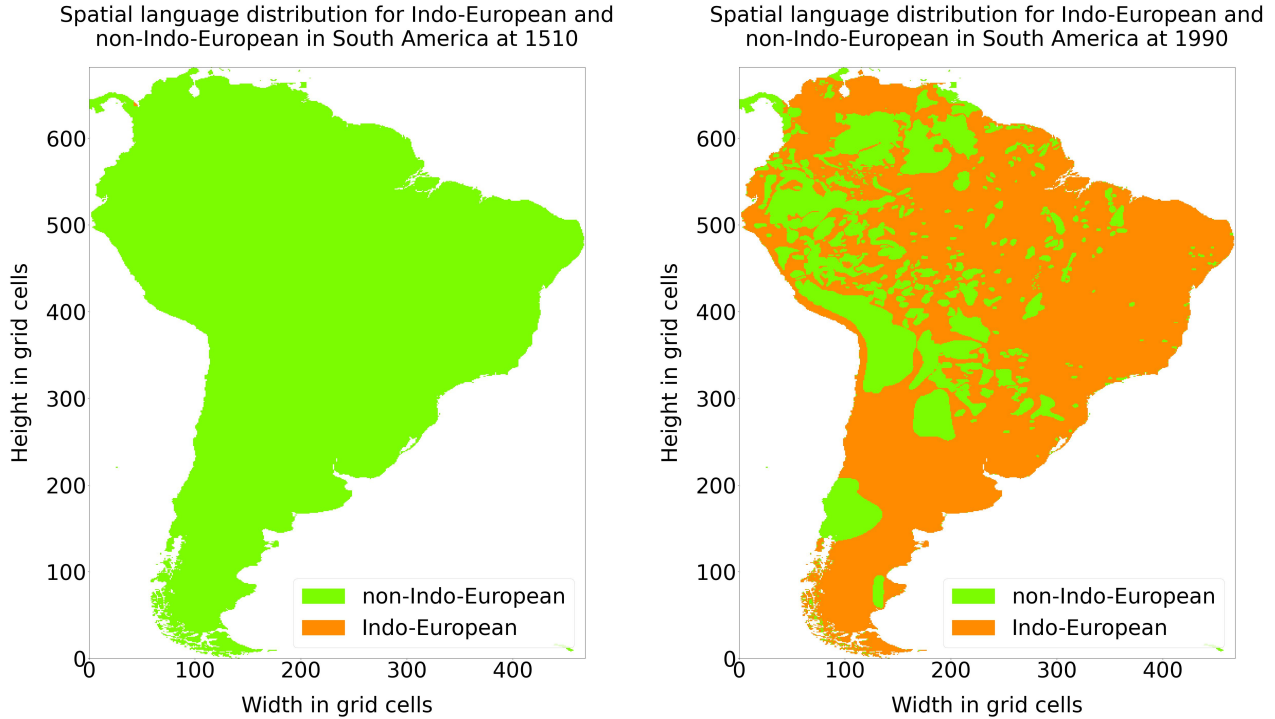


Figure 4.4: Rasterized binary "TOC" and "C" spatial language distribution maps of South America between which the CA interpolates.

this binary approach is to reduce the computation time within the Bayesian inference's MCMC later on. The binary TOC and C maps, between which a first potential interpolation is inferred, can be seen in Figure 4.4

The CA runs with equation (5) (see chapter 4.1.1) and 241 time steps. The first and the last time step represent the TOC and C map, which leaves 239 in-between time steps which each represent two real years. This amount of time steps is chosen because, between the various constellations I tested, it allows for the most regular progression of the interpolation, i.e., for the Indo-European language family to most fluidly reach the extension known from the C map. Furthermore, three different grids are iterated at each time step and compared to the C map to find the most fitting one. This relatively low number of comparisons is due to the fact that the interpolation inferred through the CA is only a non-deterministic input data set for the MCMC. Using more comparisons would therefore be obsolete – as the MCMC later maximizes the likelihood – and increase computation time with only minor gain.

Computation time is also a general issue when running the CA with the real language data. In order to reduce it, the language grids are hot-encoded at every time step t and fed into the CA to infer the most probable language grid at the subsequent time step $t+1$. The hot-encoding was chosen because, within the CA, it allows to perform a convolution for each of the two language family values – 0 and 1 – in order to obtain the probability with which each of the two values is copied towards $P_{i,j}$. The convolution is conducted with a 3x3 Moore neighbourhood mask as described in chapter 4.1.1. However, the convolution mask has reduced diagonal spreading to avoid too rectangular shapes for the inferred spatial language distributions. Specifically,

that means that the convolution mask does not have the shape $[[1, 1, 1], [1, 1, 1], [1, 1, 1]]$, but $[[\frac{1}{\sqrt{2}}, 1, \frac{1}{\sqrt{2}}], [1, 1, 1], [\frac{1}{\sqrt{2}}, 1, \frac{1}{\sqrt{2}}]]$.

After having inferred and compared three different language grids and the one to be kept for time step $t+1$ has been chosen, this language grid's hot-encoding is reversed. Indeed, a 2-dimensional version of the grid is needed to potentially implement further European settlement seeds, to write out the grid into a .TXT file as input for the MCMC and to plot the intermediate step of the interpolation process.

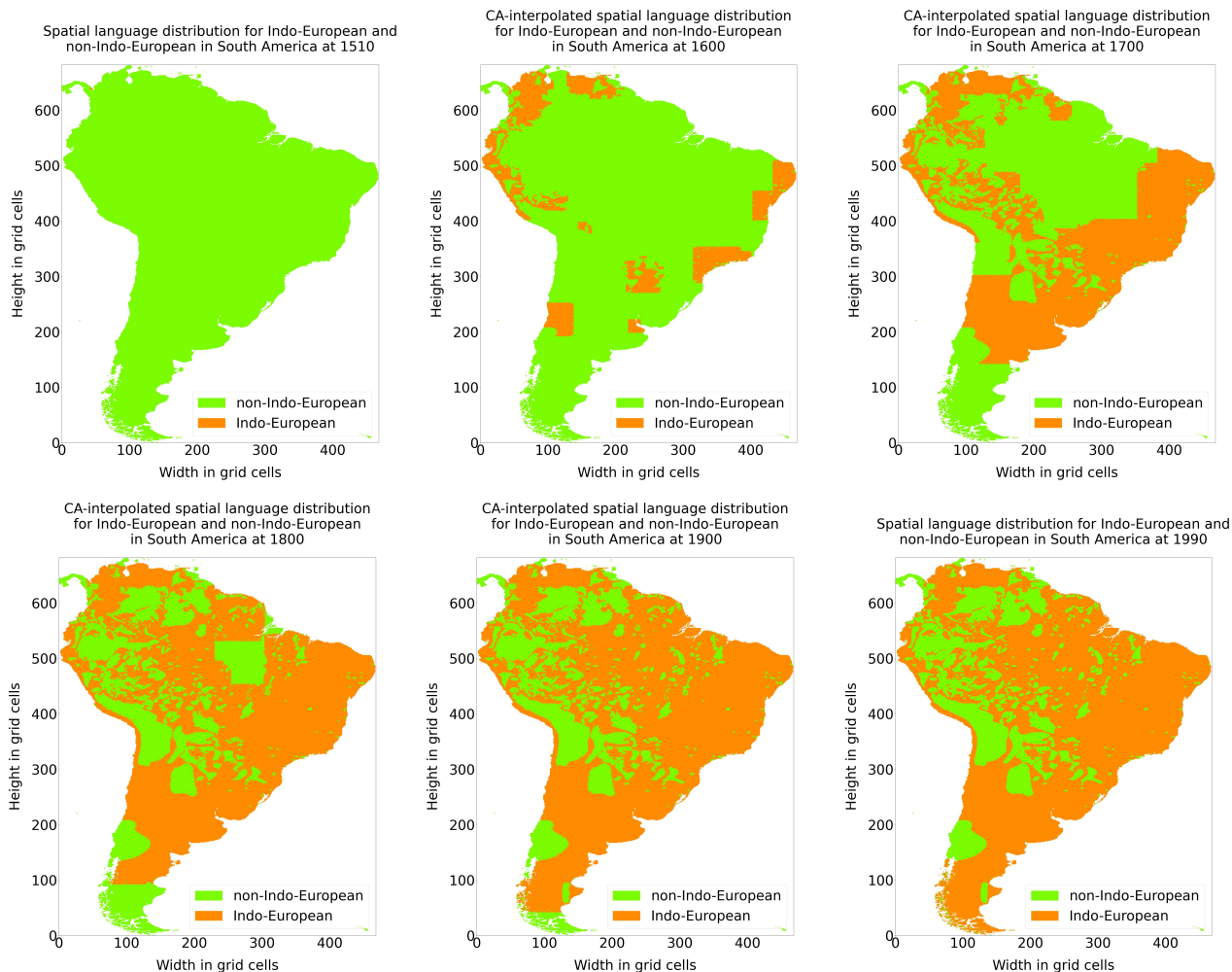


Figure 4.5: CA-induced interpolation steps for binary spatial language distributions at 100-year-intervals.

Let's start with the potential implementation of further European settlement seeds. Since one time step corresponds to 2 real years, not every new-founded city can be added in the correct year. Instead, some cities have to be implemented into the inferred map depicting the year closest after the corresponding settlement's foundation date. For example, the settlement seed for Santa Marta can only be implemented into the map representing 1526 despite the city already having been founded in 1525 (Kline et al. 2023; Wallenfeldt 2022). All European settlement seeds are implemented as a 5x5 cell-sized language seed around the grid cell corresponding to the central x,y-coordinates of the corresponding settlement (see chapter 3.3). European settlement seeds are not implemented at every time step.

After the potential implementation of further European settlement seeds, I need to write out the language grid for time step $t+1$ into a .TXT file as input for the MCMC. While the integer codes for Indo-European and non-Indo-European remain the same for the MCMC data input, the grid's void values need to be encoded as . and, if European settlement seeds have been added in time step $t+1$, these have to be encoded separately as # (for more details, see chapter 4.2).

Subsequently, the inferred spatial language distribution for time step $t+1$, including potential European settlement seeds, is plotted. Afterwards, the only just plotted 2-dimensional language grid is hot-encoded and fed into the CA to allow the inference of the most probable language grid at the subsequent time step $t+2$. So the process begins anew until a spatial language distribution has been inferred for all 239 in-between time steps.

The so-produced complete evolutionary language history containing 239 inferred time steps and the given TOC and C maps is a first potential interpolation of spatial language distributions between the the binary TOC and C maps for South America. While this evolutionary language history is still non-probabilistic, it is the so-called “starting history” which will be fed as input data into the MCMC (see chapter 4.2). Excerpts of the CA-induced binary “starting history” can be seen in Figure 4.5.

4.2 MCMC for Bayesian inference

The output of the CA, a first evolutionary language history called “starting history” and saved in 241 .TXT files (one for each time step), is the input needed to run the Markov Chain Monte Carlo (MCMC) method within my probabilistic interpolation method for spatial language distributions. Indeed, an MCMC needs good input data to start its journey through parameter space. The use of the MCMC is necessary since my Bayesian inference model contains many parameters and MCMC is typically the tool of choice in such cases.

The “starting history” is a complete evolutionary language history of South America comprising the 239 inferred in-between language distributions and the two given spatial language distributions at 1510 A.D. and 1990 A.D.

The question my Bayesian model wants to answer is: “What is the distribution of probable evolutionary histories between 1510 A.D. and 1990 A.D. in South America, given the prior knowledge of the spatial language distributions between 1512 A.D. and 1988 A.D. and the input data of the spatial language distribution at 1990 A.D.?” Knowing that my Bayesian model uses sampling as data-generating process, all parameters can be inferred from the input data by using an MCMC. To do so, in a first step, the Bayesian theorem is established as described in equation (2) on page 9. In the theorem, the term “in-between evolutionary language history” does not refer to the complete evolutionary language history, but only to the 239 interpolated in-between language distributions. Furthermore, the abbreviation SLD for “spatial language distributions” is used to keep the equations short. Finally, the number of parameters, i.e., the parameter space, is $n * (T - 2) = 76'921'416$ with $n=319'176$ grid cells and $T=241$ total time steps.

- **Posterior probability:** The posterior probability is the probability distribution of the in-between evolutionary language history of South America between 1512 A.D. and 1988 A.D. given the input data of the spatial language distributions at 1990 A.D. Or, put in a more technical way, the posterior probability is, given the aforementioned input data, the

probability distribution of the $n^*(T-2)$ parameters.

$$P(\text{parameters} \mid \text{data}) = P(\text{SLD between 1512 and 1988} \mid \text{SLD at 1990}) \quad (6)$$

- **Likelihood function:** The likelihood function is the probability of getting the spatial language distribution of South America at 1990 A.D. given the in-between evolutionary language history of South America between 1512 A.D. and 1988 A.D. Or, put in a more technical way, the likelihood function is the probability of getting the spatial language distribution of South America at 1990 A.D. given the probability distribution of the $n^*(T-2)$ parameters. This probability $a_{i,j}$ for each cell $P_{i,j}$ at the final time step T can be computed based upon the previous time step $T-1$ and using an equation similar to the basic transition rule for the CA:

$$a_{i,j} = \begin{cases} \frac{1}{9} & \text{for a neighbour at } T \text{ or } P_{i,j} \text{ itself at } T-1 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$P(\text{data} \mid \text{parameters}) = P(\text{SLD at 1990} \mid \text{SLD between 1512 and 1988}) \quad (8)$$

- **Prior probability:** The prior probability is the strong and informative probability distribution of the in-between evolutionary language history of South America between 1512 A.D. and 1988 A.D. based upon the previous knowledge of the history's starting point, i.e., the spatial language distributions at 1510 A.D. Or, put in a more technical way, the prior probability is the probability distribution of the $n^*(T-2)$ parameters where the probability $a_{i,j}$ for each cell $P_{i,j}$ at each time step t can be computed based upon the previous time step $t-1$ and using basically the same equation as for the likelihood function (see [7](#)):

$$a_{i,j} = \begin{cases} \frac{1}{9} & \text{for a current neighbour or previous } P_{i,j} \text{ itself} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$$P(\text{parameters}) = P(\text{SLD between 1512 and 1988} \mid \text{SLD at 1510}) \quad (10)$$

- **Marginal likelihood function:** The normalizing constant which can be omitted is the input data, i.e., the spatial language distribution at 1990 A.D.

$$P(\text{data}) = P(\text{SLD at 1990}) \quad (11)$$

In a next step, I would implement the MCMC fitting this Bayesian model. However, due to time and expertise issues, I did not implement the MCMC myself, but instead used the one developed by Dr. Takuya Takahashi. An outline for this MCMC, programmed in C++, can be found in the following GitHub repository: https://github.com/takuya-tkhs/language_diffusion/blob/main/MCMC_idea_for_publication.pdf.

I will however give a short overview over MCMC in general and some specific characteristics of the one implemented by Dr. Takuya Takahashi in order to make my future results more easily understandable.

MCMC is a highly efficient algorithm to draw samples from a high-dimensional target distribution (McElreath 2016). A fact which makes MCMC a very promising approach in my attempt to develop a probabilistic interpolation method for language distributions since my target distribution in the Bayesian model - the posterior distribution according to equation (6) - is indeed highly complex. While the MCMC simulates its samples out of the real-world target distribution, the new samples are always based upon information from the previous sample (McElreath 2016). This has however the effect that the samples are correlated instead of independent (McElreath 2016), an important notion to keep in mind when analyzing the MCMC's results. Finally, the MCMC reaches a stationary state at some point, i.e., it reaches a point where, within acceptable error, the probability distribution of the posterior will not change anymore. Reaching this point is called "convergence" and describes a "final and good" target distribution. To reach convergence, a lot of experimenting concerning the necessary amount of sample steps is necessary (McElreath 2016).

Additionally, a few more characteristics are included in the MCMC implemented by Dr. Takuya Takahashi. On the one hand, due to computation time limits, the MCMC can only deal with binary data, i.e., only take in two different "language families": Indo-European and non-Indo-European. However, four different values are still included in the "starting history" and sampled by the MCMC: 0, 1, # and . While the two numbers represent non-Indo-European and Indo-European respectively, the . stands for voids, i.e., cells with no value. As within the CA, the only cells containing voids are the ocean cells. The # however represents at a time step t all grid cells in which Indo-European settlement seeds have been implemented at that time step. The freshly implemented settlement seeds are marked specifically so that they cannot be eradicated by the MCMC in that same time step t . From time step $t+1$ onward, the settlement seeds marked with # at time step t are then represented by 1.

5 Results

The MCMC for the Bayesian model samples 100 evolutionary histories which are saved in 100 .TXT files containing each the 241 time steps of one history. These 100 evolutionary histories are the results of the newly developed probabilistic interpolation method for spatial language distributions.

For further processing and visualizing the results, the 100 .TXT files are read into Python and the data saved in two dictionaries. The first dictionary represents the stacked samples while the second dictionary represents the stacked time steps. This allows to visualize the spatial extension probability of the Indo-European language family as dominant language in South America for each of the 241 time steps: for each cell at each time step, the probability of the spatial extension is calculated based upon the 100 evolutionary histories. A selection of the 241 time steps can be seen in Figure [5.1](#).

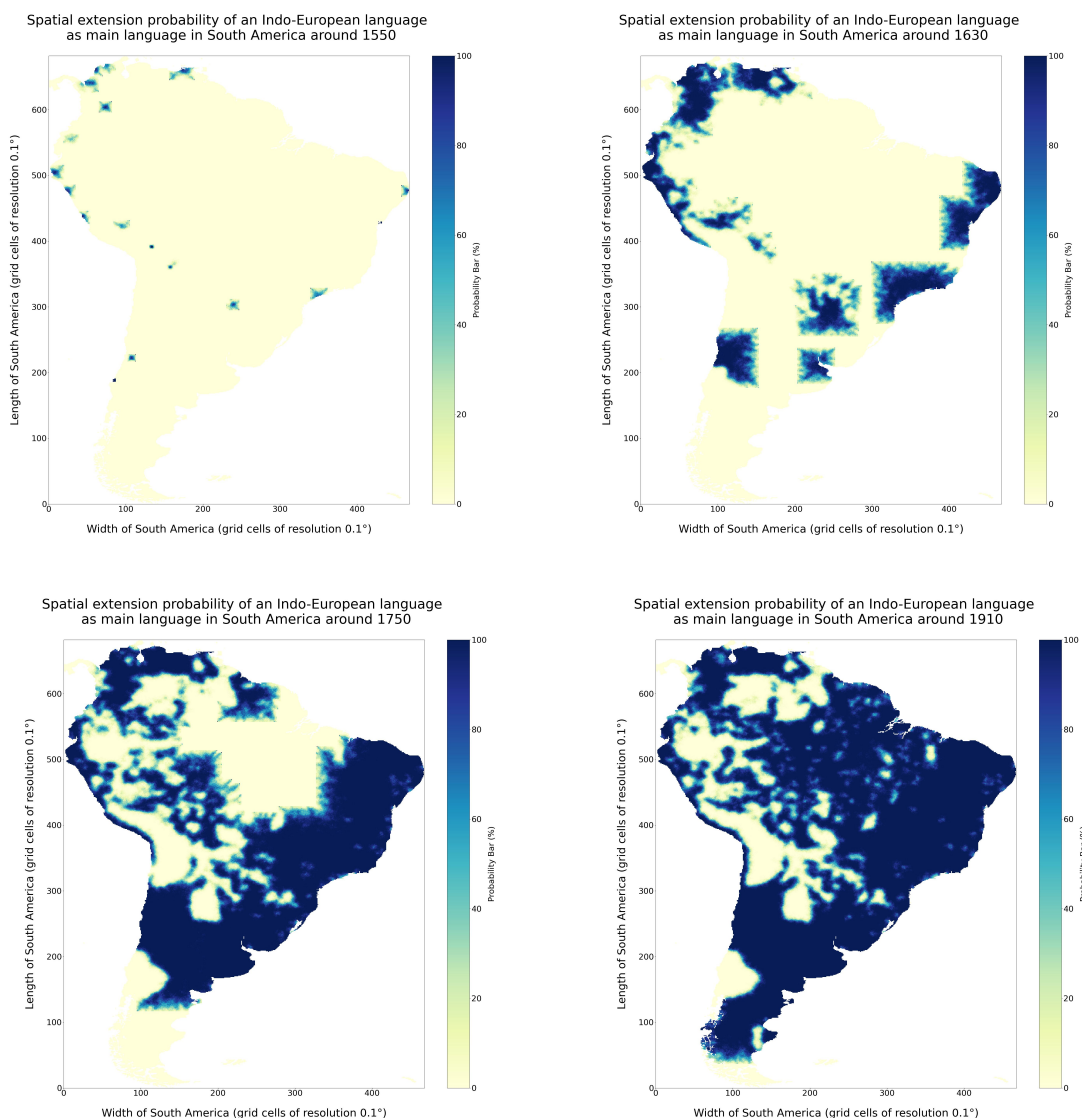


Figure 5.1: Spatial extension probability of the Indo-European language family as dominant language in South America for selected years.

The depicted spatial extension probabilities (see Figure 5.1) show that, for all the interpolated evolutionary histories, the spreading patterns of the Indo-European language family always seem to unfold from the coasts towards the centre of the continent. Furthermore, the Indo-European language family seems to first spread in the Northern half of South America with a strong focus on the Western coast. Indeed, the North-East of the continent looks mostly untouched until the 1750s with the Indo-European language family only spreading in the lower half of modern-day Brazil. While these inferred spreading patterns are of course also related to the placement of the European settlement seeds, they nevertheless show that the performed spreading mechanism via neighbouring cells is able to produce meaningful results. Indeed, the interpolated evolution histories follow the historical colonization pattern of South America (see Figure 5.2), a pattern which is most likely strongly related to the language spreading of Indo-European on the continent.

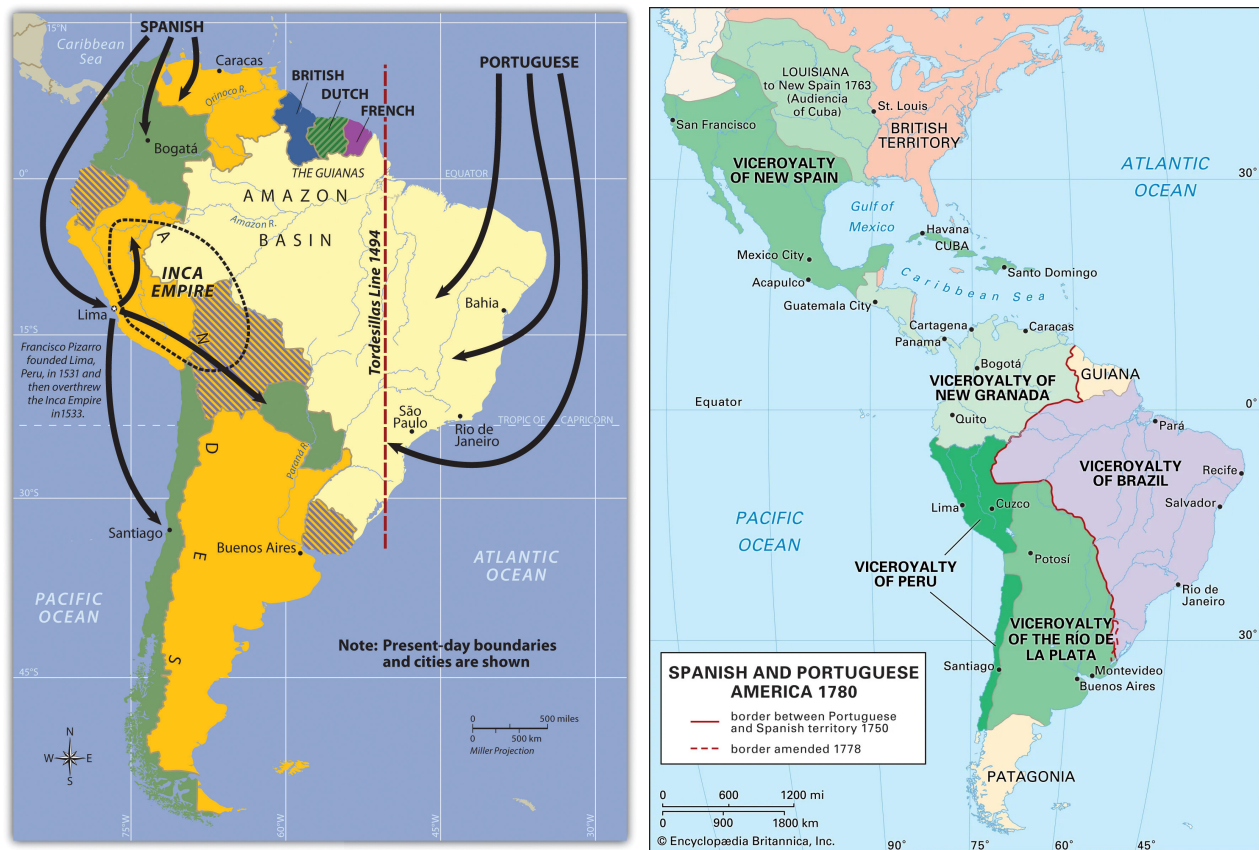


Figure 5.2: Early colonization patterns in the 16th century (left) and Iberian colonies around 1780 in South America (right). (Sources: Dastrup 2020 (left), Britannica 2019b (right))

The spatial extension probability around 1550 furthermore vividly depicts how the Indo-European language family spreads from the implemented European settlement seed points: while the centres of the seed points have a spatial extension probability close to 100%, the surrounding neighbouring cells have very diverse probabilities and proceed according to various spreading patterns. For example, Indo-European mostly expands towards the West of Cumaná, a settlement in North-Eastern Venezuela, while progressing in a circular shape around Bogotá.

The spatial extension probability around 1630 however shows intriguing square-shaped spatial language distributions. This is due to a modelling restriction of the CA, which will be discussed in more detail in chapter 6.2.2.

The spatial extension probability around 1750 indicates that the Indo-European language family already dominates in the areas it also dominates today after roughly 250 years, i.e., after half the time between the “Time Of Contact” and today, with the exceptions of the Amazon Basin and the Southern tip. The interpolated spreading and rise in importance of Indo-European in South America hence mostly already happened in the first two centuries of colonization, with the spreading rate seeming to slow down afterwards. This indicates that the interpolation method is most likely heavily dependent on the choice of the European settlement seeds. Indeed, the Amazon Basin and the Southern tip are the only regions of South America where no settlement seeds were implemented. Therefore, a good starting point to improve the interpolation method would probably be to reconsider the chosen European settlement seeds - and their number - with historical expertise.

Finally, the spatial extension probability around 1910 shows that the areas in which Indo-European is dominant are rather well defined towards the end of the interpolation process: the borders between them and the non-Indo-European areas are sharply delimited with only the border cells between the two areas having in-between probabilities.

However, the lowest divergences between the spatial extension probabilities of the 100 evolutionary histories happen already around 1850 as well as at the very beginning of the interpolation process (see Figure 5.3). The explanation for the latter is that, at the beginning of the interpolation process, only very few settlement seeds are yet implemented. Since those are furthermore still small – as each time step only allows to spread into the neighbouring cells –, the Indo-European language family can only spread from a very limited amount of cells. Therefore, the spreading is rather straight forward without many divergences between the different evolutionary histories.

The explanation for the low divergence around 1850 – with an approximate Indo-European grid cell proportion of 70% – is most probably the reaching of a local maximum within the interpolation process. A hypothesis which is backed by the previous observation that, already around 1750 and with the exception of the Amazon Basin and the Southern tip, the Indo-European languages seem to be dominant in almost all of the areas where they are also dominant nowadays. Furthermore, some consequences of a local maximum can be observed after 1850 concerning the proportion of grid cells within South America primarily speaking Indo-European. Indeed, in some evolutionary histories, Indo-European languages then also seem to spread into cells which are not filled with Indo-European in the final, given spatial language distribution from 1990. This would explain why one can observe the grid cell proportion dipping down again around the late 1980s for the evolutionary histories where this happens (see Figure 5.4). In other evolutionary histories, after 1850, the Indo-European languages still spread mostly into cells which are also filled with Indo-European in the final, given language distribution from 1990. Therefore, since only a few grid cells are left to be filled between 1850 and 1990, the curve of their Indo-European grid cell proportion increases steadily but with a minimal slope until 1990 (see Figure 5.4).

The likely forming of a local maximum within the interpolation process is probably due to a too high number of time steps. However, the 241 time steps were necessary for the Indo-European language family to reach the Southern tip of South America between 1510 and 1990. If one decides to further work with the probabilistic interpolation method for spatial language distributions, I would therefore recommend to also insert, in consultation with a historian, seed points for European settlements founded later than 1600. Since some of them will most likely be located in the Southern tip of South America, this should then allow for a faster and more

accurate spreading pattern of Indo-European in that area. Subsequently, the number of time steps should be reducible and a potential local maximum in the interpolation process avoided.

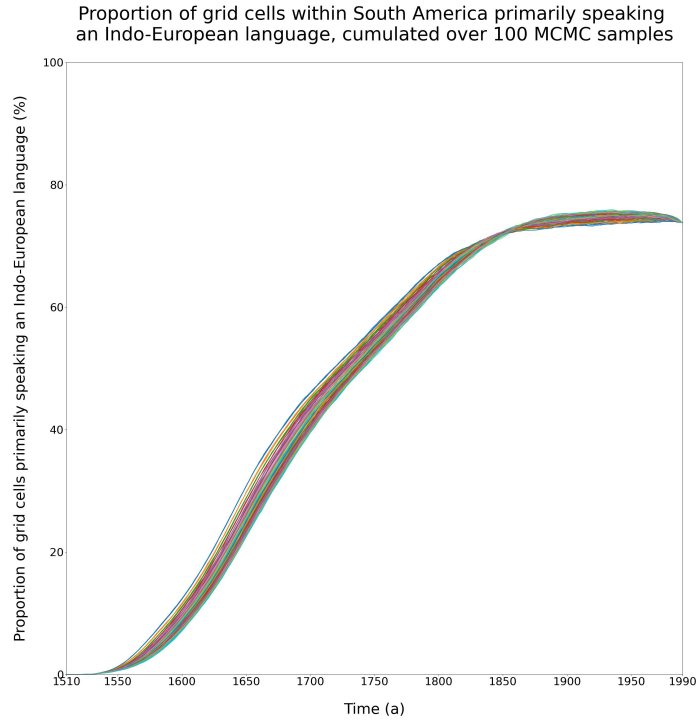


Figure 5.3: Proportion of grid cells within South America primarily speaking an Indo-European language, cumulated over 100 samples/evolutionary histories.

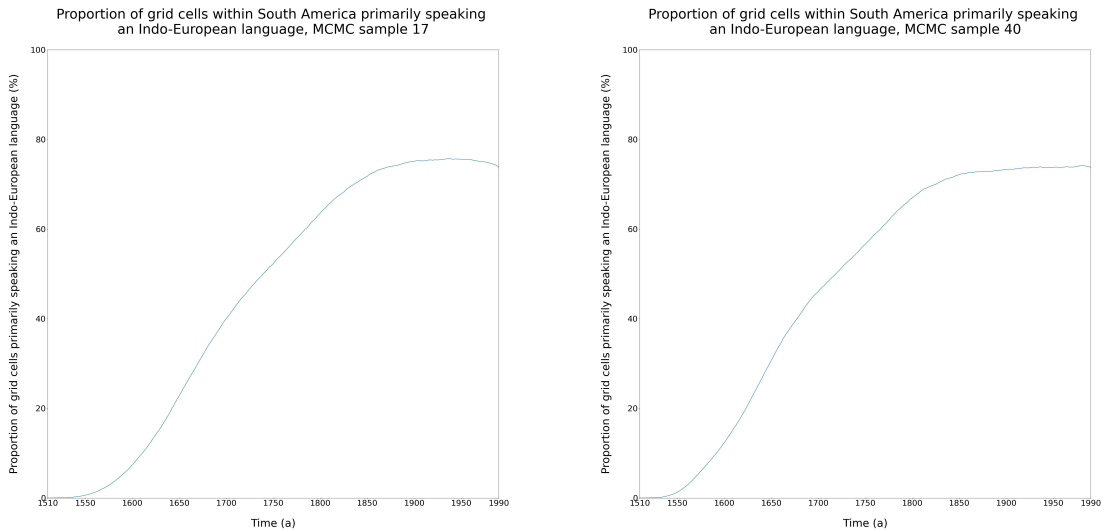


Figure 5.4: Proportion of grid cells within South America primarily speaking an Indo-European language for selected samples/evolutionary histories. Left: medium grid cell proportion increase after 1850 with a sudden dip around 1990. Right: steady, but very slow grid cell proportion increase after 1850.

6 Discussion

6.1 Choice of methods

In the following, I will shortly explain why, among all the options I had, I chose the previously presented methods (see chapter 4) for developing my probabilistic interpolation method for spatial language distributions.

6.1.1 A process-based simulation model (PBM) as overall concept

As overall concept, I decided to use a process-based simulation model (PBM) (see chapter 2.2.1) since these have recently very promisingly been introduced in several studies assessing language diversity patterns (see chapter 2.2.2). Using a PBM as overall concept furthermore gave me the freedom to freely choose my statistical method, i.e., Bayesian inference, within it. Unfortunately, the biggest asset of PBMs – focusing on assessing the causal impact of selected driving factors of language spreading – did not come into play in this thesis: due to time and computation restraints, I had to forfeit my original plan to implement the chosen factors and processes of language spreading into the MCMC in order to later causally assess their impact on the generated spatial language distributions.

However, knowing that the direct effect these factors and processes have on the spatial language distribution at the in-between steps of the interpolation is of high importance – it would not only allow to assess the importance of these factors in the regional context of South America, but also to determine potential factors which could enhance the accuracy of my probabilistic interpolation method now and in the future –, I instead give an outline on how to implement selected driving factors with real data within the MCMC (see chapter 7.1). This knowledge should prove useful in case someone decides to further use my probabilistic interpolation method for spatial language distributions.

Therefore, I still consider the concept of PBMs very important for this thesis: they did not only provide me with the original spark for my work, but will hopefully still help to fully unfold the potential of the developed interpolation method.

6.1.2 A cellular automaton (CA) as core underlying mechanism

Grid structures have already been successfully applied as basis for working with spatial language distributions in the studies by Gavin et al. (2017), Pacheco Coelho et al. (2019) and Pacheco Coelho et al. (2021). Pacheco Coelho et al. even introduced the idea of algorithmic, artificial cycles within the grid (2021). This brought me to the idea of using a cellular automaton (CA) (see chapter 2.3) as core underlying mechanism within my PBM. Especially since CA in specific have already successfully been used over a longer period within language studies, e.g., in the language shift study by Beltran et al. (2010). The grid-like nature of a CA allows to model the spatial language distributions in South America as raster with discrete time steps, e.g., 2 years. The use of these discrete time steps – algorithmic cycles, only indirectly related to real time – instead of real years was important as it offered me more experimental variability during modelling. Another advantage of using a CA was the clear language family attribution. Since I ignored multilingualism and only attributed a single language family to each area – like in the study by Pacheco Coelho et al. (2021) –, the fact that each cell of the CA can only contain a single value per time step fitted perfectly. I am aware that that this approach reduces the complexity of reality and led towards errors in the results. However, the advantage of it is simplicity, which

made inference – and therefore the development of my probabilistic interpolation method for spatial language distributions – possible.

6.1.3 Bayesian inference as statistical method

I decided to use Bayesian inference as statistical method since it is well tailored to my case of learning and deriving, i.e., interpolating, from existing real-world language data. Moreover, Bayesian inference can be extended to model processes of almost any complexity. Therefore, my interpolation between two given states consisting of about 300'000 different cells each was still feasible. Furthermore, the fact that the Bayesian inference can capture the uncertainty of the modelling process was a huge asset since the interpolation process is heavily dependent on the chosen input data, e.g., language data, settlement seed choices, making it therefore a rather uncertain process. Finally, the results of a Bayesian inference are straightforward to interpret. This was an especially big advantage as I was not modelling and implementing the Bayesian inference for this thesis myself, but instead feeding my data set into an existing MCMC.

6.2 RG: Development of a probabilistic interpolation method for spatial language distributions

In this chapter, I will discuss the research goal:

RG: How is it possible to interpolate spatial language distributions in a given area and over a certain time span with a known distribution both in the beginning and in the end?

As the results (see chapter 5) show, with the combined use of a cellular automaton and Bayesian inference, it is possible to interpolate spatial language distributions in South America over a time span of roughly 500 years given only the distributions around 1510 and 1990. While the interpolation process is of course underlying data uncertainty (see chapter 6.2.1) and restrictions due to modelling choices and limitations (see chapter 6.2.2), the overall results look very promising. To further enhance the accuracy of the sampled evolutionary histories, it would however be important to add spatial factors and processes of language spreading as the example of the Amazon Basin (see chapter 6.2.3) shows. Adding such factors and processes is also backed by literature, which acknowledges their importance in the development of spatial language distributions over time (see chapter 2.1.2). Furthermore, if one embeds the developed interpolation method into a process-based model simulation (see chapter 2.2), the importance of each added spatial factor of language spreading could also be assessed in a regional context. This would then not only allow for a generally enhanced accuracy of my interpolation method, but also for locally optimized interpolation processes. An outline on how to implement geographical factors within the developed interpolation process is therefore given in chapter 7.1.

6.2.1 Data uncertainty

The most important issue is the incompleteness of the digitized polygon data set. The large percentage of both "Unclassified" and "Bookkeeping" – language family constructs containing languages which have either not been properly classified or have in the meantime been reclassified – and the substantial amount of void polygons are a result of this. Furthermore, there are wrongly identified language families as the examples of Nuclear Trans New Guinean and Sino-Tibetan show: both language families are depicted as being part of the C map while in reality, they were never and are not present in South America. While the language data's incompleteness

is mostly a problem when representing more than two language families in the interpolation process - as suggested for future research, see chapter [7.2](#) -, it may also lead to smaller errors when only distinguishing between Indo-European and non-Indo-European language families.

The same goes for the way I chose to deal with multilingualism in a given area. Since the digitized language polygons can overlap or contain several languages and language families, I had to decide which language family to keep per given area. To do so, I ordered the polygons from the biggest to the smallest in the geodataframe. During the rasterization process of the digitized language polygons, the algorithm gives priority to the later entries, i.e., the smaller polygons, and their integer codes, i.e., codes representing a language family, are therefore attributed preferentially to each grid cell. Hence, the language family of a smaller polygon is preferred over the language family of a larger polygon in case of multilingualism. My way to deal with multilingualism attributes more importance to language families with a smaller spreading area, giving them a chance to also be represented within the language distribution. This choice, giving preferential treatment to small language families and therefore a high language diversity, influences the results in favour of such small language families. This is cemented by the fact that eight language families of middle-sized range – Andoque, Kanoê, Naduhup, Taushiro, Waorani, Misupalman, Puri-Coroado, and Timote-Cuica – are either in both TOC and C or only in C completely superimposed by several language families with smaller language ranges.

Also, overarching large-scale language classifications like Kaufman’s have recently come more and more under pressure in academia. Indeed, the authors of such classifications have often compiled information about which they had little to no personal knowledge. Furthermore, in numerous instances, the classifications are “based on little to no evidence for some of the entities they classify”(Campbell [2012](#)). Therefore, the choice of my language data is certainly worthy of discussion in a modern linguistics context.

As already mentioned when presenting my results (see chapter [5](#)), the choice of the European settlement seeds to implement is also a delicate matter. Since the current choice seems to be partially responsible for a local maximum in the interpolation process, historians should be included or at least consulted in future applications of the interpolation method.

Another uncertainty related to the European settlement seeds is the assumption that the founding of a European settlement equals the immediate dominance of an Indo-European language in that area. However, when a European settlement gets founded, it probably takes years to decades before an Indo-European language is dominant in that area. That time span is also highly dependent on both the amount of people speaking other languages in that area and the number of Indo-European speakers immigrating into the new settlement: if only very few people speaking Indigenous language live around the newly founded settlement or many Indo-European speakers immigrate, it will most likely take less long for an Indo-European language to become dominant. Although this assumption is a good first approximation, it needs to be seen with a critical eye and a more refined method to spatially locate dominant language families might be necessary in case of further application of the developed interpolation method.

6.2.2 Restrictions due to modelling choices and limitations

While both the CA and the MCMC depend on the input data and are therefore sensitive to data issues, modelling choices and limitations also lead towards restrictions and potential errors within my interpolation method for spatial language distributions.

First, the chosen resolution for rasterizing the digitized language polygons – 0.1° – appears at

the same time too high and too low. Indeed, a lower resolution and therefore lower amount of cells would have reduced the MCMC's computation time and therefore potentially allowed a higher amount of sampled evolutionary histories. It could in the future also help keeping the MCMC's computation time manageable when interpolating between more than two language families. However, as the example of Belém (see chapter [6.2.3](#)) shows, the current resolution already leads to errors in more detailed analysis. A problem which would only intensify if the resolution gets reduced.

As already mentioned when presenting the results (see chapter [5](#)), the interpolated evolutionary histories all share a rather square-shaped language spreading pattern. Since a CA works with a neighbourhood, the induced spreading pattern reflects the shape of its neighbourhood. In my case - as it is usual - I used a square neighbourhood, i.e., the Moore neighbourhood. Despite weighing the diagonal cells of the Moore-neighbourhood with $\frac{1}{\sqrt{2}}$ instead of 1 to reduce the square-shaped spreading pattern, the Indo-European language family still strongly expands in square-shaped spreading patterns. Despite the CA delivering only the “starting history” for the MCMC, the MCMC does not seem to be able to sample the square-shaped spreading pattern completely out of the evolutionary histories.

Another potential weakness of the CA is the fact that a grid at a random time step t can have more in common with C than the last iterated time step at $t=240$. However this has been adjusted/accounted for with the usage and implementation of the Bayesian inference.

An important issue is that the CA struggles to reach C if it only considers the language family of each neighbouring cell at time step t . In that case, the interpolated evolutionary history only consists of randomly scattered patterns. To avoid this and obtain a meaningful “starting history” for the MCMC, the basic transition rule was adjusted to also consider the language family of each neighbouring cell in C (see chapter [4.1.1](#)). As this variation of my CA uses non-random transition rules, it can never be used as the sole method for inference. However, in combination with Bayesian analysis, it does not pose a problem for the interpolation method.

The MCMC has a very high computation time due to the extremely high number of parameters – 76'921'416, see chapter [4.2](#). This leads to some severe limitations: running the MCMC over three days results in 100 evolutionary histories where each history can only contain binary data, i.e., two language families. To minimize this limitation, the interpolation was reduced to only two language family groups: the spreading of the Indo-European language family at the cost the non-Indo-European language families since colonization heavily impacted South America in the analyzed time span. Discovering more about this process and the way Indo-European drove out Indigenous language families seems not only important from a linguistic perspective, but also with respect to history and social justice.

The MCMC does, furthermore, not properly test for convergence. However, when analysing the 100 samples, it does not seem like the MCMC converges yet. The 100 samples are therefore most likely not enough to reach a stationary state and a “final and good” target distribution, i.e., the posterior distribution of the Bayesian model. For future application of the method, the computation time for the MCMC should be reduced in order to experiment with higher sample numbers and eventually reach a “final and good” distribution of the spatial language distributions between 1512 and 1988 given the spatial language distribution at 1990.

6.2.3 Optimization potential: Example of the Amazon Basin

A good example to analyze how well my developed probabilistic interpolation method for spatial language distributions works is the Amazon Basin (see Figure 6.1). Indeed, the Amazon Basin is a mostly landlocked region whose climate and environment make it rather difficult to access, especially for people not used to it like the European colonizers. The colonization of the Amazon Basin – and subsequently the related spreading of the Indo-European language family – therefore progressed rather slowly and often only at later stages, with a huge peak happening as recently as in the 20th century (Wood et al. 1988). I would therefore expect my interpolation method to represent this slow expansion of the Indo-European language family within the Amazon Basin.



Figure 6.1: Map of the Amazon Basin depicting the Amazon river and its most important tributary rivers as well as important local centres. (Source: Kmusser 2013)

To test this hypothesis, I map the times by which a Indo-European language first reaches selected settlements in the Amazon Basin. The chosen settlements are, with one exception, all located in Brazil since Brazil comprises most of the Amazon Basin. The mentioned exception is Iquitos (Peru), the city situated at the source of the Amazon river. This settlement represents both the beginning of the Amazon river and the other South American countries comprising parts of the Amazon Basin. The selected Brazilian settlements are all state capitals laying in the Amazon Basin while also being important local centres. However, I had to leave out Belém, the capital of Pará, as due to the rasterization resolution, Belém is unfortunately mostly categorized as void ocean cells in my evolutionary histories and no spatial language information is available for it. In total, five different settlements were selected (see Figure 6.2): Macapá, Manaus, and

Iquitos, which are all located on the Amazon river, as well as Rio Branco and Porto Velho, which are located on smaller tributary rivers of the Amazon.

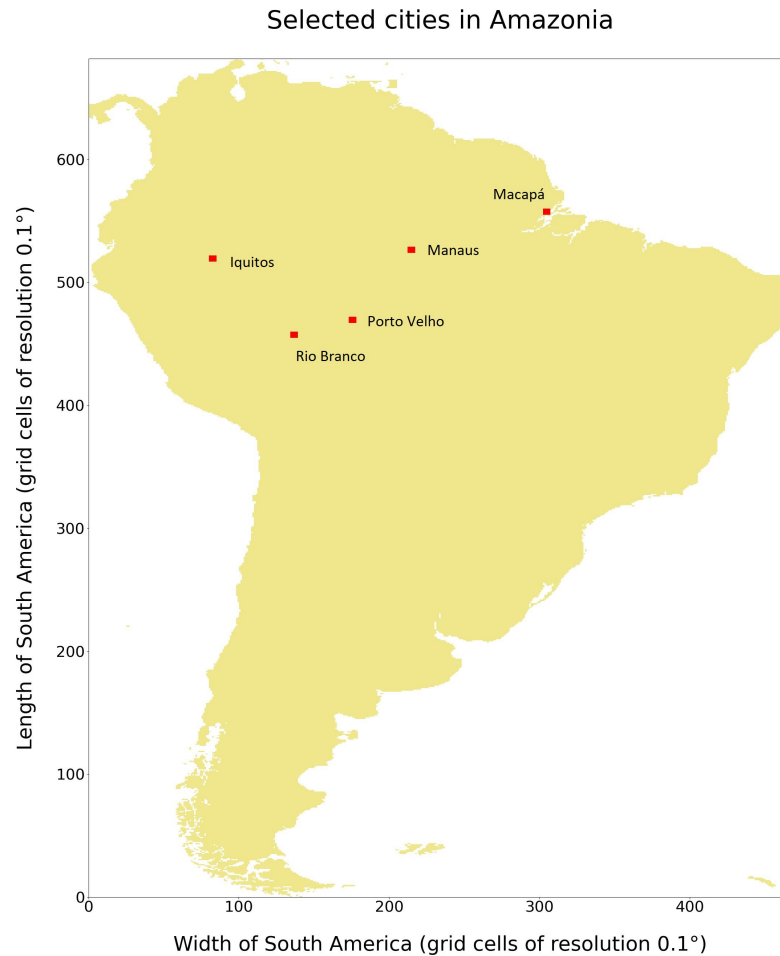


Figure 6.2: Selected cities in the Amazon Basin.

The years, extracted from 100 samples, by which an Indo-European language first reaches these five settlements, are represented in Figure 6.3. The median years in which an Indo-European language first reaches these five settlements are represented in table 6.1. To compare the sampled year values to reality, I use the approximate founding years of the settlements. I cannot use the real years in which an Indo-European language first reached the selected settlements since this information does not exist. Using the founding years is therefore the best available information I can get to compare my sampled data with. Comparing the sampled median years to the settlements' founding years gives a first approximation of the method's accuracy (see chapter 6.2.2), but is in no means a final assessment. Especially for the Amazon Basin region, since for many settlements, e.g., Rio Branco or Porto Velho, the exact founding date is seemingly unknown. More refined methods to assess the sampled years by which an Indo-European language first reaches selected settlements will therefore be necessary in case one wants to continue using my interpolation method. However, since I implemented my settlement seed points (see chapter 3.3) according to the same logic – settlement founding equals dominant Indo-European language

– there is at least continuity within the entire implementation process of my interpolation method.

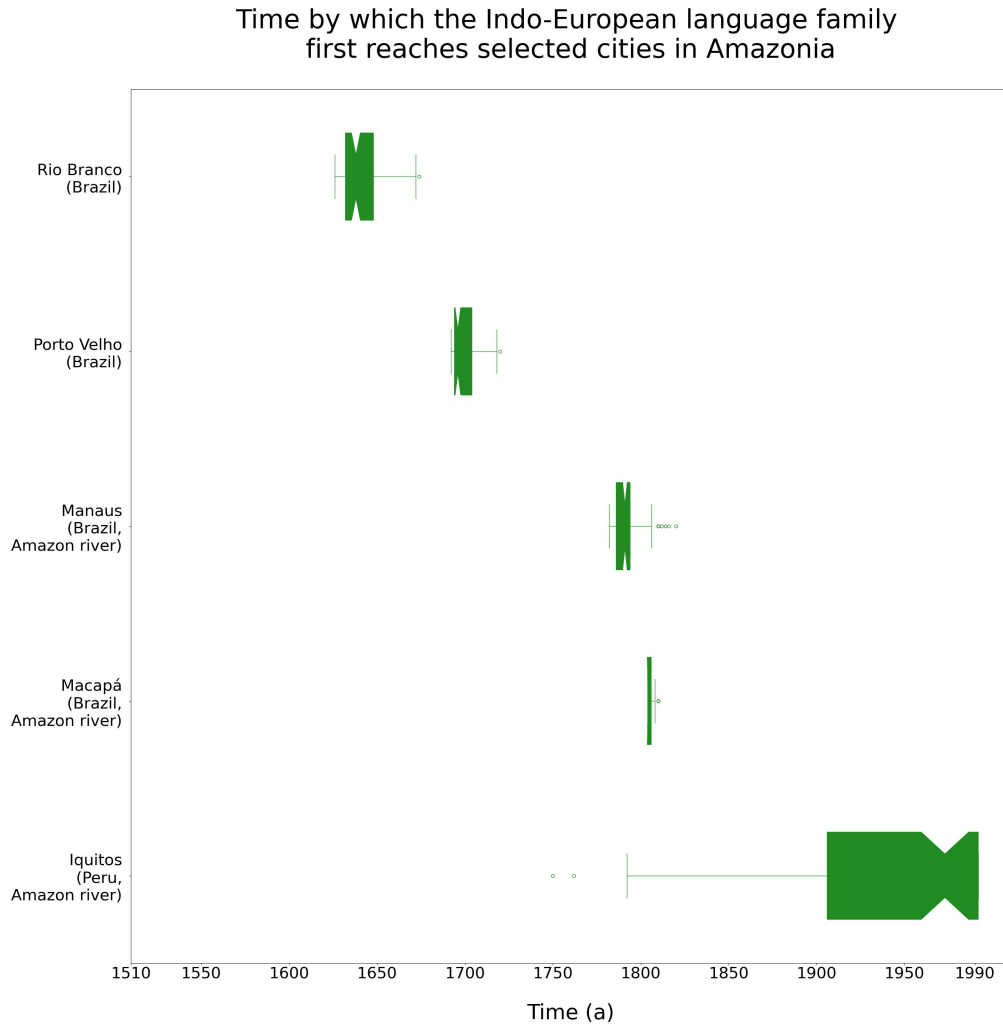


Figure 6.3: Time by which a Indo-European language first reaches selected settlements in the Amazon Basin based upon 100 samples.

Settlement	Median year	Approximate founding year
Rio Branco (Brazil)	1638	\simeq 1860s
Porto Velho (Brazil)	1696	\simeq 1900s
Manaus (Brazil, Amazon)	1791	1669
Macapá (Brazil, Amazon)	1804	1856
Iquitos (Peru, Amazon)	1973	1864

Table 6.1: Summary of the median years by which a Indo-European language reaches selected settlements in the Amazon Basin. Furthermore, the approximate founding years of the settlements are added. (Source for founding dates: *Encyclopaedia Britannica*, see Table [A2](#) in the appendix)

However, the first approximation of the method’s accuracy, based on the selected settlements in the Amazon Basin, already shows some interesting results. On the one hand, it is clear that the interpolation reflects reality since European colonizers and therefore Indo-European language families seem to have reached the Amazon Basin rather late. On the other hand, the settlements along the Amazon river itself are found to speak Indo-European way earlier in reality than in the interpolated evolutionary histories. This indicates the Amazon river to be an important geographical factor which accelerated the colonization and therefore language spreading along its course. Implementing geographical factors into the MCMC seems therefore a valid next step if someone wants to further develop my interpolation method. This is further cemented by the sampled median years around which Indo-European languages first reached the remote areas around Rio Branco and Porto Velho: the sampled years are way earlier than the settlements’ founding dates. The reason for this is that currently, my interpolation method does not take into account the important geographical factors slope and habitat barriers. Indeed, since the Spanish had already founded many settlements on the West coast of South America by the mid of the 16th century – e.g. Lima –, which are also used as seed points in my interpolation method, Indo-European spreads rather fast towards the Western edge of the Amazon Basin in my evolutionary histories. This is exactly where Rio Branco and Porto Velho are located. However, if the geographical factors slope and habitat barriers were included, the Indo-European language spreading towards the Western edge of the Amazon Basin would most likely be slowed down by both the Andes and the switch from mountainous grasslands towards tropical and subtropical forests (Dinerstein et al. [2017](#)). Rio Branco and Porto Velho would then be reached later, closer to their founding year. Finally, the large lower whisker of the boxplot for Iquitos indicates large local differences between the 100 evolutionary histories. Hence, even when considering mere neighbourhood spreading without any geographical factors, a lot of variance is already possible. This is interesting as it shows that, even when considering mere neighbourhood spreading without any geographical factors, a lot of variance is already possible.

In summary, the example of the Amazon Basin shows that, while my developed interpolation method works well in overall, its accuracy would probably strongly benefit from implementing geographical factors. An outline on how to do this is therefore presented in chapter [7.1](#)

7 Future Research

7.1 Implementation of geographical factors

A very interesting and important subsequent work to this thesis would be to implement geographical factors and processes of language spreading into the MCMC. This would not only allow to assess the importance of these factors in the regional context of South America, but also to determine potential spatial factors which could enhance the accuracy of the developed probabilistic interpolation method. In order to facilitate such a task, in the following, a possible workflow to implement some pre-selected driving factors within the MCMC is provided.

7.1.1 Selected factors and processes of language spreading

According to the studies by Gavin et al. (2017), Pacheco Coelho et al. (2019) and Pacheco Coelho et al. (2021), the most important overall driving factors and processes of language spreading are population size (with the related notion of environmentally limited group size per area, i.e., carrying capacity) as well as ecological and climatic factors. This aligns with the overall driving factors and processes of language spreading discussed in a broader literary context, see chapter 2.1.2. However, it is very important to notice that global factors and processes do not seem to exist: indeed, their importance is mostly regionally defined since in reality, the interweaving of several factors and therefore the corresponding local mix is responsible for language spreading (Gavin et al. 2013). Furthermore, the study by Gavin et al. (2017) shows that only three factors and processes can already correctly predict about 50 percent of the spatial language distributions. Finally, Pacheco Coelho et al. (2019) address the problem that sociocultural and historical factors and processes are too complicated to include in grid-based structures and are therefore mostly left out in these three studies: only population size is indirectly included through the concept of carrying capacity, a value which is calculated based upon natural factors like average rain.

This means that the most important driving factors and processes defined by these studies are only a mild recommendation for this thesis since the key factors and processes might regionally differ in South America. However, I will still orient my choice by them. Furthermore, while the terms “spatial language distribution” and “language spreading” are normally defined in relation to languages (see chapter 2.1.1), I will use them for language families and make the assumption that the distribution and expansion of language families are similar to those of languages. Concerning the number of chosen factors of language spreading, I orient my choice by the low, but still successful, number chosen by Gavin et al. (2017) in their study. Therefore, only two climatic or ecological driving factors of language spreading will be selected. These are mountain slopes and habitat barriers (see chapter 2.1.2), since a correct implementation of the ambiguous role coastlines and notably rivers play in language spreading (see chapter 2.1.2) is beyond the scope of this outlook – especially with a complex river system such as the Amazon. Furthermore, sociocultural or historical factors are neglected. This due to the usage of an underlying grid-based structure and because the processing of the numerous data about natural factors necessary to calculate a valid carrying capacity would also exceed the scope of this research outlook. However, this impacts the amount of insufficiently explained spatial language distributions since South America is strongly influenced by the sociocultural impact of colonization between 1510 A.D. and 1990 A.D. It seems therefore recommendable to consider implementing sociocultural and historical factors in further grid-based research. An extensive way of calculating the necessary carrying capacity for it can be found in the paper by

7.1.2 Weighted adjacency lists for geographical costs

Graphs are generally defined as comprising a finite number of nodes and a finite number of – normally ordered – pairs called edges or arcs with the latter being able to contain weights or costs. A typical example of a graph with ordered arcs comprising costs can be seen in Figure 7.1. Instead of a graphical depiction, graphs can also be stored in so-called adjacency lists. These usually contain as many entries as the graph contains nodes and list all outgoing arcs per node (see Table 7.1). An alternative way of depicting an adjacency list is to have one entry per ordered arc and list the start node, end node and weight for each arc (see Table 7.2) (Pressl 2012).

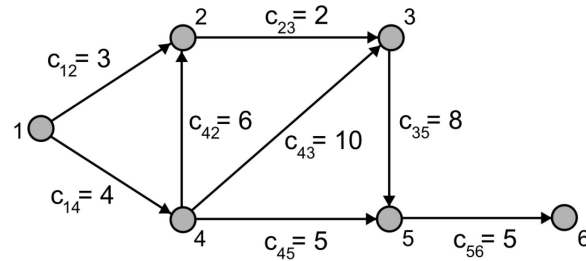


Figure 7.1: Example of a graph with 6 nodes and ordered arcs comprising costs. E.g., the arc between the nodes 1 and 2 has the costs $c_{i,j}=3$. (Source: Pressl 2012)

Node	Outgoing arcs
1	1-2, 1-4
2	2-3
3	3-5
4	4-2, 4-3, 4-5
5	5-6
6	/

Table 7.1: Typical adjacency list containing as many entries as the graph contains nodes and listing all outgoing arcs per node. Adjacency list based upon the example in Figure 7.1

Within the MCMC, the spatial language distributions can be seen as grid-like graphs with the original $c = 468 \times 682 = 319'176$ grid cells from the “starting history” being the nodes of the graph. In compliance with the Moore neighbourhood used in the CA, ordered arcs with costs then part from each node P to the node’s eight neighbours as well as the node itself (see Figure 7.2). The costs $c_{P,j}$ from each node P to its neighbouring nodes are in the current version of the MCMC all the same and normalized to 1: each cell has the same probability $a_{i,j}$ (as defined in equations (7) and (9) on page 24) to be copied into $P_{i,j}$. However, using a grid-like graph stored within adjacency lists has the huge advantage of allowing different costs for each arc and therefore also allowing to have very different probabilities $a_{i,j}$ for the cells to be copied into $P_{i,j}$. This is especially interesting for implementing geographical factors as a much higher variability of the geographical factors can be implemented through the various costs. Indeed, the probability $a_{i,j}$ of each cell to be copied into $P_{i,j}$ does not depend on its value, but on the geographical costs to switch from $P_{i,j}$ to that cell. While the value of a cell stays unchanged

during a time step t , the costs to reach it always change according to the node from where one tries to reach it.

Arc	Start node	End node	Costs/Weights per arc
1-2	1	2	3
1-4	1	4	4
2-3	2	3	2
3-5	3	5	8
4-2	4	2	6
4-3	4	3	10
4-5	4	5	5
5-6	5	6	5

Table 7.2: Adjacency list containing as many entries as the graph contains ordered arcs and listing the start node, end node and weight per arc. Adjacency list based upon the example in Figure 7.1.

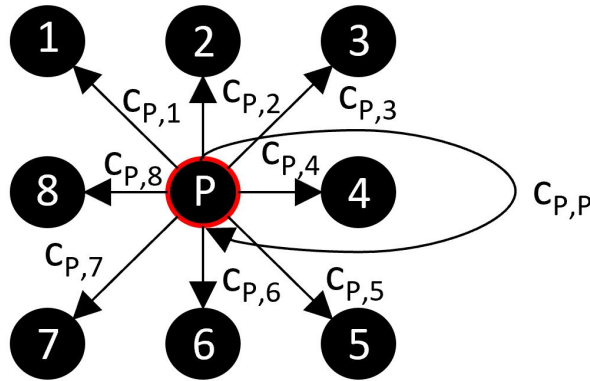


Figure 7.2: Nine ordered arcs with costs $c_{P,j}$ leaving a central node P according to a Moore neighbourhood.

Another advantage of the spatial language distributions being stored as graphs instead of grids in the MCMC is that the grid-like nature of the graphs could, at some point, even be discarded in favour of a more sparse graph. This would then allow to also envision language spreading with given costs towards non-neighbouring cells. However, this will not be further explored in this thesis where the developed probabilistic method for interpolating spatial language distributions is tied to language spreading between neighbours only.

7.1.3 Outline on implementing selected geographical factors

The first selected factor of spatial language spreading should be ecosystem barriers. The first step towards its implementation is to choose a fitting habitat data set, e.g., the one presented in chapter 3.4.1. This data set then needs to be rasterized within the same extent – the bounding box defined in chapter 3.1 – and with the same resolution, 0.1° , as the two language data sets TOC and C. Subsequently, each cell of the rasterized habitat data set contains a numerical value representing a habitat. The rasterized habitat data set is fed into the MCMC in addition to the “starting history” and will be overlaying each spatial language distribution at each time step t in order to assess the geographical costs of copying a neighbouring cell given habitat information.

The rasterized habitat data set is modelled as a grid-like graph and stays the same for all time steps. This under the assumption that the habitats in South America have been mostly constant over the past 500 years. This is a huge simplification of reality given habitat changes are prone to have occurred during that time span - e.g., through deforestation. However, this assumption of habitat consistency is deemed a necessary step to allow habitat information to be included in the MCMC.

As described in chapter [2.1.2](#), people and therefore languages tend to move along similar habitats if possible. Hence, if copying a neighbouring cell equals crossing a habitat barrier in the overlaying habitat data set, the geographical costs for the copying are higher than for copying a neighbouring cell lying within the same habitat as $P_{i,j}$. Since there do not seem to exist any numerical values on the preference of moving along the same habitats, I arbitrarily fix the costs for switching towards a different habitat at 10 and the costs for switching towards the same habitat at 1. These values can easily be changed. However, in case several geographical factors are taken into account, it is important that the values chosen for the habitat costs are around the same magnitude then the costs of the other geographical factors. The probability $a_{i,j}$ for a cell to be copied into $P_{i,j}$ based upon habitat costs can be expressed through:

$$a_{i,j} = \begin{cases} \frac{\frac{1}{c_{P,j}}}{\sum_{n=1}^9 \frac{1}{c_{P,j}}} & \text{for a current neighbour or previous } P_{i,j} \text{ itself} \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

with:

$$c_{habitat_{P,j}} = \begin{cases} 10 & \text{for switching to a different habitat} \\ 1 & \text{for switching to the same habitat} \end{cases} \quad (13)$$

The second factor of spatial language spreading I select are mountain slopes. The steps towards their implementation are similar to the ones for the habitat barriers. First you have to choose a fitting DEM data set, e.g., the one presented in chapter [3.4.2](#). The slopes data set then also needs to be rasterized within the same extent and with the same resolution as the two language data sets TOC and C. Subsequently, each cell of the rasterized slopes data set contains a numerical value representing the cell's elevation. The rasterized habitat data set is fed into the MCMC in addition to the "starting history" and will be overlaying each spatial language distribution at each time step t in order to assess the geographical costs of copying a neighbouring cell given slope information. The rasterized slopes data set is modelled as a grid-like graph and stays the same for all time steps under the assumption that the topography in South America to have been constant over the past 500 years. This is again a huge simplification of reality given topography changes are prone to have occurred during that time span - e.g., through land slides. However, this assumption of topography consistency is deemed a necessary step to allow slopes information to be included in the MCMC.

As described in chapter [2.1.2](#), language spreading in the mountains occurs bottom-up with the valley languages slowly eradicating the high altitudes' languages. A first set of geographical costs for slopes is equal to the change in elevation when switching from the central node to one of the neighbouring nodes in the overlaying slopes data set. Knowing that each cell of the underlying spatial language distributions has a cell size of 10 km x 10 km due to the rasterization with 0.1°, one can calculate the slope in percent. The slope is already a more refined geographical

cost then a simple change in elevation. It is however important to maintain both negative and positive slopes for the geographical costs. It is therefore suggested to categorize the slopes in a third step. This has one big advantage: the slope categories do not need to follow the slopes' metric ordinal scale and so-created slope costs can therefore better represent the bottom-up language spreading. An exemplary slope categorization following the Canadian Governments categorizations (Canadian State 2013) can be seen in table 7.3. The lowest geographical costs occur at steep uphill slopes (15% to 60%), representing the tendency that languages spread bottom-up in more mountainous regions. Too steep uphill regions (> 60%) have higher costs though since they are less accessible and therefore probably also less populated. Areas with flat and gentle slopes (-9% to 9%) are considered neutral middle ground concerning the costs. Finally, the steeper the slope goes downhill, the higher the geographical costs.

Category / Geographical cost $c_{P,j}$	Slope range (in %)
1	15% to 30%
2	30% to 60%
3	9% to 15%
4	> 60%
5	-9% to 9%
6	-15% to -9%
7	-30% to -15%
8	-60% to -30%
9	< -60%

Table 7.3: Table showing the slope categories, i.e., the geographical costs given slope information, for certain slope ranges

The probability $a_{i,j}$ for a cell to be copied into $P_{i,j}$ based upon slope costs can therefore be expressed through:

$$a_{i,j} = \begin{cases} \frac{1}{\sum_{n=1}^9 \frac{1}{c_{P,j}}} & \text{for a current neighbour or previous } P_{i,j} \text{ itself} \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

with $c_{slopes_{P,j}}$: see table 7.3

In case the two geographical factors, i.e., habitat barriers and slopes, are both to be added simultaneously, the two corresponding rasterized data sets need to be implemented into the MCMC in addition to the “starting history”. It is in that case important that all the involved geographical factors have about the same magnitude. Otherwise, the geographical factor with the significantly higher magnitude will outshine the other. Furthermore, it is of course also possible to include more – or different – spatial factors of language spreading than the two presented here. The outline on how to implement selected geographical factors is indeed merely meant to give an idea of and starting point for useful future research related to the here presented probabilistic interpolation method for spatial language distributions.

7.2 Running the interpolation with non-binary language data

It is possible to run the CA to interpolate between non-binarised pre-processed language grids for TOC and C. This results in a potential interpolation with the spatial language distributions

containing 110 language families instead of 2 (see Figure 7.3). However, such an interpolation will then need an updated MCMC which has reduced computation time for non-binary cell values. This is crucial since the CA alone does not allow for a deterministic and therefore valid interpolation. Furthermore, a different approach towards dealing with multilingualism would be advisable since the here presented approach leads to a loss in information and potential errors when dealing with more than two language families.

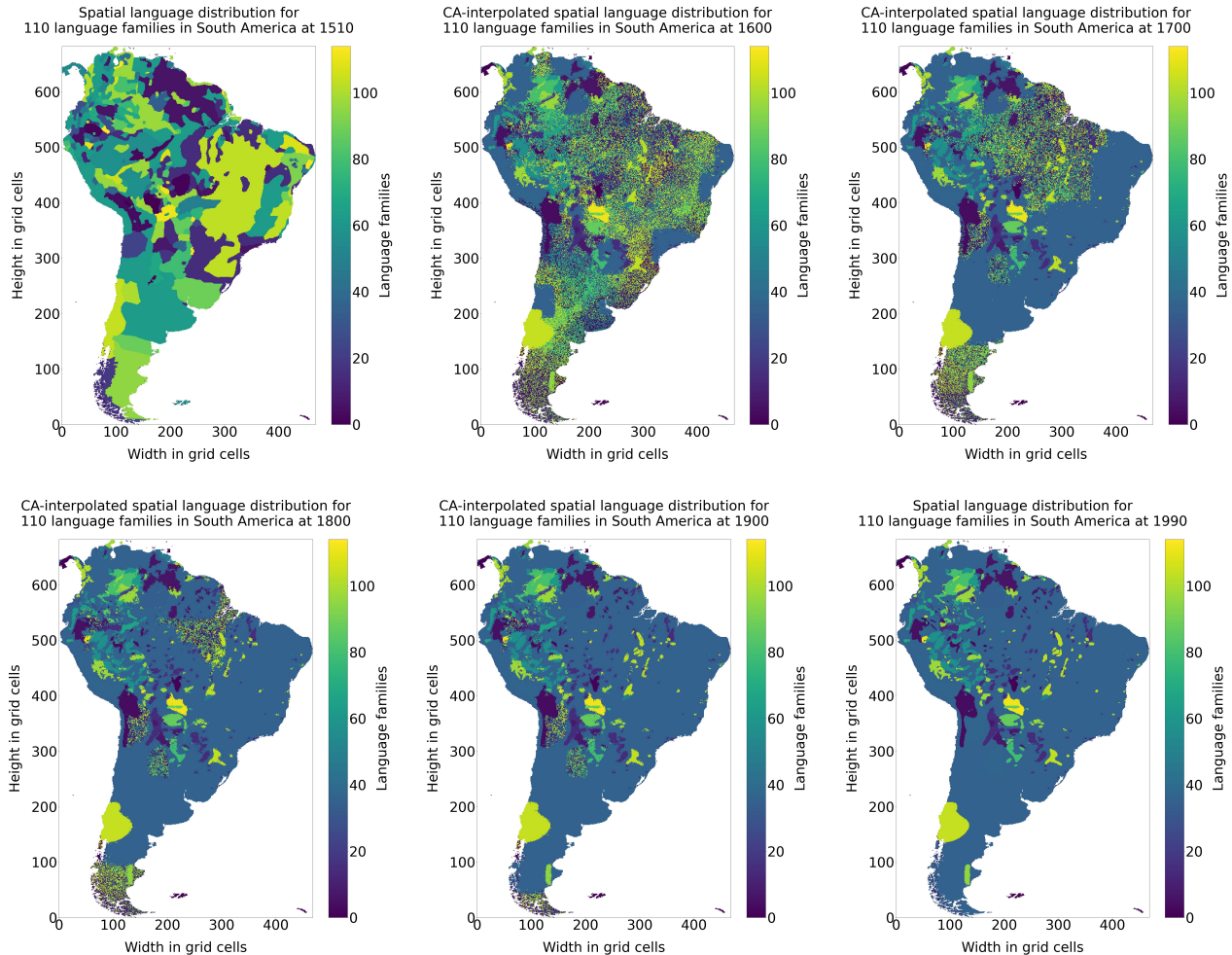


Figure 7.3: CA-induced interpolation steps for spatial language distributions containing 110 language families at 100-year-intervals.

The CA-inferred interpolation for multiple language families also needs improvement through the MCMC’s likelihood function and prior probability since in areas where the Indigenous language families are getting driven out and mixed up, the interpolated spatial language distributions just show random scatter noise (see Figure 7.3). This should however improve if these distributions are implemented as “starting history” into an MCMC. Moreover, one could start to slowly raise the number of included language families instead of switching directly from 2 to 110 language families in order to reduce the random scatter noise.

The big asset of being able to run the here presented interpolation method with several language families is that it makes the method more robust and polyvalent: the spreading patterns of not only a single language family, but of several families could be simulated in parallel. This would allow for more in-depth knowledge about an area’s language richness and distribution over time

as well as its related cultural development. Subsequently, the interpolation method could then also become interesting for areas where, contrary to South America, basic knowledge about language spreading already exists.

8 Conclusion

Based on theories originating from research on language diversity and inspired by macro ecology, a probabilistic method for interpolating spatial language distributions in a given area and over a certain time span has been developed. This method is composed of a cellular automaton as core underlying mechanism and Bayesian inference as statistical method. The output of the former, called “starting history”, is the input for the MCMC, the tool of choice to implement Bayesian inference in case the Bayesian model, like in the present case, contains many parameters. The interpolation was executed for South America between the existing spatial language distributions of “Time of Contact”, i.e., 1510 A.D., and “Contemporary”, i.e., 1990 A.D. The interpolation was furthermore performed on the scale of language families, i.e., the spreading of the Indo-European language family at the cost of the Indigenous language families was inferred.

While the interpolation method is subject to data uncertainties and restrictions due to modelling choices and limitations, it allows for a good first approximation of the potential spreading pattern of the Indo-European language family in South America between 1510 and 1990. However, the additional implementation of geographical factors within the MCMC will most likely further improve the accuracy of the interpolation method and is therefore highly recommended in case of further use of the interpolation method. Furthermore, if the computation time constraints within the MCMC are solved, the interpolation method can also be used to infer spatial language distributions for several language families.

Related to the case study of South America, the here developed probabilistic method for interpolating spatial language distributions could help to fill existing knowledge gaps in current phylogenetic diversity studies for South America and at contributing to a better understanding of related cultural developments on a continent heavily influenced by European colonisation during the interpolated time period. In a broader context, the developed interpolation method should be transferable to other areas and therefore be able to set the basis for a globally applicable model allowing to conduct linguistic research in various regions across the globe.

References

- Asher, Ronald Eaton and Moseley, Christopher (2007). “Introduction”. In: *Atlas of the World’s Languages*. Ed. by Ronald Eaton Asher and Christopher Moseley. 2nd ed. Publication Title: Atlas of the World’s Languages. Milton Park, Abingdon, Oxfordshire, England, UK: Routledge, pp. 1–3. ISBN: 978-1-315-82984-5. DOI: [10.4324/9781315829845-1](https://doi.org/10.4324/9781315829845-1).
- Beltran, Francesc; Herrando, Salvador; Estreder, Violant; Ferreres, Adoración; Adell, Marc-Antoni and Ruiz-Soler, Marcos (2010). “A Language Shift Simulation Based on Cellular Automata”. In: *Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models*. Ed. by Emmanuel G. Blanchard and Danièle Allard. Hershey, Pennsylvania, USA: IGI Global, pp. 136–151. ISBN: 978-1-61520-883-8. DOI: [10.4018/978-1-61520-883-8.ch007](https://doi.org/10.4018/978-1-61520-883-8.ch007).
- Bentz, Christian; Dediu, Dan; Verkerk, Annemarie and Jäger, Gerhard (2018). “The evolution of language families is shaped by the environment beyond neutral drift”. In: *Nature Human Behaviour* 2.11, pp. 816–821. ISSN: 2397-3374. DOI: [10.1038/s41562-018-0457-6](https://doi.org/10.1038/s41562-018-0457-6).
- Bouckaert, Remco R.; Bowern, Claire and Atkinson, Quentin D. (2018). “The origin and expansion of Pama–Nyungan languages across Australia”. In: *Nature Ecology & Evolution* 2.4, pp. 741–749. ISSN: 2397-334X. DOI: [10.1038/s41559-018-0489-3](https://doi.org/10.1038/s41559-018-0489-3).
- Britannica, The Editors of Encyclopaedia Britannica (2007a). *Huancavelica | Peru | Britannica*. URL: <https://www.britannica.com/place/Huancavelica-Peru>.
- Britannica, The Editors of Encyclopaedia Britannica (2007b). *Trujillo | Peru | Britannica*. URL: <https://www.britannica.com/place/Trujillo-Peru>.
- Britannica, The Editors of Encyclopaedia Britannica (2012a). *Acre | state, Brazil | Britannica*. URL: <https://www.britannica.com/place/Acre-state-Brazil>.
- Britannica, The Editors of Encyclopaedia Britannica (2012b). *Amapá | state, Brazil | Britannica*. URL: <https://www.britannica.com/place/Macapá>.
- Britannica, The Editors of Encyclopaedia Britannica (2012c). *São Vicente | Brazil | Britannica*. URL: <https://www.britannica.com/place/Sao-Vicente-Brazil>.
- Britannica, The Editors of Encyclopaedia Britannica (2014a). *Macapá | Brazil | Britannica*. URL: <https://www.britannica.com/place/Macapá>.
- Britannica, The Editors of Encyclopaedia Britannica (2014b). *Pôrto Velho | Brazil | Britannica*. URL: <https://www.britannica.com/place/Porto-Velho>.
- Britannica, The Editors of Encyclopaedia Britannica (2014c). *Rio Branco | Britannica*. URL: <https://www.britannica.com/place/Rio-Branco/additional-info>.
- Britannica, The Editors of Encyclopaedia Britannica (2015a). *Cumaná | Venezuela | Britannica*. URL: <https://www.britannica.com/place/Cumana>.
- Britannica, The Editors of Encyclopaedia Britannica (2015b). *Iquitos | Peru | Britannica*. URL: <https://www.britannica.com/place/Iquitos>.
- Britannica, The Editors of Encyclopaedia Britannica (2015c). *Potosí | Bolivia | Britannica*. URL: <https://www.britannica.com/place/Potosi-Bolivia>.
- Britannica, The Editors of Encyclopaedia Britannica (2016). *Olinda | Brazil | Britannica*. URL: <https://www.britannica.com/place/Olinda>.
- Britannica, The Editors of Encyclopaedia Britannica (2018a). *Concepción | Chile | Britannica*. URL: <https://www.britannica.com/place/Concepcion-Chile>.
- Britannica, The Editors of Encyclopaedia Britannica (2018b). *Piura | Peru | Britannica*. URL: <https://www.britannica.com/place/Piura-Peru>.
- Britannica, The Editors of Encyclopaedia Britannica (2019a). *Manaus | History, Population, & Facts | Britannica*. URL: <https://www.britannica.com/place/Manaus>.

- Britannica, The Editors of Encyclopaedia Britannica (2019b). *Spanish viceroyalties and Portuguese territories*. URL: <https://www.britannica.com/art/Latin-American-architecture>.
- Britannica, The Editors of Encyclopaedia Britannica (2020a). *Asuncion: History, Capital, & Facts | Britannica*. URL: <https://www.britannica.com/place/Asuncion>.
- Britannica, The Editors of Encyclopaedia Britannica (2020b). *Santa Marta | Colombia | Britannica*. URL: <https://www.britannica.com/place/Santa-Marta>.
- Britannica, The Editors of Encyclopaedia Britannica (2021). *Salvador | Brazil | Britannica*. URL: <https://www.britannica.com/place/Salvador-Brazil>.
- Britannica, The Editors of Encyclopaedia Britannica (2022a). *Cochabamba | History & Facts | Britannica*. URL: <https://www.britannica.com/place/Cochabamba>.
- Britannica, The Editors of Encyclopaedia Britannica (2022b). *Quito | History, Map, Population, & Facts | Britannica*. URL: <https://www.britannica.com/place/Quito>.
- Britannica, The Editors of Encyclopaedia Britannica (2022c). *Santiago | History, Map, Population, & Facts | Britannica*. URL: <https://www.britannica.com/place/Santiago-Chile>.
- Britannica, The Editors of Encyclopaedia Britannica (2023a). *Cuzco | Peru | Britannica*. URL: <https://www.britannica.com/place/Cuzco>.
- Britannica, The Editors of Encyclopaedia Britannica (2023b). *La Paz | History, Elevation, Population, & Facts | Britannica*. URL: <https://www.britannica.com/place/La-Paz-Bolivia>.
- British Library, N.N. (2018). *Maps of the Americas | The British Library*. URL: <https://www.bl.uk/collection-items/maps-of-the-americas-c-1687>.
- Burns, Edward Bradford; James, Preston E.; Martins, Luciano; Schneider, Ronald Milton and Momsen, Richard P. (2023). *Brazil - History | Britannica*. URL: <https://www.britannica.com/place/Brazil/History>.
- Calvert, Peter A.R.; Eidt, Robert C. and Donghi, Tulio Halperin (2023). *Argentina - Sports and recreation | Britannica*. URL: <https://www.britannica.com/place/Argentina/Sports-and-recreation>.
- Campbell, Lyle (2012). "Classification of the indigenous languages of South America". In: *The Indigenous Languages of South America*. Ed. by Lyle Campbell and Verónica Grondona. The World of Linguistics [WOL] 2. Berlin ; Boston: De Gruyter Mouton, pp. 59–166. ISBN: 978-3-11-025513-3. DOI: [10.1515/9783110258035.59](https://doi.org/10.1515/9783110258035.59).
- Campbell, Lyle (2019). "How Many Language Families are there in the World?" In: *Anuario del Seminario de Filología Vasca "Julio de Urquijo"* 52.1/2, p. 133. ISSN: 2444-2992, 0582-6152. DOI: [10.1387/asju.20195](https://doi.org/10.1387/asju.20195).
- Campbell, Lyle and Grondona, Verónica (2012). "Preface". In: *The indigenous languages of South America: a comprehensive guide*. Ed. by Lyle Campbell and Verónica María Grondona. The World of Linguistics [WOL] 2. Berlin ; Boston: Mouton de Gruyter, pp. ix–xi. ISBN: 978-3-11-025513-3. DOI: [10.1515/9783110258035](https://doi.org/10.1515/9783110258035).
- Canadian State, Canada Agriculture and Agri-Food (2013). *Slope Gradient*. URL: <https://sis.agr.gc.ca/cansis/nsdb/slc/v3.2/cmp/slope.html>.
- Carmagnani, Marcello A.; Johnson, John J.; Caviades, César N. and Drake, Paul W. (2023). *Chile - Recreation | Britannica*. URL: <https://www.britannica.com/place/Chile/Recreation>.
- Connolly, Sean R.; Keith, Sally A.; Colwell, Robert K. and Rahbek, Carsten (2017). "Process, Mechanism, and Modeling in Macroecology". In: *Trends in Ecology & Evolution* 32.11, pp. 835–844. ISSN: 01695347. DOI: [10.1016/j.tree.2017.08.011](https://doi.org/10.1016/j.tree.2017.08.011).
- Cooper, Robert L. (1982). *Language Spread: Studies in Diffusion and Social Change*. Ed. by Robert L. Cooper. Bloomington, Indiana, USA: Indiana University Press.

- Cubero-Hernández, Antonio; Raony Silva, Eudes and Arroyo Duarte, Silvia (2022). “Urban Layout of the First Ibero-American Cities on the Continent through Conventual Foundations: The Cases of Santo Domingo (1502) and Panama Viejo (1519)”. In: *Land* 11.12, p. 2144. ISSN: 2073-445X. DOI: [10.3390/land11122144](https://doi.org/10.3390/land11122144).
- Das, Debasis (2011). “A Survey on Cellular Automata and Its Applications”. In: *Communications in Computer and Information Science*. Series Title: Communications in Computer and Information Science, pp. 753–762. DOI: [10.1007/978-3-642-29219-4_84](https://doi.org/10.1007/978-3-642-29219-4_84).
- Dastrup, R. Adam (2020). *Introduction to World Regional Geography*. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. URL: <https://slcc.pressbooks.pub/worldgeography/>.
- Davies, Thomas M.; Burr, Robert N.; Moore, John Preston; Pulgar-Vidal, Javier and Kus, James S. (2023). *Peru - Discovery and exploration by Europeans / Britannica*. URL: <https://www.britannica.com/place/Peru/History>.
- Dinerstein, Eric; Olson, David; Joshi, Anup; Vynne, Carly; Burgess, Neil D.; Wikramanayake, Eric; Hahn, Nathan; Palminteri, Suzanne; Hedao, Prashant; Noss, Reed; Hansen, Matt; Locke, Harvey; Ellis, Erle C.; Jones, Benjamin; Barber, Charles Victor; Hayes, Randy; Kormos, Cyril; Martin, Vance; Crist, Eileen; Sechrest, Wes; Price, Lori; Baillie, Jonathan E. M.; Weeden, Don; Suckling, Kierán; Davis, Crystal; Sizer, Nigel; Moore, Rebecca; Thau, David; Birch, Tanya; Potapov, Peter; Turubanova, Svetlana; Tyukavina, Alexandra; Souza, Nadia de; Pintea, Lilian; Brito, José C.; Llewellyn, Othman A.; Miller, Anthony G.; Patzelt, Annette; Ghazanfar, Shahina A.; Timberlake, Jonathan; Klöser, Heinz; Shennan-Farpón, Yara; Kindt, Roeland; Lillesø, Jens-Peter Barnekow; Breugel, Paulo van; Graudal, Lars; Vogé, Maianna; Al-Shammari, Khalaf F. and Saleem, Muhammad (2017). “An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm”. In: *BioScience* 67.6, pp. 534–545. ISSN: 0006-3568, 1525-3244. DOI: [10.1093/biosci/bix014](https://doi.org/10.1093/biosci/bix014). URL: <https://academic.oup.com/bioscience/article-lookup/doi/10.1093/biosci/bix014>.
- Eberhard, David M.; Simons, Gary F. and Fennig, Charles D. (2022). *Ethnologue: Languages of the World. Twenty-fifth edition*. URL: <https://www.ethnologue.com/>.
- Ebert, Christopher (2019). *The Dutch in South America and the Caribbean*. DOI: [10.1093/OBO/9780199766581-0214](https://doi.org/10.1093/OBO/9780199766581-0214). URL: <https://www.oxfordbibliographies.com/display/document/obo-9780199766581/obo-9780199766581-0214.xml>.
- Evers, Jeannie (2023). *First Contact in the Americas / National Geographic Society*. URL: <https://education.nationalgeographic.org/resource/first-contact-americas>.
- Gavin, Michael C.; Botero, Carlos A.; Bown, Claire; Colwell, Robert K.; Dunn, Michael; Dunn, Robert R.; Gray, Russell D.; Kirby, Kathryn R.; McCarter, Joe; Powell, Adam; Rangel, Thiago F.; Stepp, John R.; Trautwein, Michelle; Verdolin, Jennifer L. and Yanega, Gregor (2013). “Toward a Mechanistic Understanding of Linguistic Diversity”. In: *BioScience* 63.7, pp. 524–535. ISSN: 1525-3244, 0006-3568. DOI: [10.1525/bio.2013.63.7.6](https://doi.org/10.1525/bio.2013.63.7.6).
- Gavin, Michael C.; Rangel, Thiago F.; Bown, Claire; Colwell, Robert K.; Kirby, Kathryn R.; Botero, Carlos A.; Dunn, Michael; Dunn, Robert R.; McCarter, Joe; Pacheco Coelho, Marco Túlio; Gray, Russell D. and Hurlbert, Allen (2017). “Process-based modelling shows how climate and demography shape language diversity”. In: *Global Ecology and Biogeography* 26.5, pp. 584–591. ISSN: 1466-822X, 1466-8238. DOI: [10.1111/geb.12563](https://doi.org/10.1111/geb.12563).
- GLOBE, Task Team and others (1999). *The Global Land One-kilometer Base Elevation (GLOBE) Digital Elevation Model, Version 1.0*. Publisher: National Oceanic and Atmospheric Administration, National Geophysical Data Center, 325 Broadway, Boulder, Colorado 80305-3328, U.S.A. URL: <https://www.ngdc.noaa.gov/mgg/topo/globe.html>.

- Gotelli, Nicholas J.; Anderson, Marti J.; Arita, Hector T.; Chao, Anne; Colwell, Robert K.; Connolly, Sean R.; Currie, David J.; Dunn, Robert R.; Graves, Gary R.; Green, Jessica L.; Grytnes, John-Arvid; Jiang, Yi-Huei; Jetz, Walter; Kathleen Lyons, S.; McCain, Christy M.; Magurran, Anne E.; Rahbek, Carsten; Rangel, Thiago F.L.V.B.; Soberón, Jorge; Webb, Campbell O. and Willig, Michael R. (2009). “Patterns and causes of species richness: a general simulation model for macroecology”. In: *Ecology Letters* 12.9, pp. 873–886. ISSN: 1461023X, 14610248. DOI: [10.1111/j.1461-0248.2009.01353.x](https://doi.org/10.1111/j.1461-0248.2009.01353.x).
- Greenhill, Simon J. (2014). “Demographic correlates of language diversity”. In: *The Routledge Handbook of Historical Linguistics*. Ed. by Claire Bowerman and Bethwyn Evans. 1st ed. . Milton Park, Abingdon, Oxfordshire, England, UK: Routledge, pp. 555–578. ISBN: 978-0-415-52789-7. DOI: [10.4324/9781315794013](https://doi.org/10.4324/9781315794013).
- Grollemund, Rebecca; Branford, Simon; Bostoen, Koen; Meade, Andrew; Venditti, Chris and Pagel, Mark (2015). “Bantu expansion shows that habitat alters the route and pace of human dispersals”. In: *Proceedings of the National Academy of Sciences* 112.43, pp. 13296–13301. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.1503793112](https://doi.org/10.1073/pnas.1503793112).
- Hammarström, Harald; Forkel, Robert; Haspelmath, Martin and Bank, Sebastian (2022). *Glottolog 4.7*. Published: Max Planck Institute for Evolutionary Anthropology. DOI: [10.5281/zenodo.7398962](https://doi.org/10.5281/zenodo.7398962).
- Heckel, Heather D.; Martz, John D.; McCoy, Jennifer L. and Lieuwen, Edwin (2023). *Venezuela - Sports and recreation | Britannica*. URL: <https://www.britannica.com/place/Venezuela/Sports-and-recreation>.
- IPBES, Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (2023). *Ecoregion*. URL: <https://www.ipbes.net/glossary-tag/ecoregion>.
- Kaufman, Terrence (1990). “Language History in South America: What We Know and How to Know More”. In: *Amazonian Linguistics: Studies in Lowland South American Languages*, pp. 13–73. URL: https://www.researchgate.net/publication/288841782_Language_history_in_South_America_What_we_know_and_how_to_know_more.
- Kaufman, Terrence (2007). “South America”. In: *Atlas of the World’s Languages*. Ed. by R.E. Asher and Christopher Moseley. 2nd ed. Milton Park, Abingdon, Oxfordshire, England, UK: Routledge, pp. 61–95. ISBN: 978-1-315-82984-5. DOI: [10.4324/9781315829845-3](https://doi.org/10.4324/9781315829845-3).
- Keeling, David J.; Pastor, José M.F. and Tulchin, Joseph S. (2022). *Buenos Aires - History | Britannica*. URL: <https://www.britannica.com/place/Buenos-Aires/History>.
- Keen, Benjamin (2023). *Vasco Nunez de Balboa | Accomplishments, Route, & Facts | Britannica*. URL: <https://www.britannica.com/biography/Vasco-Nunez-de-Balboa>.
- Kline, Harvey F.; Garavito, Clemente; Parsons, James J.; Gilmore, Robert Louis and McCreevey, William Paul (2023). *Colombia - Sports and recreation | Britannica*. URL: <https://www.britannica.com/place/Colombia/Sports-and-recreation>.
- Kmusser (2013). *Map of the Amazon River drainage basin with the Amazon River highlighted*. CC BY-SA 3.0 Wikimedia Commons. URL: <https://upload.wikimedia.org/wikipedia/commons/f/ff/Amazonrivermap.svg>.
- Lameli, Alfred (2009). “32. Linguistic atlases - traditional and modern”. In: *Volume 1 Theories and Methods*. Ed. by Peter Auer and Jürgen Erich Schmidt. Berlin ; Boston: De Gruyter Mouton, pp. 567–592. ISBN: 978-3-11-022027-8. DOI: [10.1515/9783110220278.567](https://doi.org/10.1515/9783110220278.567).
- Mace, Ruth and Pagel, Mark (1995). “A latitudinal gradient in the density of human languages in North America”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 261.1360, pp. 117–121. ISSN: 0962-8452, 1471-2954. DOI: [10.1098/rspb.1995.0125](https://doi.org/10.1098/rspb.1995.0125).

- MacLeod, Murdo J.; Pozo Vélez, Homero and Knapp, Gregory W. (2023). *Ecuador - Media and publishing* / *Britannica*. URL: <https://www.britannica.com/place/Ecuador/Media-and-publishing>.
- McElreath, Richard (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 1st edition. Boca Raton, Florida, USA: Chapman and Hall/CRC. ISBN: 978-1-315-37249-5. DOI: [10.1201/9781315372495](https://doi.org/10.1201/9781315372495).
- McFarren, Peter J. and Arnade, Charles W. (2023). *Bolivia - Press and telecommunications* / *Britannica*. URL: <https://www.britannica.com/place/Bolivia/Press-and-telecommunications>.
- Moore, Joslin L.; Manne, Lisa; Brooks, Thomas; Burgess, Neil D.; Davies, Robert; Rahbek, Carsten; Williams, Paul and Balmford, Andrew (2002). “The distribution of cultural and biological diversity in Africa”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269.1501, pp. 1645–1653. ISSN: 0962-8452, 1471-2954. DOI: [10.1098/rspb.2002.2075](https://doi.org/10.1098/rspb.2002.2075).
- Mufwene, Salikoko S. (2006). “Language Spread”. In: *Encyclopedia of Language & Linguistics*. Ed. by Keith Brown. 2nd ed. Amsterdam, Netherlands: Elsevier, pp. 613–616. ISBN: 978-0-08-044854-1. DOI: [10.1016/B0-08-044854-2/01291-8](https://doi.org/10.1016/B0-08-044854-2/01291-8).
- N.N. (2023a). *Archaeological Site of Panama Viejo and Historic District of Panama*. URL: <https://whc.unesco.org/en/list/790/>.
- N.N. (2023b). *GeoHack (0; 0)*. URL: https://geohack.toolforge.org/geohack.php?params=N_E.
- N.N. (2023c). *Historic Centre of Lima*. URL: <https://whc.unesco.org/en/list/500/>.
- N.N. (2023d). *Historic Centre of Salvador de Bahia*. URL: <https://whc.unesco.org/en/list/309/>.
- N.N. (2023e). *Historic Centre of the Town of Olinda*. URL: <https://whc.unesco.org/en/list/189/>.
- Nettle, Daniel (1998). “Explaining Global Patterns of Language Diversity”. In: *Journal of Anthropological Archaeology* 17.4, pp. 354–374. ISSN: 02784165. DOI: [10.1006/jaar.1998.0328](https://doi.org/10.1006/jaar.1998.0328).
- NGS, National Geographic Society (2022). *Biomes, Ecosystems, and Habitats*. URL: <https://education.nationalgeographic.org/resource/biomes-ecosystems-and-habitats>.
- Nichols, Johanna (1997). “Modeling Ancient Population Structures and Movement in Linguistics”. In: *Annual Review of Anthropology* 26.1, pp. 359–384. ISSN: 0084-6570, 1545-4290. DOI: [10.1146/annurev.anthro.26.1.359](https://doi.org/10.1146/annurev.anthro.26.1.359).
- Nickson, R. Andrew; Painter, James E.; Williams, John Hoyt; Butland, Gilbert James and Service, Elman R. (2022). *Paraguay - Daily life and social customs* / *Britannica*. URL: <https://www.britannica.com/place/Paraguay/Daily-life-and-social-customs>.
- Oppenheim, Gary M and Jones, Manon W (2019). *Languages without armies: Dialect alignment in word production*. Tech. rep. Bangor, Wales, UK: School of Psychology, Bangor University. URL: https://oppenheim-lab.bangor.ac.uk/pubs/oppenheimJones_submitted2019_dialectPriming.pdf.
- Pacheco Coelho, Marco Túlio; Haynie, Hannah J.; Bower, Claire; Colwell, Robert K; Greenhill, Simon J.; Kirby, Kathryn R.; Rangel, Thiago F. and Gavin, Michael C. (2021). *Demographic shifts, inter-group contact, and environmental conditions drive language extinction and diversification*. preprint. SocArXiv. DOI: [10.31235/osf.io/xqr2u](https://doi.org/10.31235/osf.io/xqr2u).
- Pacheco Coelho, Marco Túlio; Pereira, Elisa Barreto; Haynie, Hannah J.; Rangel, Thiago F.; Kavanagh, Patrick; Kirby, Kathryn R.; Greenhill, Simon J.; Bower, Claire; Gray, Russell D.; Colwell, Robert K.; Evans, Nicholas and Gavin, Michael C. (2019). “Drivers of geographical patterns of North American language diversity”. In: *Proceedings of the Royal Society B*:

- Biological Sciences* 286.1899, p. 20190242. ISSN: 0962-8452, 1471-2954. DOI: [10.1098/rspb.2019.0242](https://doi.org/10.1098/rspb.2019.0242).
- Pérez Morales, Edgardo (2020). *Cartagena de Indias*. DOI: [10.1093/OBO/9780199766581-0183](https://doi.org/10.1093/OBO/9780199766581-0183). URL: <https://www.oxfordbibliographies.com/display/document/obo-9780199766581/obo-9780199766581-0183.xml>.
- Phillipson, Robert (2008). “Language Spread”. In: *Volume 3*. Ed. by Ulrich Ammon; Norbert Dittmar; Klaus J. Mattheier and Peter Trudgill. Vol. 3. Sociolinguistics / Soziolinguistik. De Gruyter Mouton, pp. 2299–2306. ISBN: 978-3-11-019987-1. DOI: [10.1515/9783110184181.3.10.2299](https://doi.org/10.1515/9783110184181.3.10.2299).
- Pressl, Bettina (2012). “Knowledge-based route planning. Data modeling and multi-criteria time-constrained route optimization for people with disabilities”. PhD thesis. Institute of Navigation (INAS), Graz, Austria: Graz University of Technology.
- Ranacher, Peter; Van Gijn, Rik and Derungs, Curdin (2017). “Identifying probable pathways of language diffusion in South America”. In: Wageningen, Netherlands: s.n. DOI: [10.5167/UZH-137656](https://doi.org/10.5167/UZH-137656). URL: <https://www.zora.uzh.ch/id/eprint/137656>.
- Sarkar, Palash (2000). “A brief history of cellular automata”. In: *ACM Computing Surveys* 32.1, pp. 80–107. ISSN: 0360-0300, 1557-7341. DOI: [10.1145/349194.349202](https://doi.org/10.1145/349194.349202).
- Schneider, Ronald Milton; Geiger, Pedro P. and Passos Guimarães, Alberto (2022). *Rio de Janeiro | History, Population, Map, Climate, & Facts | Britannica*. URL: <https://www.britannica.com/place/Rio-de-Janeiro-Brazil>.
- Sundberg, Minna (2014). *A comprehensive overlook of the Nordic languages in their Old World language families*. URL: <http://www.sssscomic.com/comic.php?page=196>.
- Wallenfeldt, Jeff (2018). *What Is the Difference Between South America and Latin America?* URL: <https://www.britannica.com/story/what-is-the-difference-between-south-america-and-latin-america>.
- Wallenfeldt, Jeff (2022). *Bogota | Elevation, Population, History, & Facts | Britannica*. URL: <https://www.britannica.com/place/Bogota>.
- Webster, Richard A.; Nowell, Charles E. and Magdoff, Harry (2023). *Western colonialism - The French | Britannica*. URL: <https://www.britannica.com/topic/Western-colonialism/The-French>.
- “Colonization and frontier expansion in Amazônia” (1988). In: *The Demography of Inequality in Brazil*. Ed. by Charles H. Wood and Jose Alberto Magno Carvalho. Cambridge Latin American Studies. Cambridge: Cambridge University Press, pp. 221–236. ISBN: 978-0-521-10246-9. DOI: [10.1017/CB09780511759901.011](https://doi.org/10.1017/CB09780511759901.011).
- Yassemi, S.; Dragičević, S. and Schmidt, M. (2008). “Design and implementation of an integrated GIS-based cellular automata model to characterize forest fire behaviour”. In: *Ecological Modelling* 210.1-2, pp. 71–84. ISSN: 03043800. DOI: [10.1016/j.ecolmodel.2007.07.020](https://doi.org/10.1016/j.ecolmodel.2007.07.020).

Appendix

A1 European settlements: References

Settlement	References
Santa María la Antigua del Darién	Cubero-Hernández et al. 2022 Keen 2023
Panamá Viejo* (Panama City)	Cubero-Hernández et al. 2022 N.N. 2023a
Nueva Toledo (Cumaná)	Britannica 2015a Heckel et al. 2023
Santa Marta	Britannica 2020b Kline et al. 2023
Piura	Britannica 2018b Davies et al. 2023
São Vicente	Britannica 2012c Burns et al. 2023
Cartagena*	Kline et al. 2023 Pérez Morales 2020
Cuzco*	Britannica 2023a Davies et al. 2023
Quito*	Britannica 2022b MacLeod et al. 2023
Trujillo	Britannica 2007b Davies et al. 2023
Lima	Davies et al. 2023 N.N. 2023e
Asunción	Britannica 2020a Keeling et al. 2022 Nickson et al. 2022
Olinda	Britannica 2016 Burns et al. 2023 N.N. 2023c
Chuquisaca* (Sucre)	McFarren et al. 2023
Santa Fé de Bacatá* (Bogota)	Wallenfeldt 2022 Kline et al. 2023
Santiago de Chile	Britannica 2022c Carmagnani et al. 2023
Potosí	Britannica 2015c Davies et al. 2023 McFarren et al. 2023
Nuestra Señora de la Paz* (La Paz)	Britannica 2023b McFarren et al. 2023
Salvador	Britannica 2021 Burns et al. 2023 N.N. 2023d
Concepción	Britannica 2018a Carmagnani et al. 2023
Huancavelica	Britannica 2007a Davies et al. 2023
Rio de Janeiro	Schneider et al. 2022
Caracas	Heckel et al. 2023
Cochabamba	Britannica 2022a McFarren et al. 2023
Buenos Aires	Calvert et al. 2023 Keeling et al. 2022

Table A1: References for the implemented European settlements. Settlements with a * have been built upon Indigenous dwellings or cities.

A2 Amazonian settlements: References

Settlement	Reference
Rio Branco (Brazil)	Britannica 2012a; Britannica 2014c
Porto Velho (Brazil)	Britannica 2014b
Manaus (Brazil, Amazon)	Britannica 2019a
Macapá (Brazil, Amazon)	Britannica 2012b; Britannica 2014a
Iquitos (Peru, Amazon)	Britannica 2015b

Table A2: References for the implemented Amazonian settlements.

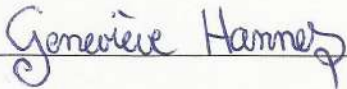
Personal Declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

Place, Date

Zurich, April 27, 2023

Signature

A handwritten signature in blue ink, reading "Genevieve Hamner", is written over a horizontal line. The signature is cursive and includes a decorative flourish at the end of the name.